Dissecting the regulatory role for a MUC5B polymorphism involved in Idiopathic Pulmonary Fibrosis

Amaranta Ú. Armesto Jiménez

Thesis for the degree of Philosophiae Doctor

Supervisor and advisor:
Magnús Karl Magnússon

Doctoral committee:
Thorarinn Gudjonsson
Erna Magnúsdottir
Eiríkur Steingrímsson
Gunnar Gudmundsson

August 2018

UNIVERSITY OF ICELAND
SCHOOL OF HEALTH SCIENCES
FACULTY OF MEDICINE

Hlutverk erfðabreytileika í stjórnsvæði MUC5B gensins og áhrif hans á lungnatrefjun

Amaranta Ú. Armesto Jiménez

Ritgerð til doktorsgráðu

Umsjónarkennari og leiðbeinandi:
Magnús Karl Magnússon

Doktorsnefnd:
Thorarinn Gudjonsson
Erna Magnúsdottir
Eiríkur Steingrímsson
Gunnar Gudmundsson

Ágúst 2018

UNIVERSITY OF ICELAND
SCHOOL OF HEALTH SCIENCES
FACULTY OF MEDICINE

ISBN 978-9935-9421-1-1

Printing by Haskolaprent ehf.

Reykjavik, Iceland 2018

Ágrip

Lungnatrefjun af óþekktum uppruna (IPF) er alvarlegur og ólæknandi sjúkdómur. Lifun eftir greiningu er um 3-5 ár og eina læknandi meðferðin er lungnaígræðsla, sem ekki allir þola. Nýverið hafa komið fram lyf á borð við pirferidone og nintedanib sem hafa náð að hægja á framgangi sjúkdómsins með því að hamla boðferla sem leiða til trefjunar. Þó er það þannig að IPF greinist oftast seint í sjúkdómsferlinu þannig að líkur á læknandi meðferð eru minni. Því er afar mikilvægt að uppgötva nýja áhættuþætti sjúkdómsins sem hægt er að nota til greiningar fyrr í ferlinu. Nýleg erfðamengis – tengslagreining á IPF sjúklingum samanborið við heilbrigða einstaklinga leiddi í ljós að einbasabreytileiki (G-T) (rs35705950) í stýrisvæði MUC5B gensins var til staðar í 45% IPF sjúklinga samanborið við 11% heilbrigðra. Þessi breytileiki er í dag stærsti þekkti áhættuþáttur fyrir IPF. G-T breytingin hefur einnig verið tengd við snemkomna trefjun í lungum, sem bendir til tengsla stökkbreytinga í MUC5B og lungnatrefjunar. Fram til þessa hefur ekki verið ljóst hvernig þessi einbasabreytileiki hefur þessi áhrif. Í þessu verkefni eru sameindalíffræðilegar afleiðingar stökkbreytingarinnar á tjáningarstjórn MUC5B í þekjufrumum lungna útskýrðar, en meðal annars er notast við CRISPR-Cas9 erfðabreytingaraðferðina. Niðurstöður rannsóknarinnar sýna fram á aukna tjáningu í viðurvist T- einbasabreytileikans. Þessi tjáningaraukning virðist drifin af tveimur atburðum; CpG metyleringar eyja eyðileggst (sem áður hefur verið tengd við bælingu á tjáningu) og nýr bindistaður myndast fyrir hvetjandi umritunarþáttinn C/EBPβ. Ennfremur sýnum við gögn sem tengja T erfðabreytileika MUC5B við lykilþætti frumuálags, svo sem minnkaða frumufjölgun, aukna tjáningu á umritunarþættinum C/EBPζ og aukinn fjölda hvatbera. Þetta verkefni hefur varpað nýju ljósi á rs35705950 stökkbreytinguna í IPF og tengt oftjáningu á MUC5B við sjúkdóminn með beinum hætti. Þessar niðurstöður tengja MUC5B beint við IPF og leggja grunn að nýrri sýn á því hvernig þessi einbasabreytileiki tengist tilurð IPF.

Lykilorð:

CEBPB, lungnatrefjun, metylering, MUC5B rs35705950, stökkbreytingar.

Abstract

Idiopathic pulmonary fibrosis is a severe lung disease with unknown aetiology. The lifespan is approximately 3 to 5 years after diagnosis, and the treatments available such as lung transplant are high risk and are not suitable for all patients. In recent years, new drugs such as pirferidone and nintedanib have proved to slow down IPF progression by targeting profibrotic signaling pathways. However, IPF is usually detected in advanced stages, diminishing the beneficial effect of new treatments. Therefore, to find novel risk factors and preclinical characteristics of IPF has become an important task in the field. A recent genome wide association study (GWAS) performed in IPF patients compared to controls revealed a common *MUC5B* cis-regulatory region polymorphism (rs35705950), where a G is replaced by a T allele in around 45% of IPF patient compared to an 11% allele frequency in controls. This polymorphism has become the strongest risk factor (genetic and otherwise) associated with IPF. Furthermore, the presence of the minor T allele has been associated to subclinical lung fibrosis, suggesting an important role of the *MUC5B* polymorphism early in IPF development.

However, the mechanism underlying the effect of the *MUC5B cis*-regulatory polymorphism remains unknown. In this study we have focused on the molecular consequences of the polymorphism on the regulation of *MUC5B* expression in human bronchial epithelial cells. The results shown in this project corroborate the gain of function role of the minor T allele. Also, we have identified a combined mechanism involving both methylation and direct transcriptional regulation mediated by the polymorphic variant on MUC5B overexpression. On the one hand, the minor T allele destroys a CpG methylation site, previously associated with *MUC5B* inhibition. On the other hand, the replacement of the wild type G allele with a T allele generates a novel C/EBPβ activating transcription binding site. Using two different approach models that include the use of luciferase reporter vectors and CRISPR-Cas9 editing technique we show the differential methylation pattern between both alleles and analyze the effect on *MUC5B* expression. Also, based on CRISPR edited cell lines we show the binding of C/EBPβ in *MUC5B* cis-regulatory domain only in presence of the minor T allele. Finally, we show preliminary data that correlates the MUC5B overexpression in T allele cells to characteristics of the cellular stress, as reduced proliferation ratio,

overexpression of the ER stress marker C/EBPζ transcription factor and increased number of mitochondria.

Altogether, the findings presented in this project assessed the bases for a better understanding of the rs35705950 T allele effect in MUC5B overexpression that would lead to specific treatments in the future as well as validate the use of rs35705950 polymorphism as a risk marker for IPF development.

Keywords:

CEBPB, Idiopathic pulmonary fibrosis, methylation, MUC5B, rs35705950 polymorphism.

Acknowledgements

"Science, my lad, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth."

— Jules Verne, A Journey to the Center of the Earth

The work presented in this thesis was carried out at the Stem Cell Research Unit (SCRU), Biomedical Center, Faculty of Medicine, University of Iceland, and supported by the Icelandic Research Council (RANNÍS) project grant.

