COMPARISON OF RISK LOCI IN FAMILIAL AND SPORADIC PULMONARY FIBROSIS CASES

by

MATTHEW OLIVER GRACEY

B.S., Colorado State University, 2002

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Master of Science
Epidemiology Program
2016
This thesis for the Master of Science degree by

Matthew Oliver Gracey

has been approved for the

Epidemiology Program

by

Tasha Fingerlin, Chair

John Hokanson

Ivana Yang

Date: May 20, 2016
ABSTRACT

Idiopathic pulmonary fibrosis (IPF) is a disease of unknown etiology that has a poor prognosis for affected individuals. Multiple risk factors and genes are associated with IPF suggesting a complex multifactorial pathogenesis. It has also been proposed that there are familial and sporadic forms of IPF. The aims of this study were to examine whether the impact of 10 known pulmonary fibrosis risk loci of IPF differs between familial and sporadic pulmonary fibrosis and to evaluate the relationship between smoking exposure and a family history of pulmonary fibrosis.

IPF case baseline characteristics were compared by family history of disease with continuous measures using two-sample t-tests and with categorical measures using chi-square tests. Mixed models were used to further evaluate the relationship between smoking and family history of pulmonary fibrosis. Multivariable logistic regression models were used to evaluate the relationship between risk loci and family history of disease. Interaction terms were added to the logistic regression models to evaluate if gene-gene or gene-smoking associations differed by family history of disease.

A total of 1838 pulmonary fibrosis cases were genotyped with 381 (21%) having a family history of pulmonary fibrosis and 1457 (79%) having no family history of disease. Age did not vary between familial and sporadic cases (p=0.53). The frequency of current or former smokers was marginally higher in sporadic cases (p=0.046). Mean pack years of smoking was higher in sporadic cases than it was in familial cases (p <0.004). The frequency of the risk alleles did not differ in familial and sporadic pulmonary fibrosis cases when evaluated individually and collectively (p>0.17). No risk alleles showed significant pairwise interaction differences in familial and sporadic cases.
(p>0.15). Marginal interaction differences between familial and sporadic cases were found in one risk loci (rs4727443) and a history of smoking (p=0.03).

In conclusion, no difference in risk allele frequency for these 10 loci was observed in a large population of familial and sporadic pulmonary fibrosis non-Hispanic white (NHW) cases while exposure to tobacco smoke was significantly higher in sporadic cases.

The form and content of this abstract are approved. I recommend its publication.

Approved: Tasha Fingerlin
ACKNOWLEDGEMENTS

I would foremost like to acknowledge my mentor in epidemiology and thesis chair Dr. Tasha Fingerlin. I am very appreciative of the large amount of time and patience she dedicated to guide me through my introduction to genetic epidemiology. I would also like to thank Dr. John Hokanson and Dr. Ivana Yang for their valuable insights into epidemiological research and for their participation on my thesis committee. I thank the Familial Pulmonary Fibrosis study for taking me on as a research assistant in 2014. Having the opportunity to speak to individuals and family members affected by IPF greatly enriched this thesis research project and master’s program for me. Finally, I’d like to thank all of my other professors and peers in the Epidemiology program at Anschutz. I’m excited to share your important research with my high school students and hopefully inspire a few of them to be part of the next generation of public health researchers.
# TABLE OF CONTENTS

## CHAPTER

### I. BACKGROUND ............................................................................................................................... 1

- Significance .................................................................................................................................. 1
- Study Aims ................................................................................................................................... 9
- Innovation ................................................................................................................................... 10
- Approach .................................................................................................................................. 11

### II. MANUSCRIPT .............................................................................................................................. 17

- Introduction .................................................................................................................................. 17
- Methods ....................................................................................................................................... 18
- Results ........................................................................................................................................ 21
- Discussion ................................................................................................................................... 26

### III. GENOTYPE ANALYSIS .................................................................................................................. 29

- Introduction .................................................................................................................................. 29
- Methods ....................................................................................................................................... 30
- Results ........................................................................................................................................ 32
- Discussion ................................................................................................................................... 36

### IV. CONCLUSION .............................................................................................................................. 38

- REFERENCES ................................................................................................................................... 40
## LIST OF FIGURES

**Figure**

Figure 1. Natural History of Disease for IPF ........................................................................................................... 2

Figure 2. DAG of Hypothesized IPF Disease Development Pathways ................................................................. 8

Figure 3. Study Population ...................................................................................................................................... 11

Figure 4. DAG of Hypothesized Interaction Between Smoking History and Risk Loci ........................................ 14

Figure 5. DAG of Hypothesized Interaction Between Risk Loci for IPF and rs35705950 .................................. 14

Figure 6. Odds of Risk Loci in Familial and Sporadic Cases vs. Controls ............................................................. 24

Figure 7. Correlation Between 10 Risk Loci of IPF .............................................................................................. 33

Figure 8. Comprehensive Set Hierarchical Cluster Solution of Training Sample ............................................ 34

Figure 9. Limited Set Hierarchical Cluster Solution of Training Sample ............................................................. 35
CHAPTER I

BACKGROUND

Significance

Idiopathic interstitial pneumonias (IIPs) are a group of pulmonary diseases of unknown etiology that affect the interstitial space and epithelium of the lung. IIPs are characterized by inflammation and fibrosis (scarring) of the alveoli and pulmonary airways. The most common and severe IIP is idiopathic pulmonary fibrosis (IPF). Pulmonary fibrosis is a term that describes progressive scarring of the interstitial space of alveoli in the lung. Some causes of pulmonary fibrosis are known including asbestos exposure (asbestosis), silica exposure (silicosis) and toxicity from many drugs, including some chemotherapy medications. While IPF is pulmonary fibrosis of unknown etiology, many genetic, demographic and environmental risk factors have been associated with IPF. Moreover, IPF has been described as a complex chronic disease that may contain multiple subtypes with different disease initiation and disease progression factors.

The natural history of disease in IPF cases is diagrammed in Figure 1. Subclinical IPF is ultimately caused by an interaction of genetic risk factors, repeated damage to the interstitium of the lung by environmental particulates and aging. Subclinical IPF is characterized by inflammation and scarring of the interstitial space that may be detectible only through radiographic measures. Clinical, symptomatic IPF is characterized by dyspnea, cough, restrictive impairment of pulmonary function, inspiratory crackling and clubbing of the nails. Diagnosis of definite IPF is made through lung biopsy. If a surgical biopsy is not conducted, a diagnosis of probable IPF is made using high resolution computerized tomography (HRCT) imaging, spirometry measures of lung capacity and other diagnostic criteria including: age over 50, unexplained dyspnea, duration of illness greater than 3 months, crackling upon inspiration and an absence of the aforementioned exposures known to cause pulmonary fibrosis. Histological patterns seen in HRCT scans of the lungs of an IPF patient
include: honeycombing (enlarged spaces due to fibrotic tissue), focal ground glass (partial filling of alveoli), bronchiectasis (thickening of bronchial walls) and architectural distortion of the lungs\(^1\). As a result, gas exchange is reduced in IPF patients both in waking and sleeping states\(^9\). Deterioration of the respiratory system can be acute (significant deterioration occurring in less than a month) or subacute (significant deterioration occurring in a few months)\(^10\).

Figure 1. Natural History of Disease for IPF

Treatment options for clinical IPF are limited. In general, the progression of clinical IPF is slower (mean survival time is higher) in those IPF patients that were never smokers, are at normal BMI, are younger and do not have other comorbid diseases that impair lung function or impair physical activity\(^2,11\). Anti-inflammatory medicines, notably Prednisone, and immune suppressants, notably Azathioprine, have been used historically to treat IPF. These drugs have been shown in some trials to have side effects that may actually reduce survival time\(^12\). A recently released anti-fibrotic medicine, Pirfenidone, has been shown in clinical trials to slow disease progression in IPF patients\(^13\).
One clinical trial found that short-term prognosis was maintained through an exercise program in IPF patients while the IPF cases that didn’t exercise deteriorated over the same time period. The study, however, did not demonstrate significant long-term improvements in prognosis. In late stage pulmonary fibrosis lung transplantation may be used. Double-lung transplant recipients still only have a 5-year survival of approximately 55%.

If untreated, the mean survival time for IPF cases is between 2 and 4 years. Approximately 20-30% of IPF cases survive past 5 years after diagnosis. A study of Medicare IPF patients in the United States found that median survival time was 3.8 years between 2001 and 2011. Most IPF patients ultimately die from respiratory failure or other related conditions including heart failure or infection.