During my doctoral studies I have had the priviledge to work with many brilliant and enthusiastic scientists who have inspired me in many ways during my time at the University of Iceland. This has turned my PhD period into an invaluable experience.

First of all, I would like to thank my supervisor Magnús Karl Magnusson for his guidance and support. Thanks to his encouragement and enthusiasm for science I have grown not only as a scientist but also as a person. I could not be more fortunate. Also, I would like to thank Þórarinn Guðjónsson who welcomed me as another one of his students from the beginning. His academic and moral support has been invaluable.

I would like to thank my doctoral committee Erna Magnúsdóttir, Eiríkur Steingrímsson and Gunnar Guðmundsson for their support and helpful discussions during this project.

Thanks to my collegues at the stem cell research unit: Anna Karen, Ari, Bryndis, Bylgja, Eirikur, Erika, Gunnhildur, Halla, Halli, Jennifer, Jón Pétur, Kata, Sigga, Sophie, Sævar and Zuzana. Sincerely thanks to make me feel at home despite being 3.000km away from my home country. Furthermore, I would like to thank all my collegues at the Biomedical Center for their collaboration and advice even if they were not directly related to my project.

My deepest gratitude to my friends, specially to my partner, for his support and unquestionable friendship and love during my doctorate.

Finally, I would like to thank my family, who taught me to always have high goals and to trust in myself. Thanks to them I have learned that there is no limit in what you can do. They also taught me to ignore doubts and explore the

world. What I am today would not have been possible without your eternal support, love and patience.

Contents

List of abbreviations

5'AZA2'	5-aza-2′-deoxycytidine
A549	Adenocarcinoma 549 Cell Line
$A549^{CRISPR}$	Heterozygous [G/T] A549 cell line modified by CRISPR
$A549^{WT}$	A549 control CRISPR-Cas9 cell line
ALI	Air-Liquid Interface
ATI	Alveolar cell type I
ATII	Alveolar cell type II
BCi_NS1.1	Bronchial Cell line from a Non-Smoker patient 1.1
BEGM	Bronchial Epithelial Growth Media
Bp	Base pair
bZIP	Basic Leucine Zipper Domain
CEBPB	CCAAT/Enhancer-Binding Protein β
CF	Cistyc Ffibrosis
ChIP	Chromatin Inmunoprecipitation
COPD	Chronic Obstructive Pulmonary Disease
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CT	Computed Tomography scan
DAB	3,3′-Diaminobenzidine staining
DPLD	Diffuse parenchymal lung diseases
DSP	Desmoplakin
EGF	Epithelial Growth Factor
EM	Electronic Microscopy
EMT	Epithelial to Mesenquimal Transition

eQTL	Expression Quantitative Trait Loci
ER	Endoplasmic Reticulum
FGFR	Fibroblast Growth Factor Receptors
gRNA	guide RNA used in CRISPR-Cas9 editing
GWAS	Genome Wide Association Study
HRCT	High Resolution Computed Tomography scan
IF	Inmunofluorescense
IF-P	Inmunofluorescense staining of Paraffin embebed samples
IIP	Idiopathic Interstitial Pneumonia
IL-13	Interleukin-13
IPF	Idiopathic Pulmonary Fibrosis
LAP	Liver-Enriched Transcriptional Activator Protein
LIP	Liver-Enriched Transcriptional Inhibitor Protein
LUC	Luciferase
M.Sss1	CpG Methyltransferase
MUC	Mucin
MUC5AC	Human Mucin 5AC
MUC5B	Human Mucin 5B
OR	Odd Ratio
PARN	Poly(A)-Specific Ribonuclease
PAX4	Paired Box 4
PCK	Phosphoenolpyruvate Carboxykinase
PCR	Polymerase Chain Reaction
PDGF-R	Platelet Derived Growth Factor Receptor
PI3k	Phosphatidylinositol-4,5-Bisphosphate
PMA	Phorbol 12-myristate 13-acetate
RLA	Relative Luciferase Activity

RTEL	Regulator of Telomere Elongation Helicase 1
RT-qPCR	Quantitative reverse transcription PCR
SD	Standart Deviation
SFTPC	Pulmonary-associated Surfactant Protein C
SFTPC2	Pulmonary-associated Surfactant Protein C2
siRNA	Small Interfering RNA
SMG	Submucosal Gland
SNP	Single Nucleotide Polymorphism
SPhk1	Sphingosine Kinase 1
TERC	Telomerase RNA Component
TERT	Telomerase Reverse Transcriptase
TF	Transcription Factor
TFBS	Transcription Factor Biding Site
TGF-α	Transforming Growth Factor Alpha
TGF-β1	Transforming Growth Factor Beta 1
TNFAIP3	TNF Alpha Induced Protein 3
TOLLIP	Toll-like Receptors
UIP	Usual Interstitial Pneumonia
VEGF	Vascular Endothelial Growth Factor
WT	Wild Type

List of figures

List of tables

List of original papers

This thesis is based on the following original publication, which is referred to in the text by the Roman numeral (I):

I. **A *MUC5B* polymorphism associated with Idiopathic Pulmonary Fibrosis mediates overexpression through decreased CpG methylation and C/EBPβ recruitment driving transcriptional activation.**

   Amaranta U. Armesto, Erna Magnusdottir, Ari J. Arason, Olafur A. Stefansson, Gunnar Gudmundsson, Thorarinn Gudjonsson, Magnus K. Magnusson.

In addition, unpublished data are presented.

All papers are reprinted by kind permission of the publishers.

Declaration of contribution

Paper #*1. A MUC5B polymorphism associated with Idiopathic Pulmonary Fibrosis mediates overexpression through decreased CpG methylation and C/EBPβ recruitment driving transcriptional activation.*

I devised experimental strategies along with my supervisor Magnus Karl Magnusson. I performed all the cell culture and molecular biology experiments with technical input and suggestions from colleges. I wrote the paper along with my supervisor and it was further edited by Erna Magnusdótir and Thorarinn Gudjonson.

1 Introduction

Idiopathic pulmonary fibrosis or IPF is a lung disease of unknown aetiology that gradually diminishes lung function leading to a respiratory failure and consequent death, most often 3 to 5 years after diagnosis. IPF is a heterogeneous disease and is thought to origin in the alveolar region. Even if the main characteristic is the proliferation of myofibroblast that impede the normal lung function, many different cell types are involved and affected in IPF. Genetic studies have implicated several pathways as critical in the development of the disease. Rare mutations with high effect have been linked to both telomere function and surfactant production, suggested to lead to either cellular senescence and/or endoplasmic reticulum (ER) stress in the airway epithelium leading to a profibrotic response.

For a better understanding of IPF pathogenesis, I will present an overview of the lung morphology and the different cell types involved in IPF.

1.1 Overview of the respiratory system: lung morphology and cell types.

The respiratory system consists of the paired lungs and a series of passages that leads to and from the lungs. Within the lung, the air passages branch into increasingly smaller tubes.

Traditionally, the respiratory system is divided into a conducting portion and a respiratory portion. The conducting portion includes the nasal and oral cavities, nasopharynx, oropharynx, larynx, trachea and paired main (primary) bronchi. The main bronchi undergo extensive branching in order to form the bronchioles, the terminal part of the conducting airways (Figure 1). The respiratory portion includes the respiratory bronchioles, alveolar ducts, alveolar sacs and alveoli (Figure 1).

The respiratory system has three principal functions; air conduction, air filtration and gas exchange. In terms of cellular composition and tissue architecture the human respiratory epithelium and stroma varies based on proximal-distal location from the trachea to the alveoli. The cellular compartment has seven different ephitelial cell types with different proportions based on the location throughout the epithelial lining of the human airways (Figure 1) (Knight et al. 2003). These cell types are:

- Ciliated cells: A major component of the upper region of human airways (trachea and larger bronchi) while its concentration diminish down to 25% in terminal bronchioles (Toskala et al. 2005). Its function with the more than 300 cilia present in each ciliated cell is the clearance of the bronchial tree by moving the mucus layer up (Dirksen 1991).
- Brush cell: They are columnar cells that bear blunt microvilli. The basal surface of the cell is in a synaptic contact with an afferent nerve ending (epitheliodentritic synapse). Thus, the brush cells are considered a receptor cell.
- Goblet cells: Mucus or goblet cells are the major producer of mucus in upper airways. Mucus is secreted into the lumen for entrapment of foreign particles and pathogens (Wolff 1986). The mucus production is finely balanced, and its disruption is associated with several lung diseases such as chronic obstructive pulmonary disease (COPD), cystic fibrosis (CF), asthma and Idiopathic pulmonary fibrosis (IPF) (Lumsden et al. 1984, Kirkham et al. 2002, Plantier et al. 2011). The number of goblet cells is increased in the submucosal glands (SMGs), a structure that lies in the submucosa in the larger airways and extend to the luminal surface. The SMGs consist of goblet, serous, basal-like and myoepithelial cells and is the major producer of mucins (MUC5B and MUC5AC), that is excreted to the luminal surface to protect the airway (Lynch et al. 2014).
- Basal cells: These are the progenitor cells for the differentiated epithelial layer and reside basally in the pseudostratified epithelium, in close connection with the basement membrane. Similar to ciliated cells, the number of basal cells diminishes through the distal airways down to the the bronchioles (Boers et al. 1998, Rock et al. 2009).
- Club (clara) cells: The progenitor role of the basal cells is replaced by club cells in the bronchioles down to the alveoli. It is debated if club cells are able to differenciate into alveolar cell types, in addition to the differentiation into ciliated and goblet cells (Kim et al. 2005, Rawlins et al. 2009). Another function of the club cells is to secrete surface-active and protective surfactants in healthy lung that reduces the surface tension during the gas exchange.
- Pulmonary neuroendocrine cells: They have been shown to sense and react to airway hypoxia and contribute to wound repair and

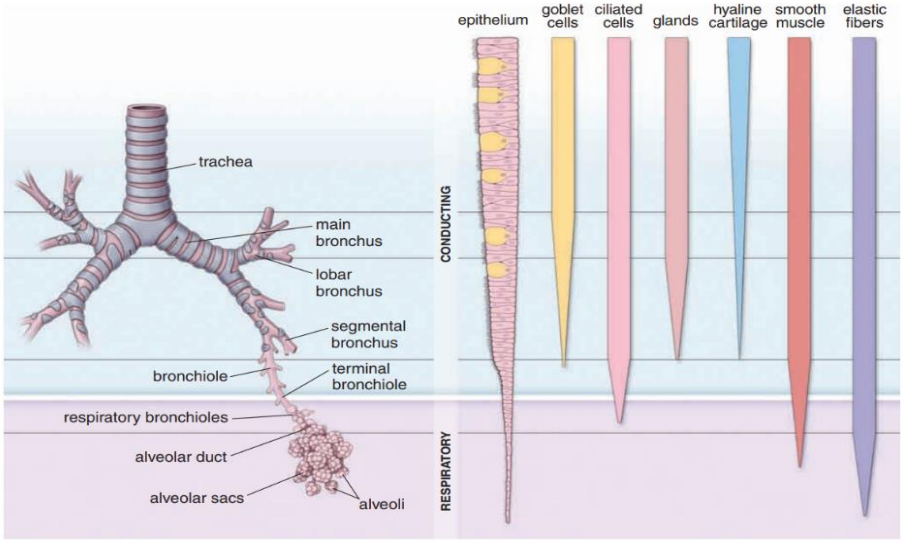


Figure 1. Division of the bronchial tree and summary of its histological features. (Ross et al. 2016)

repopulation of differentiated ephitelial cells subsequent to injury. Neuroendocrine cells can be found alone or in bundles known as a neuroephitelial bodies (Adriaensen et al. 2003).

The alveolar ducts and sacs are composed of different cell types, including:

- Type I alveolar cells: They are extremely thin and form the major component of the alveoli. Their role is to facilitate gas exchange using their large and thin body which is in close contact with the pulmonary vasculature (Stone et al. 1992).
- Type II alveolar cells: These cells have a cuboidal structure and are mainly located between type I alveolar cells. Their function is to secrete surfactants to maintain a low surface tension and prevent alveolar fluid accumulation. It has been shown that Type II alveolar cells can also transdiferentiate into Type I alveolar cells (Barkauskas et al. 2013, Desai et al. 2014).

1.2 Idiopathic pulmonary fibrosis (IPF).

Idiopathic interstitial pneumonias (IIP) represent a group of diffuse parenchymal lung diseases with varying patterns of fibrosis and inflammation. They all fall within the broad category of diffuse parenchymal lung disease (DPLD) (American Thoracic et al. 2002). Within IIPs, idiopathic pulmonary fibrosis (IPF) is the best defined and most common phenotype. It is a severe lung disease characterized by progressive diminution in lung function based on an underlying fibrotic process in the lung parenchyma. The severity of the progressive nature of IPF leads to a median survival of less than 3 years, similar to what is seen in stage 1 and 2 non-small cell lung cancer. The other IIP phenotypes include cryptogenic organising pneumonia, acute interstitial pneumonia, nonspecific interstitial pneumonia, desquamative interstitial pneumonia, and respiratory bronhiolitis interstitial lung disease (Figure 2) (Bradley et al. 2008). Idiopathic pulmonary fibrosis can also be divided into two subtypes that include sporadic and familial IPF. Traditionally, the cases have been divided between these two categories based on a minimun of two consanguineous members of the same family that develop IPF. Usually, familial pulmonary fibrosis is characterized by the presence of rare mutations related to IPF, such as in telomere genes *TERT* and *TERC* or the surfactant protein genes *SFTPC* and *SFTPC2*.