**Impact of Idiopathic Pulmonary Fibrosis**

The prevalence of IPF in the general population of the United States has been estimated to be between 16.3 to 42.7 per 100,000. The incidence of IPF in United States Medicare patients between 2001 and 2011 was estimated at 93.7 per 100,000 person years. The same study found that the prevalence of IPF increased in their study population between 2001 and 2011 from 202.2 cases per 100,000 people in 2001 to 494.5 cases per 100,000 people in 2011. The incidence of IPF varies by ethnicity and geographic location. Medicare patients that reported Hispanic or White ethnicity were more likely to have IPF than those that reported Black ethnicity. Studies of Asian countries have reported much lower rates of IPF than have studies of IPF in Northern European countries and the United States.

The financial burden of IPF is large. Medical costs for IPF patients on Medicare have been estimated to be 134% higher as compared to other Medicare patients. The overall annual financial burden of IPF in the United States has been estimated to be over $1 billion.
**Comorbidities**

IPF has many comorbidities that affect the disease development and survival time of IPF\(^2\). The following diseases are known to be comorbid with IPF: emphysema, sarcoidosis, lung cancer, gastroesophageal reflux disease (GERD), diabetes mellitus, cardiovascular disease, hypertension, ischemic heart disease, depression, osteoporosis\(^{15,20,21}\). For example, small studies have found that cases with combined emphysema and pulmonary fibrosis have shorter survival time than in IPF or emphysema alone\(^{22,23}\).

**Demographic and Exposure Risk Factors**

Male gender and older age are both risk factors of both the development of IPF and poorer outcomes from IPF. The incidence of IPF in Medicare patients in the United States between 2001 and 2011 was higher in male patients and older patients\(^{15}\). The median age of diagnosis of IPF is 66 approximately years. It is not conclusive that male gender is an inherent risk factor of IPF or just associated with greater exposure to other risk factors of IPF\(^2\). A history of smoking has also been associated with IPF disease development, but studies that have evaluated the role of smoking and disease progression have had mixed results\(^2,8,24\).

**Genetic Risk Factors**

Genetic studies have found multiple genes with alleles that are associated with both the development of IPF and survival time in IPF cases. These genetic risk variants include: \(TERT\), \(TERC\), \(SFTPC\), \(SFTPA2\), \(MUC5B\) and \(HLA\)\(^7,25\).

\(TERT/TERC\)

\(TERC\) and \(TERT\) are found on chromosome 3 and 5 respectively and both encode for proteins that are involved in the maintenance of telomere length\(^7\). A recent study found that telomere length was significantly shorter in IPF patients vs. controls and that shorter telomere length was associated...
with shorter mean survival time in IPF patients\textsuperscript{26}. The role of shortened telomeres in the development of IPF is not well understood\textsuperscript{7}.

\textit{SFTPC/SFTPA2}

\textit{SFTPC} and \textit{SFTPA2} are found on chromosome 8 and 10 respectively and both encode for surfactant proteins. It has been hypothesized that mutations in these two genes lead to loss of epithelial cells in the interstitial space of the lung and that those cells are replaced with fibrotic connective tissue\textsuperscript{7}.

\textit{MUC5B}

\textit{MUC5B} is found on Chromosome 11 and encodes for a mucus gel forming protein. A single nucleotide polymorphism (SNP) in the promoter region of \textit{MUC5B} is approximately 6 times more likely to be found in IPF cases than controls\textsuperscript{27}. This SNP (rs35705950) has been shown to be the most predictive of IPF as compared to other SNPs found in or near \textit{MUC5B}. \textit{MUC5B} expression was found to be 14.1 times higher in IPF patients as compared to controls in a lung expression study\textsuperscript{28}. Interestingly, the risk variant of rs35705950 is associated with improved disease prognosis in IPF cases\textsuperscript{29}. The protein product of \textit{MUC5B} has a role in the response to antigen in the lung but the role of \textit{MUC5B} and development of fibrosis is not fully understood\textsuperscript{30}.

\textit{HLA}

\textit{HLA} is a complex of genes found on Chromosome 6 that encode for proteins involved in antigen presentation in the adaptive response of the immune system. \textit{HLA} risk variants for IPF are associated with decreased diffusing capacity of the lung (DLCO)\textsuperscript{31}.

\textit{GWAS}

Genome-wide association studies (GWAS) identify single nucleotide polymorphisms (SNPs) that are associated with disease. An individual SNP is a locus in the human genome in which variation exists\textsuperscript{32}. GWAS typically include a replication study to provide more confidence that the
risk alleles are associated with disease and not a result of chance. A 2013 GWAS and subsequent imputation study of IIP cases and controls found 10 loci in the human genome strongly associated with IIP case status. The odds of a higher risk variant load for these 10 loci in IPF cases as compared to controls when adjusting for age and gender are reported in Table 1 based on the most strongly-associated SNP within each locus.

Table 1. Risk Loci for IPF

<table>
<thead>
<tr>
<th>SNP</th>
<th>Locus</th>
<th>Risk Allele</th>
<th>IPF Case Risk Allele Count</th>
<th>Control Risk Allele Count</th>
<th>IPF Cases vs. Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12696304</td>
<td>3q26</td>
<td>G</td>
<td>871 767 200 1838</td>
<td>1110 784 124 2018</td>
<td>1.34 (1.20,1.50) &lt;0.0001</td>
</tr>
<tr>
<td>rs2609261</td>
<td>4q22</td>
<td>T</td>
<td>1022 680 136 1838</td>
<td>1279 669 70 2018</td>
<td>1.34 (1.19,1.50) &lt;0.0001</td>
</tr>
<tr>
<td>rs10069690</td>
<td>5p15</td>
<td>C</td>
<td>58 555 1184 1797</td>
<td>126 734 1109 1969</td>
<td>1.46 (1.29,1.65) &lt;0.0001</td>
</tr>
<tr>
<td>rs7887</td>
<td>6p21</td>
<td>C</td>
<td>169 761 821 1751</td>
<td>289 833 804 1926</td>
<td>1.24 (1.19,1.38) &lt;0.0001</td>
</tr>
<tr>
<td>rs2076295</td>
<td>6p24</td>
<td>G</td>
<td>394 811 513 1718</td>
<td>537 895 401 1833</td>
<td>1.29 (1.17,1.43) &lt;0.0001</td>
</tr>
<tr>
<td>rs4727443</td>
<td>7q22</td>
<td>A</td>
<td>520 830 364 1714</td>
<td>648 909 270 1827</td>
<td>1.27 (1.14,1.41) &lt;0.0001</td>
</tr>
<tr>
<td>rs9325507</td>
<td>10q24</td>
<td>T</td>
<td>351 874 561 1786</td>
<td>476 973 521 1970</td>
<td>1.18 (1.06,1.30) 0.002</td>
</tr>
<tr>
<td>rs35705950</td>
<td>11p15</td>
<td>T</td>
<td>632 931 140 1703</td>
<td>1454 346 17 1817</td>
<td>5.11 (4.38,5.95) &lt;0.0001</td>
</tr>
<tr>
<td>rs1981997</td>
<td>17q21</td>
<td>G</td>
<td>53 469 1194 1716</td>
<td>79 647 1101 1827</td>
<td>1.33 (1.17,1.52) &lt;0.0001</td>
</tr>
<tr>
<td>rs12610495</td>
<td>19p13</td>
<td>G</td>
<td>735 748 215 1698</td>
<td>923 748 142 1813</td>
<td>1.35 (1.21,1.51) &lt;0.0001</td>
</tr>
</tbody>
</table>

*logistic regression; adjusted for age and gender

Four of these loci are found in or near the previously reported genes associated with IPF: HLA (rs7887), TERT (rs10069690), TERC (rs12696304) and MUC5B (rs35705950). Six of these loci were novel (Table 2). FAM13A is a gene previously associated with chronic obstructive pulmonary disease (COPD). The products of DSP are involved in cell-cell adhesion and have been shown to be involved with the response of epithelial cells to mechanical stress. The products of DPP9 also play a role in cell-cell adhesion. OBFC1, like TERC and TERT, is associated with telomere length maintenance. The region of association on Chromosome 17 contains several genes. While the most associated SNP is near the MAPT gene, the localization of that signal, and therefore the biological role of the relevant gene is not yet clear.
Table 2. Genes Associated with Risk Loci