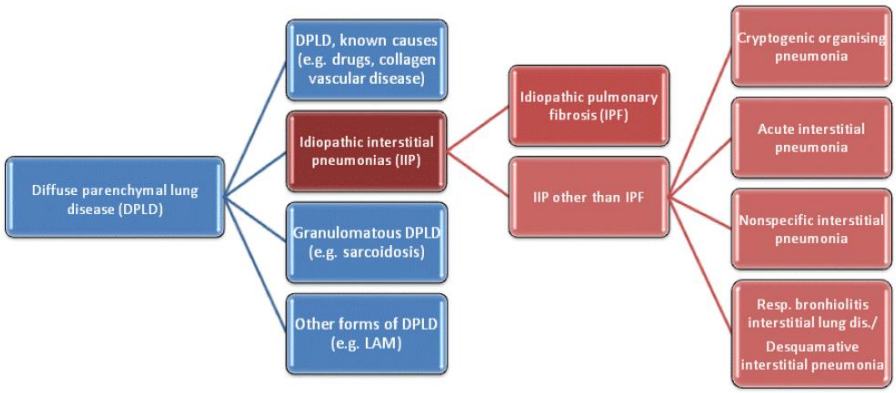


Figure 2. Diffuse parenchymal lung diseases classification.

Idiopathic pulmonary fibrosis affects 5 million world-wide, being more frequent in men and has increased prevalence with age (Navaratnam et al. 2011, Hutchinson et al. 2015). Death occurs in up to 50% of patients within 3-5 years after diagnosis (2000, Raghu et al. 2011, Hutchinson et al. 2015). In recent years the incidence of IPF has been increasing over time. Even though IPF is still considered an underdiagnosed disease (Olson et al. 2007, Navaratnam et al. 2011, Raghu et al. 2011, Hutchinson et al. 2014), its prevalence keeps growing through the time. The increasing prevalence of IPF could be partially explained by the aging population and the better characterization of the disease that has lead to earlier and more accurate diagnosis (Olson et al. 2007, Hutchinson et al. 2014).

Several studies have been carried out to establish a uniform incidence of IPF between countries. However, the task has been challenging. Even if there are many epidemiological studies that have estimated the incidence, they are limited by non-standarized parameters and variability between countries. Hutchinson et. al. (2015) have summarized 32 different studies around the globe and report an incidence range of 3–9 cases per 100.000 per year in Europe and North America. Also, they reported a lower incidence in East Asia and South America. Most of the studies corroborate an increased incidence over time.

While the incidence is increasing, IPF has limited treatment options. The only curative treatment is a bilateral lung transplant for qualified IPF patients, with a 50% 3 years survival ratio post-transplant (Neurohr et al. 2010). However, there is a relatively small number of patients that qualify for this treatment and the rate of mortality as a consequence of the transplant is high (Mason et al. 2007). Other conventional options include the use of oxygen, sildenafil for pulmonary hypertension and the treatment of gastric acid reflux aspiration (Collard et al. 2007, Mermigkis et al. 2007, Idiopathic Pulmonary Fibrosis Clinical Research et al. 2010). In the past decade novel treatment trials for IPF have focused on inhibition of fibroproliferation pathways. Two novel drugs targeting these pathways have been shown to slow down IPF progression, Pirferidone and Nintedanib. Pirferidone is known to interfere with profibrotic signaling through TGF-β1 inhibition possibly by inhibiting the upregulation of HSP47 and Col1 RNA in fibroblasts (Margaritopoulos et al. 2016), while Nintedanib is a triple-kinase inhibitor targeting platelet-derived growth factor receptors (PDGF-R), fibroblast growth factor receptors (FGFR) and vascular endothelial growth factor (VEGF)-family tyrosine kinase receptors (Bonella et al. 2015).

Despite these advances, the improvement in quality-of-life of the novel drugs is very limited due to the late diagnosis of IPF in most patients. For this reason, it is important to identify risk factors for IPF, such as genetic predictors that can increase the diagnositic yield at early stages.

1.3 Clinical and pathological features of IPF.

The pathological hallmark of IPF on imaging and histology is usual interstitial pneumonia (UIP). UIP is characterized by basal predominant fine reticulonodular abnormality, wich may be present for many years before clinical presentation with IPF. However, these features are relatively nonspecific (Mathieson et al. 1989) and therefore the use of high-resolution CT (HRCT) scans is crucial for IPF evaluation (Mayo 2009). The main features on HRCT are reticular abnormalities, traction bronchiectasis and honeycombing, with predominant basal and peripheral distribution, and with absence features that would suggest an alternative diagnosis (Raghu et al. 2011). Nevertheless, HRCT doesn't differentiate IPF from asbestosis or chronic hypersensitivity pneumonitis. Identification of underlying cause for UIP therefore remains an important task to achive better IPF diagnosis.

The histological characteristics of IPF include progressive scarring of the lung interstitium dominated by collagen types-1 and -2 fibrosis and proliferation of fibroblast/myofibroblast in the stroma and regeneration of adjacent alveolar epithelium (Raghu et al. 2011, Travis et al. 2013). A heterogeneous insterstitial fibrosis, fibroblastic foci and subpleural microscopic honeycombing are the main histological characteristics of IPF. It is thought that fibroblastic foci represent early and characteristic features of IPF. An injury in the type 1 alveolar epithelial cells may lead to the development of numerous foci of injury and repair resulting in the progressive accumulation of extracellular matrix. The fibroblastic foci may lead to an uncontrolled proliferative process that can severly affect lung function (Selman et al. 2001).

Another main feature of IPF is alveolar cell type II (AT II) hyperplasia. It is hypothesized that these cells can convert to myofibroblasts by epithelial to mesenchymal transition (EMT) (Barkauskas et al. 2013). AT II is the major progenitor of alveolar cell type I (or AT I), that represents 90% of alveolar region (Selman et al. 2008). The replacement of AT I by AT II is also hypothesized to increase the profibrotic signal and collagen accumulation, leading to fibroblast and myofibroblast accumulation (Willis et al. 2005, Selman et al. 2006).

1.4 Genetic and environmental determinants of pulmonary fibrosis.

IPF is considered a disease with a heterogenous aetiology, that develops as a result of a combination of genetic and non-genetic (exogenous and endogenous) risk factors. Due to the complexity of IPF disease and variability between individuals, the contribution of each factor has not been clearly elucidated.

1.4.1 Risk of IPF.

Traditionally, the risk of developing IPF has been associated with numerous environmental factors that include the exposure to metal and wood dust (Iwai et al. 1994, Hubbard et al. 1996, Baumgartner et al. 2000, Hubbard et al. 2000), viruses (Stewart et al. 1999, Tang et al. 2003, Lawson et al. 2008), asbestos (Brody 1993) and drugs (Erwteman et al. 1977, Musk et al. 1979, Hubbard et al. 1998). Of the environmental factors, the most relevant is cigarette smoking (Spira et al. 2004) which has been shown to increase IPF risk even after smoking cessation. These finding suggest that repetitive lung injury is an important factor in IPF development.

Of the genetic risk factors, a limited number of rare variants has been associated with IPF, including variants in *TERT* (Armanios et al. 2007, Tsakiri et al. 2007), *TERC* (Armanios et al. 2007, Tsakiri et al. 2007), *RTEL* (Cogan et al. 2015, Stuart et al. 2015) and *PARN* (Stuart et al. 2015) all in the telomerase pathway and variants in the surfactant protein genes *SFTPC* (Thomas et al. 2002, van Moorsel et al. 2010) and *SFTPC2* (Wang et al. 2009).

In 2011, Seibold et al. detected a linkage between idiopathic interstitial pneumonia and a 3.4-Mb region of chromosome 11p15 in 82 families. They evaluated genetic variation in this region in gel-forming mucin genes expressed in the lung among 83 subjects with familial interstitial pneumonia, 492 subjects with idiopathic pulmonary fibrosis, and 322 controls. Based on their research they identified a variant (rs35705950) in a putative *cis*-regulatory region of the gene encoding *MUC5B,* associated with the development of both

Table 1. **Additive Genotypic Associations of *MUC2, MUC5AC,* and *MUC5B* Single-Nucleotide Polymorphisms (SNPs) in Subjects with Familial Interstitial Pneumonia or Idiopathic Pulmonary Fibrosis and Controls.***

SNP	Nucleotide and Amino Acid Change	Mucin Region	Nucleotide Position†	Minor-Allele Frequency			Genotypic Association Test			
				FIP (N=83)	IPF (N=492)	Controls (N=322)	Odds Ratio for FIP (95% CI)	P Value	Odds Ratio for IPF (95% CI)	P Value
				percent						
rs10902081	C→T	*MUC2* Int7	1079809	37.2	38.6	47.9	0.6 (0.4–0.9)	0.011	0.7 (0.5–0.8)	4.3×10^{-4}
rs7127117‡	T→C	*MUC2* Int7	1079879	49.3	60.0	47.4	1.0 (0.7–1.5)	0.826	1.6 (1.3–2.0)	6.9×10^{-5}
rs41453346	C→T Tyr426Tyr	*MUC2* Ex10	1080894	5.0	6.5	2.2	1.9 (0.8–4.3)	0.124	2.8 (1.6–5.2)	0.001
rs41480348	G→A Thr618Thr	*MUC2* Ex15	1082605	8.4	6.5	12.1	0.7 (0.4–1.2)	0.188	0.5 (0.4–0.8)	0.001
rs7934606‡	C→T	*MUC2* Int31	1093945	49.4	54.0	40.5	1.4 (1.0–2.0)	0.055	1.7 (1.4–2.2)	3.8×10^{-6}
rs10902089‡	A→G	*MUC2* Int31	1094357	57.9	58.8	48.5	1.5 (1.0–2.1)	0.031	1.5 (1.2–1.9)	2.9×10^{-4}
rs9667239	C/T	*MUC2-5AC* intergenic region	1143101	22.5	21.0	12.5	2.2 (1.4–3.6)	0.001	1.9 (1.4–2.7)	5.6×10^{-4}
rs55846509	G→A Arg47Gln	*MUC5AC* Ex2	1154294	3.1	5.5	1.6	1.7 (0.6–5.1)	0.316	3.6 (1.7–7.3)	0.001
rs28403537	C→T Ala497Val	*MUC5AC* Ex12	1161315	8.9	13.0	3.4	2.7 (1.3–5.3)	0.006	4.6 (2.8–7.6)	3.2×10^{-9}
MUC5AC-025447‡	C→T	*MUC5AC* Int26	826476§	20.1	21.0	13.8	1.6 (1.0–2.5)	0.053	1.6 (1.2–2.2)	0.003
rs35288961	G→T	*MUC5AC* Int46	1220462	28.8	26.6	15.9	2.2 (1.4–3.5)	3.2×10^{-4}	2.0 (1.5–2.6)	3.7×10^{-6}
rs35671223	C→T	*MUC5AC–5B* intergenic region	1227069	42.6	42.4	33.4	1.4 (1.0–2.0)	0.05	1.5 (1.2–1.9)	0.001
rs28654232	C→T	*MUC5AC–5B* intergenic region	1229227	21.6	22.8	32.9	0.6 (0.4–0.9)	0.009	0.6 (0.5–0.8)	1.1×10^{-4}
rs34595903‡	C→T	*MUC5AC–5B* intergenic region	1230393	21.5	23.3	34.8	0.5 (0.3–0.7)	0.001	0.5 (0.4–0.7)	2.4×10^{-6}
rs2672794	C→T	*MUC5B* promoter	1241005	27.2	27.5	40.4	0.5 (0.3–0.8)	0.001	0.5 (0.4–0.7)	1.9×10^{-7}
rs35705950	G→T	*MUC5B* promoter	1241221	33.8	37.5	9.1	6.2 (3.7–10.4)	3.7×10^{-12}	8.3 (5.8–11.9)	4.6×10^{-31}
rs35619543‡	G→T	*MUC5B* promoter	1242250	40.3	39.0	23.8	2.4 (1.6–3.6)	3.3×10^{-5}	2.1 (1.6–2.8)	1.5×10^{-8}
rs12804004	G→T	*MUC5B* promoter	1242299	39.2	39.4	48.9	0.6 (0.4–0.9)	0.019	0.6 (0.5–0.8)	1.2×10^{-4}
rs868903‡	T→C	*MUC5B* promoter	1242690	65.4	61.0	49.5	1.8 (1.3–2.6)	0.001	1.6 (1.3–2.1)	2.8×10^{-5}

* FIP denotes familial interstitial pneumonia, and IFP idiopathic pulmonary fibrosis.
† Nucleotide positions are based on data from human genome browser hg19, except where denoted otherwise.
‡ For these SNPs, DNA was available for 304 controls.
§ The position provided for this nucleotide is based on genome build NW_001838016.1.

Figure 3. Variants analyzed by Seibold et al (2013) at chromosome 11p15 related to IPF and IIP. The minor-allele of the single-nucleotide polymorphism (SNP) rs35705950, located 3 kb upstream of the MUC5B transcription start site, was present at a frequency of 34% among subjects with familial interstitial pneumonia, 38% among subjects with idiopathic pulmonary fibrosis, and 9% among controls (allelic association with familial interstitial pneumonia, P=1.2×10−15; allelic association with idiopathic pulmonary fibrosis, P=2.5×10−37). The odds ratios for disease among subjects who were heterozygous and those who were homozygous for the minor allele of this SNP were 6.8 (95% confidence interval [CI], 3.9 to 12.0) and 20.8 (95% CI, 3.8 to 113.7), respectively, for familial interstitial pneumonia and 9.0 (95% CI, 6.2 to 13.1) and 21.8 (95% CI, 5.1 to 93.5), respectively, for idiopathic pulmonary fibrosis.

familial interstitial pneumonia and sporadic idiopathic pulmonary fibrosis (Figure 3).