<table>
<thead>
<tr>
<th>SNP</th>
<th>Locus</th>
<th>Minor Allele</th>
<th>Risk Allele</th>
<th>Associated Gene</th>
<th>Gene Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12696304</td>
<td>3q26</td>
<td>G</td>
<td>G</td>
<td><strong>LRRC34 (TERC)</strong></td>
<td>Telomere length maintenance</td>
</tr>
<tr>
<td>rs2609261</td>
<td>4q22</td>
<td>T</td>
<td>T</td>
<td><strong>FAM13A</strong></td>
<td>Signal transduction, related to hypoxia</td>
</tr>
<tr>
<td>rs10069690</td>
<td>5p15</td>
<td>T</td>
<td>C</td>
<td><strong>TERT</strong></td>
<td>Telomere length maintenance</td>
</tr>
<tr>
<td>rs2076295</td>
<td>6p24</td>
<td>G</td>
<td>G</td>
<td><strong>DSP</strong></td>
<td>Intercellular junction</td>
</tr>
<tr>
<td>rs7887</td>
<td>6p21</td>
<td>A</td>
<td>C</td>
<td><strong>EHMT2; HLA</strong></td>
<td>Antigen presenting proteins</td>
</tr>
<tr>
<td>rs4727443</td>
<td>7q22</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs9325507</td>
<td>10q24</td>
<td>C</td>
<td>T</td>
<td><strong>OBFC1</strong></td>
<td>DNA replication</td>
</tr>
<tr>
<td>rs35705950</td>
<td>11p15</td>
<td>T</td>
<td>T</td>
<td><strong>MUC5B</strong></td>
<td>Mucus formation</td>
</tr>
<tr>
<td>rs1981997</td>
<td>17q21</td>
<td>A</td>
<td>G</td>
<td><strong>MAPT</strong></td>
<td>Microtubule associated protein</td>
</tr>
<tr>
<td>rs12610495</td>
<td>19p13</td>
<td>G</td>
<td>G</td>
<td><strong>DPP9</strong></td>
<td>Serine protease, cell adhesion</td>
</tr>
</tbody>
</table>

*The SNP either is found within the gene or in the promoter region of the gene

Gene regulation may have an important role in the relationship between these genetic risk factors and IPF development. A small number of studies that have evaluated the epigenetics of IPF have reported differences, including differences in methylation patterns and post-transcriptional regulation with microRNA (miRNA), between IPF cases and controls. An understanding of not only the genetic risk factors of IPF but also how those genes or gene products are regulated is necessary to understand the etiology of IPF.

**Familial and Sporadic Pulmonary Fibrosis**

A family history of IPF has been described as the strongest known risk factor for IPF. Consequently, there are sporadic and familial forms of IPF. When IPF is determined to be familial, it technically no longer is idiopathic, and therefore is referred to as familial pulmonary fibrosis.

Familial pulmonary fibrosis (familial PF) is defined as occurring in individuals with one or more relative, third degree or closer, with probable IIP while sporadic pulmonary fibrosis (sporadic PF) occurs in individuals with no family history of IIP. The majority of IPF cases are sporadic; familial cases account for approximately 3-20% of all IPF cases. A previous study that evaluated the clinical and histological characteristics of these two forms of IPF concluded that they were...
indistinguishable. That same study did find that as the number of family members affected by IPF increased, mean survival time for a familial PF patient decreased\(^{44}\). One study found in a small cohort of Newfoundland IPF cases that the mean age of onset of symptoms, diagnosis and death were significantly earlier in familial PF cases\(^{42}\). Familial PF follows an overall autosomal dominant inheritance pattern, however mutations in the genes associated with IPF are found in just a portion of familial PF patients and are also found in sporadic patients\(^{43}\).

**The Etiology of IPF**

The general disease development of IPF pathway is hypothesized in a directed acyclic graph (DAG) in Figure 2.

![Figure 2. DAG of Hypothesized IPF Disease Development Pathways](image)

*M: potential mediator; I: potential effect modifier

Ethnicity and family history both are directly related to the inheritance of the risk alleles of IPF. It is hypothesized that environmental exposures mediate the pathway between known risk loci and development of subclinical and clinical IPF. Environmental exposures cause damage to the
epithelium in the interstitial space of the lung and various risk alleles dictate whether that damage occurs, how that damage is repaired and if that repair is involved the development of fibrotic tissues. Therefore, it is hypothesized that environmental exposures are on the casual pathway of IPF. The inheritance of risk alleles doesn’t necessarily indicate that those alleles are expressed similarly. Histone modification, DNA methylation and post-transcriptional modification of the risk alleles could all impact the effect of the expression of that risk allele. Accordingly, gene regulation is hypothesized to act as a mediator between the risk alleles and disease development. Age seemingly modifies the relationship between risk alleles and disease because increased age would involve more exposures to particulates and more time for disease to develop into clinical phase. Finally, each different comorbidity presumably has a different relationship with IPF. It is hypothesized that some comorbidities share risk factors with IPF and may have a role in the casual pathway of IPF, while other comorbidities are only correlated with IPF due to the increased incidence of both diseases in older individuals and/or among smokers. Clearly, this hypothesized diagram greatly generalizes the variables involved in the casual pathway of IPF and future research will elucidate the actual relationships between these variables.

**Study Aims**

In this study we aim to evaluate the demographic and genetic characteristics of familial PF and sporadic PF cases. This study also aims to determine if genetic subtypes of IPF exist.

**Specific Aim 1**

The first aim of this study is to determine if smoking status, age or gender is differentially associated with familial PF and sporadic PF. *Hypothesis 1.1:* Smoking history and mean pack years differs by family history of pulmonary fibrosis. *Hypothesis 1.2:* Mean age at time of disease differs by family history of pulmonary fibrosis. *Hypothesis 1.3:* Gender differs by family history of pulmonary fibrosis.
Specific Aim 2

The second aim of this study is to determine if the association between genetic risk factors of IPF and disease differs in familial PF and sporadic PF. Hypothesis 2.1: Familial PF and sporadic PF cases have different risk allele frequencies for the 10 loci most predictive of IPF disease development (Table 2). Hypothesis 2.2: Smoking history interacts with the alleles of the 10 risk loci and familial or sporadic PF outcomes.

Specific Aim 3

The third aim of this study is to determine if genetic subtypes of IPF exist and if the risk loci of IPF interact with each other. In particular, we seek to evaluate the interaction between the MUC5B promoter risk variant (rs35705950) and the other risk loci. Hypothesis 3.1: Subgroups of IPF cases that carry different frequencies of the 10 risk loci exist. Hypothesis 3.2: The MUC5B promoter risk variant interacts with the other risk loci of IPF in cases vs. controls. Hypothesis 3.3: The MUC5B promoter risk variant interacts with the other risk loci of IPF in familial PF vs. sporadic PF cases.

Innovation

This study is, to our knowledge, the most high-powered study to date to evaluate familial and sporadic pulmonary fibrosis cases. This study would be the first to evaluate the 10 risk loci associated with IPF in familial PF and sporadic PF cases. This study also proposes a novel genetic clustering approach that could identify genetic subtypes of IPF. The characteristics of these subtypes may reveal potential gene-gene and gene-environment interactions involved in the pathogenesis of IPF. Cluster analysis has been successfully used to identify phenotypic subtypes of other lung diseases\(^{45,46}\). A clear understanding of the etiology of IPF, including subtypes of disease, is needed in order to develop effective prevention strategies for this severe chronic disease\(^{43}\). The findings of this study could justify future research into the association of IPF subtypes with detailed exposure histories and longitudinal clinical measures of IPF\(^7\).
Approach

This research project utilized a dataset of 4,389 IIP cases and controls that were genotyped for the Fingerlin et al. (2013) GWAS study.

Study Population

IIP cases who did not have an IPF diagnosis were excluded from this study population (Figure 3).

Figure 3. Study Population

Pulmonary fibrosis cases were classified as being familial if that case had one or more third-degree or closer relatives with probable or definite idiopathic interstitial pneumonia (IIP). Individuals with no family history of disease were classified as being sporadic. One case was randomly selected from a family that contained multiple cases to control for the effect of repeated genomic measures from the same family cluster. Study aims 1 and 2 were addressed using a case-only population of
familial and sporadic cases (n=1838) and aim 3 was addressed using a case-control study population (n=3856).