MUC5B expression in the lung was 14.1 times higher in subjects who had idiopathic pulmonary fibrosis than in those who did not (P<0.001). Also, the minor T allele of rs35705950 was associated with increased expression of MUC5B in the lung in unaffected subjects (expression was 37.4 times higher than in unaffected subjects homozygous for the wild-type allele, P<0.001) and they report that MUC5B protein is expressed in lesions of idiopathic pulmonary fibrosis.

Almost at the same time, two large genome-wide association studies (GWAS) of patients with both familial and sporadic pulmonary fibrosis compared to controls were conducted (Fingerlin et al. 2013, Noth et al. 2013). Both studies corroborate the linkage of the rs35705950 polymorphism to IPF and IIP, being today considered the strongest IPF risk factor, genetic or otherwise.

Other polymorphisms were identified in the GWAS studies, including the *TOLLIP* gene (close to *MUC5B* polymorphism DNA region), *FAM13A, DSP*, *OBFC1*, *ATP11A* and *DPP* locis. Altogether, these results highlight the importance of telomere maintenance (*TERT*, *TERC*, and *OBFC1*) and host defence (*MUC5B*, *ATP11A, TOLLIP*), and identified barrier function (*DSP*, *DPP*) genes as important for IPF development.

1.4.2 Rare vs common IPF risk variants.

IPF is a complex disease where many different factors (environmental and genetics) are combined to trigger its development. Focusing on genetic risk factors, both rare and common variants, contribute to IPF development. The *MUC5B cis*-regulatory variant is the strongest known risk factor for pulmonary fibrosis, but rare variants when present have a larger effect on disease risk. As presented in Figure 4, the more common variants generally have lesser effects (lower odds ratio), but because they are common, they have the potential to explain more of the population burden of disease. For example, *DSP* has an allele frequency of ~40% and a OR of 1.43, while rare variants allele frequency acounts for less than 0.1% of cases (Kaur et al. 2017). The *MUC5B* variant presents unique characteristics as its allele frequency is ~11% with a high OR (4.5-6.6), meaning that even if the variant is considered common, it presents a large effect on IPF risk.

Nevertheless, it must be emphasized that given the low incidence of IPF, the impact of common variants on the absolute risk of IPF is low.

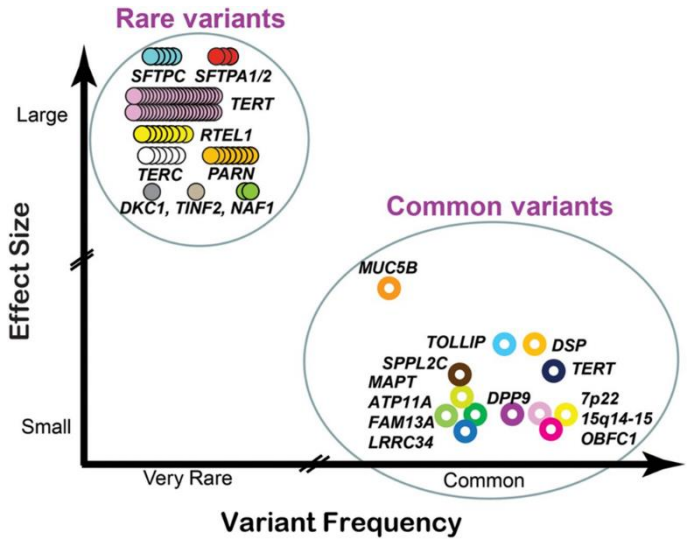


Figure 4. Relationship between variant frequency and penetrance of the risk allele (Mathai et al. 2016).

1.4.3 MUC5B variant rs35705950.

As we described above, the *MUC5B* upstream variant is the strongest known risk factor for IPF. Recent studies have examined the relevance of *MUC5B* variation in other ethnic groups, showing that the frequency of rs35705950 minor allele (a G-to-T SNP) is 11, 8 and 1% among European, South Asian and East Asian populations, respectively (Wang et al. 2014, Horimasu et al. 2015, Peljto et al. 2015), and it is almost non-existent in Africans (Genomes Project et al. 2015). Despite the different prevalence, the evidence indicates that the risk of developing IPF associated with the variant is comparable to the risk observed in European population (Hunninghake et al. 2013, Peljto et al. 2013). Furthermore, carriers of the *MUC5B* variant are at risk of getting subclinical interstitial lung abnormalities based on screening with high-resolution CT as shown by Hunninghake et al (2013). Hunninghake et al (2013) described that between the 2633 participants of their study, 7% presents interstitial lung abnormalities, and for each copy of the minor rs35705950 allele, the odds of interstitial lung abnormalities increased 2.8-fold (95% confidence interval [CI], 2.0 to 3.9; P<0.001), and the odds of definite CT evidence of pulmonary fibrosis were 6.3 times greater (95% CI, 3.1 to 12.7; P<0.001).

The *MUC5B cis*-regulatory variant has also been analysed in other lung

diseases such as asbestosis, sarcoidosis, scleroderma associated instertitial lung disease, chronic obstructive pulmonary disease (COPD), asthma and lung cancer without significant association (Peljto et al. 2012, Borie et al. 2013, Stock et al. 2013). The odds ratio (OR) for IPF associated with carrying one allele of the variant is 4.5-6.6 which is very high for such a common variant, while homozygosity leads to an OR of 9.6-20.2 (Zhang et al. 2011, Borie et al. 2013, Fingerlin et al. 2013, Noth et al. 2013, Stock et al. 2013, Wei et al. 2014, Horimasu et al. 2015). The elevated risk associated with the *MUC5B* variant is substantially higher than for most other risk variants in complex diseases, except for the human leukocyte antigen (HLA) region for some autoinmmune diseases such as type I diabetes which has OR=10 (Erlich et al. 2008).

The *MUC5B* variant is also an unusual genetic variant. It is simple in the sense that it is genetically isolated, i.e. it is not in strong linkage disequilibrium with other variants. It also carries a strong eQTL association with MUC5B expression, specifically in lung (Figure 5).