IPF cases and controls were recruited from a variety of study cohorts (Table 3). Study participants were recruited from 11 cohorts (COPDGene (n=1895), Duke University (n=323), Intermune IPF trial (n=580), the National Heart, Lung, and Blood Institute Lung Tissue Research Consortium (n=41), National Jewish Health IIP population (n=209), Penn State University (n=51), University of California San Francisco (n=72), United Kingdom-Royal Brompton (n=72), University of Pittsburg (n=304), University of Texas (n=110) and Vanderbilt University (n=236)). One study site only recruited familial cases, 6 study sites recruited only sporadic cases and 3 study sites recruited both familial and sporadic cases. The majority of controls came from the COPDGene study and the remainder came from a University of Pittsburg study. The COPDGene study population is comprised of smokers so the control population is not representative of the general population for smoking history45.

Table 3. Study Population

<table>
<thead>
<tr>
<th>Case Status (n)</th>
<th>Case Type (n)</th>
<th>COPDGene</th>
<th>Dubois</th>
<th>Duke</th>
<th>Intermune</th>
<th>LTRC</th>
<th>NIH</th>
<th>Pitt</th>
<th>UCSF</th>
<th>UK</th>
<th>UT</th>
<th>Vanderbilt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control (2018)</td>
<td></td>
<td>1895(94)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case (1838)</td>
<td>Familial PF (381)</td>
<td></td>
<td>154(40)</td>
<td></td>
<td>69(18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sporadic PF (1457)</td>
<td></td>
<td>51(4)</td>
<td>169(12)</td>
<td>580(40)</td>
<td>41(3)</td>
<td>140(10)</td>
<td>181(12)</td>
<td>35(2)</td>
<td>72(5)</td>
<td>188(13)</td>
<td></td>
</tr>
</tbody>
</table>

Study Variables

All study participants were surveyed at the time of blood and/or tissue collection. Demographic and exposure data available for all of the IPF cases included age at time of disease and gender. Smoking history (n=1116) and pack years (n=1015) were available for a majority of the IPF cases. All study participants self-reported as being non-Hispanic white. All study participants, cases and controls, were genotyped using techniques previously described in the online methods.
supplement to Fingerlin et al. (2013). The 10 risk loci (SNPs) found to be most predictive of IPF
disease development were the genetic risk factors evaluated throughout this study (Table 2). The
collection of genomic and demographic data in this study population was approved by the National
Jewish Health Institutional Review Board (IRB).

**Statistical Analysis**

Study aim 1 was addressed using univariate and regression analysis. The distribution of
gender and history of smoking in familial PF and sporadic PF cases was assessed using chi-square
and Fisher’s exact tests. The mean age and pack years of these two case types were assessed using
two-sample t-tests. Regression models were constructed to control for potential confounding
variables depending on the results of univariate analysis. A p-value cutoff for significance of 0.05
was used in these analyses.

Study aim 2 was addressed using logistic regression analysis. The odds of having a higher
number of each risk allele given a family history was modeled for each of the 10 risk alleles. Risk
estimates were reported as odds ratios with 95% confidence intervals. Confounding or precision
variables were included in these models. A p-value cutoff for significance of 0.005 was used in these
analyses to account for the 10 tests being done (=0.5/10). Interaction between smoking exposures
and the risk alleles was also assessed (Figure 4). The role of potential effect modification was
assessed by adding interaction terms to the regression models. In the case of significant (p<0.01) or
marginally significant (p<0.05) interaction terms, subgroup analyses were done to assess if
meaningful interaction existed.
Study aim 3 was addressed using both logistic regression models and cluster analysis approaches. Interaction terms of each risk loci and the $MUC5B$ promoter risk variant (rs35705950) were added to the regression models built for study aim 2 (Figure 5). In the case of significant ($p<0.01$) or marginally significant ($p<0.05$) interaction terms, subgroup analyses were done to assess if meaningful interaction existed.

Hierarchical and k-means clustering approaches using the 10 risk loci (SNPs), treated ordinarily (0,1,2), were used to attempt to determine if genotypic subgroups existed within the study population of cases and controls. The cluster analysis study population was split into equal-sized
training and validation samples. First a comprehensive clustering was implemented by clustering on all 10 SNPs. Second, a subgroup of the SNPs that explained 10% or more of the total variance were identified using factor analysis. The two clustering approaches were then applied to this smaller set of SNPs. Cluster stability was assessed in every approach using cluster stability measures including adjusted-r square and pseudo f statistic of each number of clusters. Any potentially successful clustering approaches were then tested for validity using a replication study with the validation sample.

**Study Limitations**

This study utilized secondary data that was not collected with the intention to address the aims of this study. Moreover, these data were collected by multiple cohorts and are therefore subject to several limitations. The smoking data were self-reported and therefore may be subject to differential recall bias or underreporting biases in those with a family history of disease as compared to those that do not. Due to the cross-sectional nature of this study, misclassification of family history was possible in sporadic cases and may have limited the study’s ability to draw sound associations. In an effort to reduce confounding and increase power, the study population consisted only of non-Hispanic white IPF cases. Therefore, the results of this study may not be generalizable to the IIP/IPF population at-large. By doing multiple tests of association between risk loci and family history power was lost and may have increased the probability of a type II error. This study proposed a cluster analysis approach using the 10 SNPs that are most strongly associated with IIP disease development when evaluated independently. However, there may be other loci that explain the differences between familial PF and sporadic PF. There also may other loci that interact to create a risk for IPF that have not yet been identified and therefore were not included in this study. These epistatic loci, if they exist, could potentially improve the stability and efficacy of any actual
genotypic clusters. Missing data and unmeasured confounding also limited this study in its ability to completely describe the characteristics of potential genotypic subtypes as data for known environmental risk factors, comorbidities and clinical measures of IPF were not collected for the majority of these study participants. Finally, clustering approaches are not commonly used on SNP data. Most clustering techniques perform less effectively with ordinal data. Therefore, the techniques proposed for study aim 3 may have failed to address hypothesis 3.1.
CHAPTER II

MANUSCRIPT

Introduction

Idiopathic pulmonary fibrosis (IPF) is the most common form of idiopathic interstitial pneumonia (IIP) and has a poor prognosis for affected individuals\(^6\). IPF is characterized by progressive scarring of the interstitial space of alveoli in the lung\(^40\). The majority of IPF cases experience respiratory failure within 3 to 5 years\(^1\). While the etiology of IPF is largely unknown, previous studies have identified risk factors and multiple genes associated with IPF suggesting a complex multifactorial pathogenesis\(^8,27\). Older age, male gender, environmental particulate exposure, smoking history and a family history of IPF are all associated with the development of IPF\(^8,24,40\). IPF is comorbid with many pulmonary diseases including sarcoidosis and non-pulmonary diseases including gastroesophageal reflux disease (GERD) and diabetes\(^6,8,20\). Genetic risk factors of IPF include variations in a mucus formation gene (\textit{MUC5B}), a surfactant protein gene (\textit{SFTPA2}), genes involved in telomere maintenance (\textit{TERT}, \textit{TERC}) and antigen presenting genes (\textit{HLA})\(^7,25,27,29\).

There are familial and sporadic forms of IPF. Studies have estimated that between 3% and 20% of IPF cases have a family history of disease, while the majority of IPF cases have no family history of disease\(^42-44\). Studies that have evaluated the clinical and genetic characteristics of familial pulmonary fibrosis (familial PF) and sporadic pulmonary fibrosis (sporadic PF) have had mixed results. One small cohort study found differences in the age of onset and death for familial PF and sporadic PF cases\(^42\). Another study found differences in gene expression profiles in familial interstitial pneumonia and sporadic IIP\(^49\). Other studies concluded that familial PF and sporadic PF were indistinguishable in terms of pathologic, radiologic and clinical characteristics\(^43,44\). Genetic studies have found genes associated with familial PF but to date no genetic risk factor differences between familial PF and sporadic PF have been described\(^7,28\). One lung tissue study did find gene
expression differences in familial and sporadic cases. Therefore, it has been hypothesized that common gene-exposure pathways exist in familial PF and sporadic PF yet at the same time disease initiation and progression may be influenced by other factors.

A clear understanding of the pathogenesis of IPF is needed in order to develop effective treatments and prevention strategies for this severe chronic disease. A recent genome-wide association study (GWAS) and subsequent imputation study identified 10 single nucleotide polymorphisms (SNPs) that were strongly associated with IIP/IPF. Some of these risk loci are located in or near genes previously shown to be associated with disease, while others were novel and were confirmed to be associated with disease in replication and expression studies. This study aims to use a high-powered study population of IPF cases to evaluate if the association between these 10 risk loci and disease differs in familial PF and sporadic PF cases and if those risk loci interact with the strongest known risk variant for IPF (rs35705950). Furthermore, this study aims to determine if smoking history varies by family history of disease.

**Methods**

This case-case study used cross-sectional genomic data comprised of IPF cases compiled by our research group.

**Study Population**

A total of 1838 study participants with familial or sporadic pulmonary fibrosis were included in this study. Diagnoses of pulmonary fibrosis were made using high-resolution computed tomography (HRCT) and the pulmonary fibrosis diagnostic criteria developed by the American Thoracic Society. Pulmonary fibrosis cases were classified as being familial if that case had one or more third-degree or closer relative with probable or definite idiopathic interstitial pneumonia (IIP). Individuals with no known family history of disease were classified as being sporadic. In the instance of multiple study participants from the same family, one member from each family cluster was
randomly selected for inclusion in this study. Study participants were recruited from 10 cohorts (Duke University (n=323), Intermune IPF trial (n=580), the National Heart, Lung, and Blood Institute Lung Tissue Research Consortium (n=41), National Jewish Health IIP population (n=209), Penn State University (n=51), University of California San Francisco (n=72), United Kingdom-Royal Brompton (n=72), University of Pittsburg (n=181), University of Texas (n=110) and Vanderbilt University (n=236)). One study site only recruited familial cases, 6 study sites recruited only sporadic cases and 3 study sites recruited both familial and sporadic cases (Table 4).

Table 4. Case-Case Study Population

<table>
<thead>
<tr>
<th>Case Type (n)</th>
<th>Dubois</th>
<th>Duke</th>
<th>Intermune</th>
<th>LTRC</th>
<th>NH</th>
<th>Pitt</th>
<th>UCSF</th>
<th>UK</th>
<th>UT</th>
<th>Vanderbilt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familial PF (381)</td>
<td>154(40)</td>
<td>69(18)</td>
<td>110(29)</td>
<td>48(13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sporadic PF (1457)</td>
<td>51(4)</td>
<td>169(12)</td>
<td>580(40)</td>
<td>41(3)</td>
<td>140(10)</td>
<td>181(12)</td>
<td>35(2)</td>
<td>72(5)</td>
<td>188(13)</td>
<td></td>
</tr>
</tbody>
</table>

All study participants self-reported as being non-Hispanic white. The collection of genomic and demographic data in this study population was approved at each clinical collection site and the collection of these into one study was approved by the National Jewish Health Institutional Review Board (IRB)\(^{27}\).

**Study Design**

This study utilized a IPF case-only population from the larger IIP case and control study population. Cross-sectional demographic data previously reported to be associated with IPF including age, gender, smoking history and family history of pulmonary fibrosis were collected when blood or tissue samples were collected for genotyping. The DNA preparation and genotyping done on these samples has been previously reported in detail in the online methods of Fingerlin et al. (2013). The demographic data were self-reported and were collected using different questionnaires across the ten cohorts. In some questionnaires, individuals reported if they were current, former or
never smokers and on other questionnaires they reported if they were ever or never smokers. Individuals that reported having never smoked were recorded as having 0 pack years. Individuals with incomplete smoking history data were excluded from smoking analyses.

**Data Analysis**

All data merging and analysis was conducted using SAS 9.4. The primary outcome of interest, familial PF and sporadic PF, was binary. The primary exposure of interest, the 10 risk loci (SNPs) of pulmonary fibrosis, were treated ordinally for the number of risk alleles held (0, 1 or 2) by each individual as this was the genetic model for each SNP identified in the GWAS.

Continuous baseline characteristics of familial and sporadic pulmonary fibrosis cases included the age of the study participant and the number pack years at the time of DNA collection. Categorical baseline characteristics of familial and sporadic pulmonary fibrosis cases including study site, gender and smoking history were assessed using chi-square tests. The distribution of continuous covariate data was assessed using univariate analysis. The mean age and pack years of familial and sporadic cases were compared using unpaired two-sample t-tests with Satterthwaite approximation. Mixed models were built to account for variance across the multiple study sites that collected smoking data.

Unconditional multivariable logistic regression models were constructed to test the association between family history of disease and each of the 10 individual risk loci. Categorical and continuous variables found to be associated with family history of disease were included in the regression models as precision variables. Since 10 tests of association were conducted, an *a priori* p-value cutoff for significance of 0.005 was used (=0.05/10). Unpaired two-sample t-tests with Satterthwaite approximation were used to compare the total mean risk allele frequency in familial and sporadic cases for all 10 loci.
Gene-gene interaction between a risk variant SNP (rs35705950) in the promoter region of *MUC5B* and the other 9 risk loci was assessed by adding rs35705950 as an interaction term to the regression models. In the case of significant (p<0.01) or marginally significant (p<0.05) interaction terms, subgroup analyses were done to assess if meaningful interaction existed. Mean number of rs35705950 risk alleles at the different levels of risk alleles for that loci were evaluated to determine if meaningful interaction existed between the risk allele and rs35705950.

Gene-tobacco smoke exposure interaction was evaluated by adding smoking history as an interaction term with each risk loci to the regression models. In the case of significant (p<0.01) or marginally significant (p<0.05) interaction terms, subgroup analyses were done to assess if meaningful interaction existed. Mean pack years and sample size at the different levels of the risk alleles for that risk loci were then evaluated to determine if meaningful interaction existed between the risk allele and smoking history.

**Results**

A total of 1838 pulmonary fibrosis cases were genotyped and met the inclusion criteria of this study, with 381 (21%) having a family history of pulmonary fibrosis and 1457 (79%) having no family history of disease (Table 5). The frequency of females was higher in familial pulmonary fibrosis cases (35%) than in sporadic pulmonary fibrosis cases (28%) (p=0.003). The difference in gender was most likely a result of how study participant recruitment was conducted in the 10 different cohorts. As a result, gender was included as a covariate in all models. Age did not significantly vary between familial and sporadic cases (p=0.53).
Table 5. Familial and Sporadic Pulmonary Fibrosis Cases

<table>
<thead>
<tr>
<th></th>
<th>Familial PF</th>
<th>Sporadic PF</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>381</td>
<td>1457</td>
<td></td>
</tr>
<tr>
<td>Female n(%)</td>
<td>135(35)</td>
<td>401(28)</td>
<td>0.003*</td>
</tr>
<tr>
<td>Mean Age</td>
<td>66.1</td>
<td>65.7</td>
<td>0.532*</td>
</tr>
<tr>
<td>Mean Pack-Years</td>
<td>17.0</td>
<td>29.6</td>
<td>&lt;0.0001*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Smoking Status</th>
<th>Current</th>
<th>Former</th>
<th>Never</th>
<th>Ever</th>
<th>Never</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>n(%)</td>
<td>7(3)</td>
<td>159(63)</td>
<td>88(35)</td>
<td>166(65)</td>
<td>88(35)</td>
<td>0.144*</td>
</tr>
<tr>
<td></td>
<td>32(3.7)</td>
<td>585(68)</td>
<td>245(28)</td>
<td>625(72)</td>
<td>245(28)</td>
<td></td>
</tr>
</tbody>
</table>

*p-value from two-sample t-test with Satterthwaite approximation
^p-value from chi-Square test

Smoking

The frequency of ever smokers was marginally higher in sporadic cases (p=0.046) when smoking history was categorized as being either an ever smoker or a never smoker (Table 5). There was no significant difference in smoking history between familial and sporadic cases when smoking history was categorized as current, former or never (p=0.144). In familial and sporadic cases from all study sites, mean pack years was higher in sporadic cases (29.6) than in familial cases (17.0). The difference in mean pack years was significant when the variance between the multiple study sites were accounted for in a mixed model when adjusted for gender (p <0.0001). The National Jewish and Vanderbilt study sites each collected smoking data for both familial PF and sporadic PF cases. In both study sites mean pack years was higher in sporadic cases (Table 6). This difference was still highly significant when gender and age were controlled for in both the National Jewish (p=0.004) and Vanderbilt cohorts (p=0.002). The same trend was also seen in these two cohorts in familial and sporadic cases that were current or former smokers.
Table 6. Mean Pack Years in National Jewish and Vanderbilt Cohorts

<table>
<thead>
<tr>
<th>Risk Loci</th>
<th>National Jewish</th>
<th>Vanderbilt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ever and Never Smokers</td>
<td>Familial PF</td>
<td>Sporadic PF</td>
</tr>
<tr>
<td>Case Type</td>
<td>n</td>
<td>pack years</td>
</tr>
<tr>
<td>Ever Smokers</td>
<td>Familial PF</td>
<td>20</td>
</tr>
<tr>
<td>Sporadic PF</td>
<td>89</td>
<td>69.8</td>
</tr>
</tbody>
</table>

*p-value from logistic regression, adjusted for age and gender.