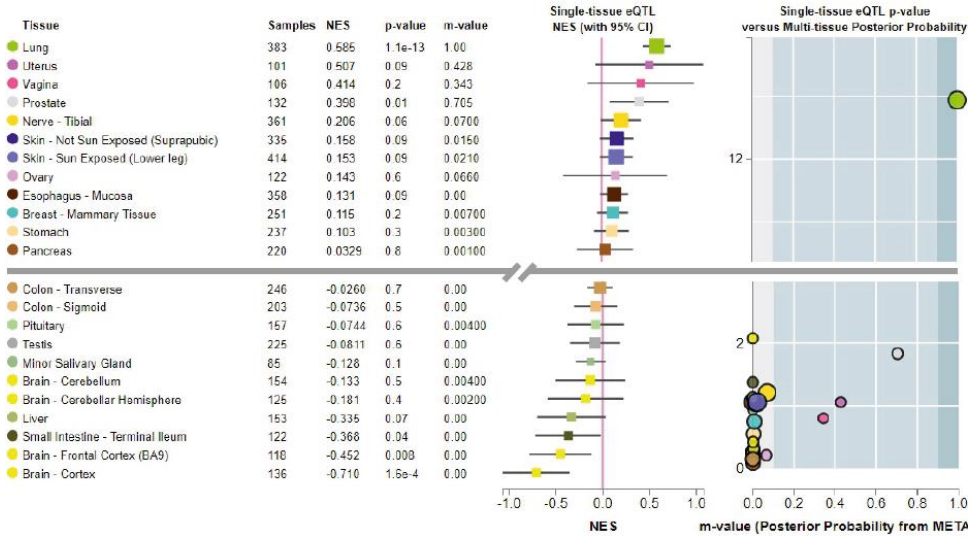


Figure 5. Multi-tissue eQTL comparison. rs35705950 polymorphism effects on MUC5B expression in different human tissues. Adapted from GTEx website. Non-relevant data has been removed.

1.4.4 Mucin gene expression and disease.

The rs35705950 risk allele is a gain-of-function variant associated with increased expression of MUC5B in unaffected subjects (Seibold et al. 2011) and presumably among subjects with IPF (Helling et al. 2017). Nevertheless, the mechanism through which MUC5B overexpression can lead to IPF remains unknown. Dysregulation of mucin expression is observed in some diseases such as gastric and colorectal cancer (Reis et al. 1999, Sylvester et al. 2001, Wakatsuki et al. 2008, Jung et al. 2010, Tamura et al. 2012, Walsh et al. 2013, Park et al. 2015). Based on the signaling pathways identified in those studies, the alteration of mucin expression has been hypothesized to lead to an aberrant host defense barrier function as drivers of disease risk. For example, knockout of mucin 2 (*MUC2*) increases expression of the Toll-like receptor (TLR) pathway (Burger-van Paassen et al. 2011), and knockout of mucin 4 (*MUC4*) is associated with increased colonization of *Helicobacter pylori* (Park et al. 2015). Erosion of barrier function due to alterations in mucin 1 (*MUC1*) and *MUC4* are important for metastasis in gastric (MUC1) (Tamura et al. 2012), ovarian (MUC4) (Ponnusamy et al. 2011) and lung cancer (MUC4) (Ponnusamy et al. 2011).

Nevertheless, one characteristic shared by mucin dysregulation discussed above is the inhibition of mucin production as a risk factor. Based on the GTEx data and previous analysis performed by other groups (Seibold et al. 2011, Helling et al. 2017), rs35705950 polymorphism is associated with higher expression of MUC5B. Considering the protective role of MUC5B in host defense (Evans et al. 2016), other groups have started to analyse a possible evolutionary selection of the polymorphism (Evans et al. 2016) to understand how this variant becomes so frequent. In fact, among patients with IPF, the

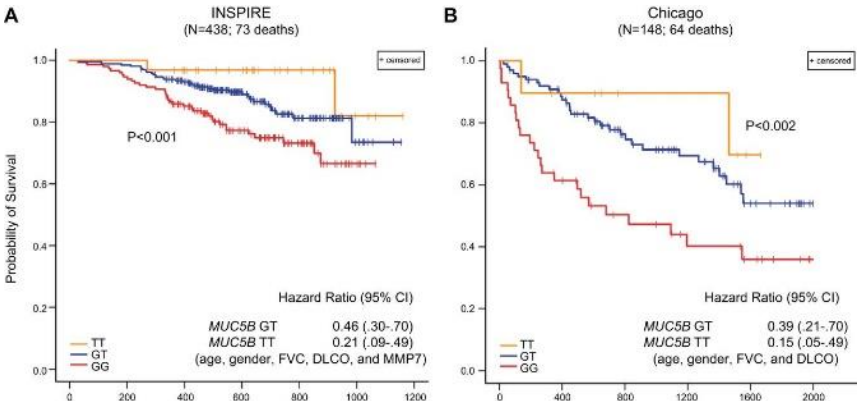


Figure 6. Probability of survival based on rs35705950 genotype. (Peljto et al. 2013)

MUC5B variant is associated with an improved prognosis (Figure 6) (Peljto et al. 2013) and higher bacterial burden (Molyneaux et al. 2014).This strengthens the idea that the *MUC5B* polymorphism provides an improvement in the host defense system. Different studies are ongoing, comparing European ancestry population (~9% baseline frequency of the rs35705950 T allele) to East Asians and Africans where the presence of the T allele is rare to estimate the age and genealogy of the risk variant to shed light on important selective events relevant to the evolution of the rs35705950 variant.

1.5 Homeostatic and pathobiological funtions of lung mucins.

Mucus is comprised of water, salt, cells, cellular debris and other macromolecules that are held together in a mesh formed by polymeric mucin glycoproteins (Thornton et al. 2008, Fahy et al. 2010). Along with cilia, mucus is the first innate defense barrier against pathogens in the airways. The quantity and qualities of mucins and the maintenance of cilia structure and motility is closely regulated in lung ephitelium. Dysregulated mucin production, aberrant cilia assembly or poor control of salt and water contents in mucus layer are related to disease development or exacerbation such as in cystic fibrosis, severe asthma, COPD and IPF (Evans et al. 2016). Mucus accumulation is linked to airflow obstruction and can lead to fatal asphyxiation and furthermore affects underlying cell function.

1.5.1 Mucin genes.

Nineteen different mucin genes have been described: *MUC1, MUC2, MUC3A, MUC3B, MUC4, MUC5AC, MUC5B, MUC6, MUC7, MUC8, MUC12, MUC13, MUC15, MUC16, MUC17, MUC19,* and *MUC20. MUC5B* is in a cluster of mucin genes at chromosome 11p15, close to *MUC5AC, MUC2* and *MUC6*. MUC5AC and MUC5B are the major component of the airway mucins.