Risk Loci

The frequency of the risk alleles for loci associated with pulmonary fibrosis did not differ in familial and sporadic pulmonary fibrosis cases when evaluated individually and adjusted for gender (Table 7). The odds of a familial case carrying a higher number of risk alleles for a locus and a sporadic case carrying a higher number of risk alleles were the same for all of the 10 risk loci. The odds of the risk loci in familial and sporadic cases vs. controls when adjusted for age and gender are illustrated in Figure 6.

Table 7. Risk Allele Frequency in Familial and Sporadic Pulmonary Fibrosis Cases

<table>
<thead>
<tr>
<th>Risk SNP</th>
<th>Risk Allele</th>
<th>Locus</th>
<th>Familial PF Risk Allele Count</th>
<th>Sporadic PF Risk Allele Count</th>
<th>OR*</th>
<th>95% CI</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12696304</td>
<td>G</td>
<td>3q26</td>
<td>176</td>
<td>166</td>
<td>39</td>
<td>381</td>
<td>695</td>
</tr>
<tr>
<td>rs2609261</td>
<td>T</td>
<td>4q22</td>
<td>215</td>
<td>148</td>
<td>18</td>
<td>381</td>
<td>807</td>
</tr>
<tr>
<td>rs10069690</td>
<td>C</td>
<td>5p15</td>
<td>10</td>
<td>106</td>
<td>255</td>
<td>371</td>
<td>48</td>
</tr>
<tr>
<td>rs7887</td>
<td>C</td>
<td>6p21</td>
<td>35</td>
<td>153</td>
<td>171</td>
<td>359</td>
<td>134</td>
</tr>
<tr>
<td>rs2076295</td>
<td>G</td>
<td>6p24</td>
<td>83</td>
<td>161</td>
<td>113</td>
<td>357</td>
<td>311</td>
</tr>
<tr>
<td>rs4727443</td>
<td>A</td>
<td>7q22</td>
<td>101</td>
<td>188</td>
<td>66</td>
<td>355</td>
<td>419</td>
</tr>
<tr>
<td>rs9325507</td>
<td>T</td>
<td>10q24</td>
<td>76</td>
<td>175</td>
<td>119</td>
<td>370</td>
<td>275</td>
</tr>
<tr>
<td>rs35709590</td>
<td>T</td>
<td>11p15</td>
<td>123</td>
<td>199</td>
<td>30</td>
<td>352</td>
<td>509</td>
</tr>
<tr>
<td>rs1981997</td>
<td>G</td>
<td>17q21</td>
<td>11</td>
<td>101</td>
<td>245</td>
<td>357</td>
<td>42</td>
</tr>
<tr>
<td>rs12610495</td>
<td>G</td>
<td>19p13</td>
<td>163</td>
<td>145</td>
<td>42</td>
<td>350</td>
<td>572</td>
</tr>
</tbody>
</table>

*odds ratio and p-value from logistic regression, adjusted for gender.
Furthermore, in the cases in which allele data were present for all 10 risk loci, no difference in mean risk allele frequency was found with familial cases carrying 51.4% of the risk alleles on average and sporadic cases carrying 51.6% of the risk alleles (Table 8).

Table 8. Mean Risk Variant Frequency

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean Total Risk Variants</th>
<th>Mean Risk Allele Frequency</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familial PF</td>
<td>321</td>
<td>10.3</td>
<td>0.514</td>
<td>0.661</td>
</tr>
<tr>
<td>Sporadic PF</td>
<td>1257</td>
<td>10.3</td>
<td>0.516</td>
<td></td>
</tr>
</tbody>
</table>

* p-value from logistic regression, adjusted for gender.
**Interaction with MUC5B Promoter SNP**

The relationship between 8 risk loci and family history of pulmonary fibrosis was not modified by the number of risk alleles held for rs35705950 (Table 9).

<table>
<thead>
<tr>
<th>SNP</th>
<th>Risk Allele</th>
<th>p-value^</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10069690</td>
<td>C</td>
<td>0.151</td>
</tr>
<tr>
<td>rs12610495</td>
<td>G</td>
<td>0.410</td>
</tr>
<tr>
<td>rs12696304</td>
<td>G</td>
<td>0.861</td>
</tr>
<tr>
<td>rs1981997</td>
<td>G</td>
<td>0.271</td>
</tr>
<tr>
<td>rs2076295</td>
<td>G</td>
<td>0.941</td>
</tr>
<tr>
<td>rs2609261</td>
<td>T</td>
<td>0.432</td>
</tr>
<tr>
<td>rs4727443</td>
<td>A</td>
<td>0.359</td>
</tr>
<tr>
<td>rs7887</td>
<td>C</td>
<td>0.862</td>
</tr>
<tr>
<td>rs9325507</td>
<td>T</td>
<td>0.646</td>
</tr>
</tbody>
</table>

^ p-value of interaction term when adjusted for gender

The relationship between rs10069690 and family history of pulmonary fibrosis is potentially modified by the number of risk alleles held for a SNP (rs35705950) in the MUC5B promoter region (Table 10). The mean number of risk alleles for rs35705950 decreased as the number of risk alleles for rs10069690 increased in familial cases but not in sporadic cases. In particular, the mean number of risk alleles for rs35705950 was highest in individuals that held no risk alleles for the rs10069690 risk variant.

<table>
<thead>
<tr>
<th>Mean number risk alleles rs35705950</th>
<th>rs10069690</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>

^ p-value of interaction term of the two SNPs; from logistic regression, adjusted for gender.
**Interaction of Risk Loci and Smoking**

One of the 10 risk variants showed potential interaction with smoking history. Familial PF cases with a higher number of risk alleles for rs4727443 had higher mean pack years while there was no trend in sporadic PF cases (Table 11). Another SNP, rs1981997 did vary by smoking history in familial PF and sporadic PF cases, however the proportion of familial PF cases with 0 risk alleles for rs1981997 and smoking history data was very small (n=5). The remaining 8 risk loci did not show significant interaction with smoking history.

<table>
<thead>
<tr>
<th>Risk alleles</th>
<th>Mean Pack Years</th>
<th>p-value^</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Familial PF</td>
<td>Sporadic PF</td>
</tr>
<tr>
<td>0</td>
<td>14.7</td>
<td>28.4</td>
</tr>
<tr>
<td>rs4727443</td>
<td>16.9</td>
<td>29.9</td>
</tr>
<tr>
<td>2</td>
<td>21.8</td>
<td>28.3</td>
</tr>
</tbody>
</table>

^ p-value of interaction term of SNP and smoking history; from logistic regression, adjusted for gender.

**Discussion**

In this study, it has been demonstrated that there was no difference in risk allele frequency for the 10 loci that are strongly associated with the development of IPF in a large population of familial PF and sporadic PF non-Hispanic white (NHW) cases. Familial PF and sporadic PF cases also had very similar total mean allele frequencies for these 10 risk loci. The results of this study demonstrate that a gene-gene interaction difference between familial PF and sporadic PF cases may exist. Familial cases with no risk allele for rs10069690, a SNP found in the intron region of the telomere length maintenance gene TERT, had more risk alleles for rs35705950, a SNP found in the promoter region of a mucus formation gene MUC5B. Inheritance of the risk allele (T) for MUC5B is a strong risk factor for the development of IPF, but is also associated with longer mean survival times.
in IPF cases. The IPF risk variant frequency of rs10069690 is very high in the general population, with 93.6% of controls in the original GWAS study having at least one risk allele for rs10069690. Consequently, the sample size of familial cases with no risk variants for rs10069690 was small (n=10) and therefore this apparent potential modification with rs35705950 should be interpreted cautiously.

This study found that mean pack years was significantly higher in sporadic PF cases than familial PF cases. To account for potential information biases due to differences in pack year data collection across the 10 study cohorts, sub-group analyses were conducted on the two cohorts that collected pack year data for both familial and sporadic cases using the same questionnaires. Sporadic cases had higher mean pack years in these two cohorts when controlling for differences in gender and age. To account for potential underreporting biases in familial cases, mean pack years in study participants was also evaluated in cases that reported that they had a history of smoking. Mean pack years was also higher in this group of ever smokers when controlling for differences in gender and age. Finally, the potential for gene-smoking interactions was evaluated. It was found that potential interaction existed between smoking history and one risk SNP, rs4727443. Familial PF cases with more risk alleles for rs4727443 also had higher pack years while this trend was not seen in sporadic PF cases. The biological role of rs4727443 is not known and therefore we cannot speculate at this time about the potential mechanism associated with the differences. These findings suggest that sporadic PF cases may require greater exposure to tobacco smoke than familial PF cases and that interactions between genetic risk factors and tobacco smoke exposure may differ between familial and sporadic PF cases.

The generalizability of the results of this study are limited by the use of a NHW only study population. The use of a NHW population did control for genomic differences that may have confounded the relationship between genetic risk variants and IPF development. The cross-sectional
nature of the data could potentially limit the validity of this study. Misclassification of some familial cases as being sporadic is likely since no follow-up data with these cases were available. In addition, not all of the cohorts used extensive questionnaires, clinical evaluation or other approaches to rigorously determine whether or not a family history existed. The use of a cross-sectional study population did allow for a large sample size and subgroup analyses. A previous study reported that mean age of diagnosis and death of IPF was earlier in familial PF cases\textsuperscript{42}. Therefore, survivor bias in familial PF cases may be present and may be driving the higher pack years reported in sporadic PF cases. This study is also limited in that only 10 loci were evaluated in these genetic analyses. Genetic differences may exist between these two forms of pulmonary fibrosis in other loci across the human genome and those loci may interact with each other. In particular, since the larger study that was used to identify the 10 loci consisted of over 70% sporadic cases, we cannot rule out that large studies conducted with primarily familial cases would identify loci more influential in that group vs. sporadic cases. Finally, data for most known environmental risk factors and comorbidities of IPF were not available for these study participants. The validity of the results reported in this study may be limited by unmeasured confounding in these pulmonary fibrosis cases.

In conclusion, these findings suggest that familial and sporadic PF share the same genetic risk factors but that differences in gene-gene interaction and gene-environmental exposures may exist. In particular, sporadic PF cases may require higher exposure to tobacco smoke for disease development to occur. IPF is a complex chronic disease that involves an interaction between genetic risk factors and dose-dependent particulate exposures\textsuperscript{7}. Future study, with more detailed exposure data, may be able to better elucidate gene-environment differences between familial and sporadic pulmonary fibrosis.
CHAPTER III

GENOTYPE ANALYSIS

Introduction

Genome-wide association studies (GWAS) identify single nucleotide polymorphisms (SNPs) that are associated with disease. An individual SNP is a locus in the human genome in which variation exists at a single nucleotide position\(^{32}\). GWAS typically include replication studies to confirm that the risk alleles they generated are associated with disease and not a result of chance\(^{33}\).

The risk alleles identified in a GWAS may be located within a gene that is directly related to the pathogenesis of a disease, or they may be located in a regulatory region of a gene or they may be associated with a nearby region of the genome that is actually involved with disease pathogenesis.

Many late-onset chronic diseases, including idiopathic pulmonary fibrosis (IPF), are polygenic\(^{50}\). The risk alleles identified in a GWAS may interact with other risk alleles of disease as part of disease development pathways\(^{48}\). Furthermore, an interaction between genotypes and lifelong environmental exposures are part of the pathogenesis of many chronic diseases\(^{50}\).

Many methods have been used to evaluate potential gene-gene interactions in SNP data, however research into the efficacy of these methods is limited\(^{48}\). The interaction between two genetic risk factors can be evaluated through stratified analysis and pairwise tests for effect modification. This study aims to determine if the most strongly associated risk locus of IPF (rs35705950) in the promoter region of \textit{MUC5B} interacts with the other 9 known risk loci for IPF.

This study also aims to determine if genotype subgroups of IPF exist using cluster analysis techniques. In exploratory analysis, cluster analysis techniques have been used to identify genotype clusters within larger datasets to both identify family clusters and to identify potential subgroups of cases that share alleles associated with disease\(^{48,51}\).
Methods

A study population of IPF cases (n=1838) and controls (n=2018) was used in these analyses (Table 12). The study population was randomly allocated into training and validation samples for cluster analysis.

Table 12. Cluster Analysis Study Population

<table>
<thead>
<tr>
<th>Sample</th>
<th>n</th>
<th>Familial PF</th>
<th>Sporadic PF</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1928</td>
<td>188(10)</td>
<td>729(38)</td>
<td>1011(52)</td>
</tr>
<tr>
<td>Validation</td>
<td>1928</td>
<td>193(10)</td>
<td>728(38)</td>
<td>1007(52)</td>
</tr>
</tbody>
</table>

Interaction Between MUC5B and Other Risk Loci

Pairwise interaction terms of 9 risk loci of IPF and the MUC5B promoter risk variant (rs35705950) were added to unconditional logistic regression models. These models were adjusted for age and gender. The role of potential effect modification was assessed by adding interaction terms to the regression models. In the case of significant (p<0.01) or marginally significant (p<0.05) interaction terms, subgroup analyses were done to assess if meaningful interaction existed.

Cluster Analysis

Hierarchical and k-means clustering techniques were used to attempt to determine if genotypic subgroups existed within the study population of cases and controls. The structure of the input data was evaluated by producing a correlation matrix between all 10 SNPs.

First, a comprehensive set clustering approach was implemented by clustering on all 10 SNPs. Second, a limited set clustering approach was implemented. In this approach clustering was conducted on a subgroup of the SNPs that each explain 10% or more of the total variance within the SNPs. These SNPs were identified using factor analysis (SAS 9.4, PROC FACTOR command) to reduce dimensionality in the comprehensive set. Observations with incomplete genomic data for each set...
of SNPs were excluded. The stability (compactness) of the clusters produced in both the comprehensive set and limited set approaches was assessed using the pseudo f statistic (ratio of between cluster sum of squares and within cluster sum of squares) and r-squared (proportion of the variation in the SNPs explained by the clustering of the study participants). The characteristics of the most stable cluster solution from each approach was analyzed using chi-square tests and one-way ANOVA. The frequency of cases and controls, familial PF cases and sporadic PF cases, gender distribution and smoking history in the different clusters was assessed. The mean age and pack years of each cluster was also evaluated. Any potentially successful clustering approaches were then tested for validity using a replication study with the validation sample.

Hierarchical Clustering

Hierarchical clustering was conducted using the PROC CLUSTER command (Ward’s Method). For hierarchical clustering, the SNPs were treated ordinally (0,1,2) for the number of risk alleles present in an individual. The ordinal data were transformed into a Euclidian distance matrix using the PROC DISTANCE command. In Ward’s clustering method, observations are clustered together one at a time to minimize the increase of within cluster variance. This process repeats itself until all observations are placed into a single cluster. The hierarchical clustering solutions were visualized in tree diagrams using the PROC TREE command.

K-means Clustering

K-means clustering was conducted using the PROC FASTCLUS command. For k-means clustering, the SNPs were treated as an interval scale (0,1,2) for the number of risk alleles present in an individual. K-means clustering is a method in which observations are placed into k number (k=2,3...n) of clusters in a multi-dimensional space. Observations are assigned to the cluster that is closest to their mean values for the variables used for clustering. With the addition of the new
observation the mean value (the centroid) of that cluster is recalculated. This process is repeated until all observations are assigned to k-number of clusters.

Results

Gene-gene interaction was assessed in 1838 IPF cases and 2018 controls.

Interaction Between MUC5B and Other Risk Loci

The relationship between two independent risk loci, rs2609261 and rs4727443, and IPF was potentially modified by the number of risk alleles held for rs35705950 (Table 13). For rs2609261, the mean number of risk alleles of rs35705950 decreased as the number of risk alleles for rs2609261 increased in cases, while in controls it increased (interaction p-value=0.05). For rs4727443, the mean number of risk alleles for rs35705950 increased as the number of risk alleles for rs4727443 increased in cases, while in controls there was no trend between the two 2 risk loci (interaction p-value=0.02). The remaining 7 risk loci had no significant interaction with rs35705950 (interaction p-value>0.11).

Table 13. Interaction with MUC5B Promoter Risk Variant

<table>
<thead>
<tr>
<th>SNP</th>
<th>risk alleles</th>
<th>Mean Number Risk Alleles rs35705950</th>
<th>p-value^</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>IPF cases</td>
<td>Controls</td>
</tr>
<tr>
<td>rs2609261</td>
<td>0</td>
<td>0.73</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.71</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.58</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0.66</td>
</tr>
<tr>
<td>rs4727443</td>
<td>1</td>
<td>0.72</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.77</td>
<td>0.20</td>
</tr>
</tbody>
</table>

^p-value of interaction term; logistic regression, adjusted for age and gender
Cluster Analysis

In study participants with data for all 10 SNPs, mean risk allele frequency was significantly higher (two-sample t-test p-value <0.0001) in IPF cases as compared to controls (Table 14). Controls, on average, carried 43.8% of the risk alleles and IPF cases carried 51.5% of the risk alleles on average.

Table 14. Risk Allele Frequencies in IPF Cases and Controls

<table>
<thead>
<tr>
<th></th>
<th>Mean Total Risk Variants</th>
<th>Mean Risk Allele Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familial PF</td>
<td>321</td>
<td>10.3</td>
</tr>
<tr>
<td>Sporadic PF</td>
<td>1257</td>
<td>8.8</td>
</tr>
<tr>
<td>Control</td>
<td>1700</td>
<td></td>
</tr>
</tbody>
</table>

Correlation between each of the 10 risk loci was very low. Most risk alleles had a Pearson Correlation Coefficient less than 0.05 and no pairs of SNPs had a coefficient greater than 0.10 (Figure 7).

Figure 7. Correlation Between 10 Risk Loci of IPF
Hierarchical Clustering

The comprehensive set of all 10 SNPs did not produce stable clusters using Hierarchical clustering. It was estimated that 50% of the variance was explained by the clusters (r-square) at 32 number of clusters. The results of this approach are illustrated in Figure 8. For reference, the y-axis contains the 917 IPF cases and 1011 controls found in the training sample that had the comprehensive set of SNP data.

The limited set of the SNPs that were estimated to explain 10% or more of the total variance found in the comprehensive set of SNPs were: rs35705950, rs12610495, rs2076295 and rs7887. The selected clustering solution with this set of 4 SNPs was n=3 clusters based on stability measures (Figure 9). For reference, the y-axis contains the 795 IPF cases and 854 controls found in the training sample that had the limited set of SNP data. Only 38% of the total variance in the observations was estimated to be explained by the 3 clusters. For the training sample, clusters 1 and 2 contained 80% of the controls cluster 3 contained 63% of the IPF cases. This clustering approach placed all observations containing risk alleles for rs35705950 into the same cluster, cluster 3. As a result, the
demographic and smoking characteristics the clusters are driven by differences in cases vs. controls.

The characteristics of these three clusters for both the training and validation samples are described in Table 15.

![Limited Set Hierarchical Cluster Solution of Training Sample]

Figure 9. Limited Set Hierarchical Cluster Solution of Training Sample

Table 15. Characteristics of Limited Set Hierarchical Cluster Solution

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Training Sample</th>
<th>Validation Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>1</td>
</tr>
<tr>
<td>Case Status</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Familial PF</td>
<td>12.5%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Sporadic PF</td>
<td>14.7%</td>
<td>23.3%</td>
</tr>
<tr>
<td>Control</td>
<td>42.7%</td>
<td>37.7%</td>
</tr>
<tr>
<td>Mean Number Risk Alleles</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>rs35705950</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>rs12610495</td>
<td>0.0</td>
<td>1.2</td>
</tr>
<tr>
<td>rs2076295</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>rs7887</td>
<td>1.3</td>
<td>1.3</td>
</tr>
<tr>
<td>Mean Age</td>
<td>60.7</td>
<td>60.4</td>
</tr>
<tr>
<td>Mean Pack-Years</td>
<td>37.8</td>
<td>37.8</td>
</tr>
<tr>
<td>Smoking Status</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Ever</td>
<td>73.1%</td>
<td>78.2%</td>
</tr>
<tr>
<td>Never</td>
<td>26.9%</td>
<td>21.8%</td>
</tr>
</tbody>
</table>
K-means Clustering

The most stable cluster solution using K-means was k=3 number of clusters using the comprehensive set of 10 SNPs. However, the r-square value of this solution was 0.19, meaning that less than 19% of the variance found with the IPF and control observations were explained by the 3 cluster solution. Furthermore, in this 3 cluster solution using the comprehensive set, controls were placed into all 3 clusters at similar frequencies (36%, 41%, 24%).

The most stable cluster solution using K-means was k=5 number of clusters using the limited set of 4 SNPs. The r-square value of this solution was 0.56, meaning that 56% of the variance found with the IPF and control observations were explained by the 5-cluster solution. However, a negative value of another measure of cluster compactness, cubic clustering criterion (CCC), indicated that the observations did not fit this clustering solution well. It was determined that k-means analysis produced no reliable clustering solution and no validation study was done.

Discussion

This study demonstrated potential gene-gene interaction in IPF cases. The number of risk alleles held for a SNP found in the promoter region of MUC5B (rs35705950) appears to have a relationship with the number of risk alleles held with two other risk loci (rs2609261 and rs4727443) in IPF cases but not in controls. MUC5B produces a mucus gel formation protein that has a role in epithelium damage repair following injury and is overly expressed in IPF cases. The risk variant of rs2609261 is located in the intron region of FAM13A, a gene involved in signaling and previously associated with COPD. The biological role of rs4727443 is unknown. There were no pairwise interactions between rs35705950 and the other 6 risk variants of IPF.

This study also demonstrated the challenges of evaluating gene-gene interactions using a larger set of genetic risk factors. The binary (presence or absence of a risk allele) or ordinal (0, 1 or 2 copies of a risk allele) nature of SNP data pose challenges when using cluster analysis techniques.
Ultimately, it was determined that the two clustering methods using both a comprehensive and a limited set of risk loci failed to produce stable and meaningful clusters. The 10 risk alleles used in this study are common in both cases and controls and as a result controls were not separated from cases using cluster analysis. Moreover, it was found that the 10 risk alleles had very low correlation with one another. This is to be expected since these alleles are found on different chromosomes and are inherited independently of one another. The low correlation between these explanatory variables also limits the performance of cluster methodology. Even if this study produced stable clusters, no clinical measures of IPF data were available for these IPF cases and controls and therefore the characteristics of the potential subgroups of IPF could not have been completely evaluated.

In conclusion, when viewed in pairs some potential gene-gene interaction was found in IPF cases. However, this study failed to determine if genotypic subgroups of IPF exist using known risk factors of IPF disease development.
CHAPTER IV

CONCLUSION

Determining the etiology of multifactorial chronic diseases is a major public health challenge. Many chronic diseases, including idiopathic pulmonary fibrosis, include lifelong interactions between genetic risk factors and dose-dependent exposures. Many chronic diseases, including IPF, most likely involve multiple gene-exposure pathways that can lead to disease development and progression.

Genome-wide association studies (GWAS) can identify novel genetic risk factors for disease but extensive subsequent study is necessary to determine the role of those risk factors in disease development pathways. Necessary subsequent study includes evaluating the epigenetic regulation of those genetic risk factors, evaluating gene-gene interactions and evaluating gene-environment interactions. Costly and time-consuming prospective study may be necessary to collect reliable and valid exposures data. Animal studies can further refine the understanding of these complex pathways.

The main goal of this research was to build on the findings of the Fingerlin et al. (2013) GWAS study through exploratory analysis and tests of association. It was shown in this study that two potential forms of IPF, familial and sporadic, do not differ for 10 highly significant risk factors of IPF. Looking at these 10 risk factors individually supports the hypothesis that familial and sporadic pulmonary fibrosis are not distinct diseases. However, marginal interaction between two risk factors, rs35705950 and rs10069690, as well as highly significant differences in pack-years of smoking may point to the potential that sporadic and pulmonary fibrosis are distinct forms of IPF.

This study demonstrated significant interaction between rs35705950 and two other risk loci in IPF cases. This suggests that the protein products of some of these risk alleles may influence the
phenotypes associated with other risk alleles in subclinical and clinical IPF. This study also
demonstrated the challenges in evaluating gene-gene interaction of multiple risk alleles. Single
nucleotide polymorphism (SNP) data from unlinked sections of the human genome can be treated
as binary, ordinal or interval data. In general, cluster analysis methods perform best with continuous
and somewhat correlated data. Cluster analysis performs less effectively with variables with very
low correlation and with collinear variables. As a result, cluster analysis methods do not perform
well with SNP data. Bioinformatics has and will continue to improve our ability to evaluate the
interaction between multiple genetic risk factors. The findings of this study suggest that, at present,
clustering of chronic disease cases might be best done with continuous phenotypic and
demographic data.
REFERENCES