The signaling pathways regulating mucin production are complex and poorly understood and not much is known specifically about *MUC5B* regulation. Vincent et al (2007) identified DNA hypermethylation of *MUC5B* as a major mechanism responsible for its silencing. They also described histone acetylation as a potential repressor of *MUC5B* expression. Different signaling pathways related to inflammation have been linked to *MUC5B* expression such as MEK1/2, PI3K, SPhk1, and MAPK14 (p38α-MAPK) by Interleukin-13 (IL-

13) activation (Atherton et al. 2003, Kono et al. 2010). Furthermore, through PMA signalling, PKC, EGF/TGF-a, Ras/Raf, Mek, ERK and Sp-1 signalling pathways have also been associated with *MUC5AC* and *MUC5B* expression (Hewson et al. 2004). In a recent study on *MUC5B* regulation by Helling et al. (Helling et al. 2017) they described the role of a well preserved FOXA2 binding site in the same regulatory domain as rs35705950 polymorphism (32bp downstream), suggesting the importance of this region for *MUC5B* regulation. The FOXA2 binding site does not overlap the rs35705950 polymorphism and no evidence of differential binding of this transcription factor was shown in carriers of the T minor allele. They showed that disruption of the FOXA2 binding site seems to significally decrease MUC5B production. How the rs35705950 polymorphism directly affects *MUC5B* expression remains unknown, but the region is likely to be an important regulatory site, given the strong effect the polymorphism has on MUC5B expression.

1.5.2 Secreted polymeric mucin structures.

Mucins are very large glycoproteins (in the mega-Dalton range) characterized by a NH_2-terminal and COOH-terminal cysteine rich domains highly conserved between species. This biochemical characteristic makes mucins thick viscoelastic polymers that can vary in assemblies from 2 to more than 20 oligomers.

Furthermore, the central portion of MUC5B and MUC5AC is comprised of a mucin glycosylation domain that is rich in serines and threonines. Attached to the serines and threonines are O-linked N-acetylgalactosamines followed by diverse carbohydrate structures that increase the heterogeneity of mucins. On the other hand, the glycosylation domain is interrupted by additional cysteine-rich CysD domains (~100aa) that form hydrophobic loops. Glycosylation is another property that affects mucin size and charge (Evans et al. 2016).

Altogether, the biochemical features of mucin glycoprotein make MUC5B difficult to study and the effects of marked MUC5B overexpression has not been directly studied in airway epithelial cells.

1.6 Clinical relevance of MUC5B.

MUC5B is a highly glycosylated protein, expressed throughout the upper and lower respiratory tract. In human upper airways MUC5B is predominantly

Human		Mouse	
Normal	**Disease**	**Normal**	**Disease**
Lung	adenocarcinoma	Lung	bacterial infection
	IPF[a]		
	cystic fibrosis[b]	cystic fibrosis[c]	
	panbronchiolitis[b]		
Trachea		Trachea	asthmatic model
Salivary glands		Salivary glands	
Nose	chronic rhinosinusitis	Nose	allergic rhinitis
Middle ear & Eustachian tube	otitis[d,e]	Middle ear & Eustachian tube	otitis
Eye (conjunctiva)		Eye (conjunctiva)	
Gallbladder[f]	hepatolithiasis	Gallbladder	
Endocervix[g,h]	endometrial tumors	Endocervix	
Breast	cancer[i]	Mammary tissue	mammary tumors
Stomach	gastric cancer		

Figure 7. MUC5B dysregulation associated diseases in human and mouse. Adapted table from Portal et al.(2017).

expressed in nasal and oral gland secretions, while in tracheobronchial conducting airways it is highly expressed in submucosal glands and in surface epithelial goblet cells but less predominantly than the related MUC5AC (Fahy et al. 2010, Evans et al. 2016). Under many disease conditions, the *MUC5B* expression pattern changes dramatically, being overexpressed and showing significantly higher expression than *MUC5AC* (Young et al. 2007, Roy et al. 2014).

It has been previously shown that MUC5B but not MUC5AC is necessary for mucociliary clearance (Roy et al. 2014), and that a $Muc5b^{-/-}$ mice accumulates apoptotic macrophages, show impaired phagocytosis and furthermore interleukin-23 (IL-23) production was reduced. By contrast, in mice that transgenically overexpress Muc5b, macrophage functions are improved (Roy et al. 2014).

However, as we mentioned before, *MUC5B* dysregulation is present in several different diseases, not only in lung but also in different types of cancer such as gastric and breast cancer (Figure 7) (Portal et al. 2017).

1.7 Proposed models of MUC5B-induced pulmonary fibrosis.

Even though the *MUC5B* variant has a well-supported genetic influence in IPF development, the mechanism underlying it is poorly understood. Currently there are at least three hypotheses that link enhanced production of MUC5B in the bronchiolo-alveolar region to the development of pulmonary fibrosis (see Evans et al.,(2016)).

The first hypothesis focuses in the high presence of MUC5B in honeycomb cysts. Excessive MUC5B production driven by the rs35705950 variant in stem cells attempting to regenerate injured bronchiolar and alveolar epithelium, disrupt normal reparative signalling in the distal lung, resulting in chronic fibroproliferation and honeycomb cyst formation. Due to the large size of the MUC5B protein, its production may carry a significant metabolic stress, which can interfere with differentiation of airway stem cells.

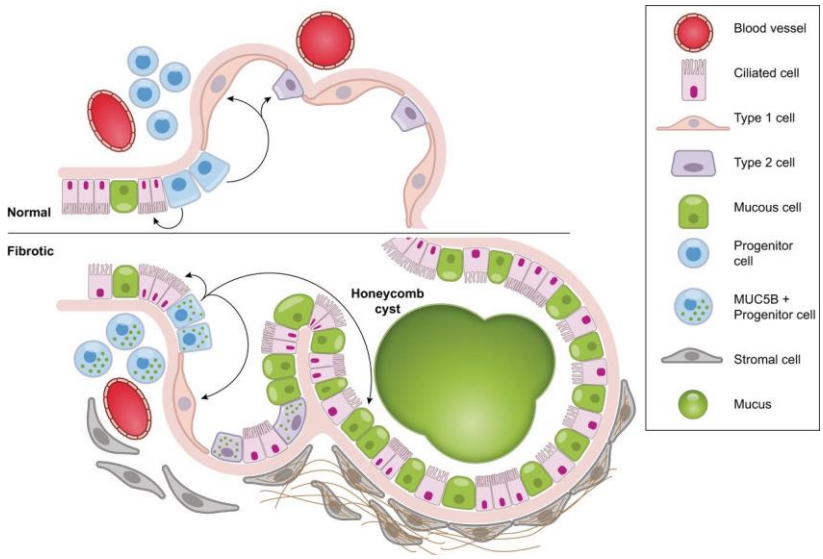


Figure 8. First proposed model for MUC5B-induced pulmonary fibrosis. (Evans et al. 2016)

The second hypothesis considers the possibility of IPF as a mucociliary disease caused by recurrent injury/inflammation/repair at the bronchoalveolar junction. MUC5B overexpression would cause reduced mucociliary function, retention of particles and enhanced lung injury. Over time, the foci of lung injury lead to scar tissue and persistent fibroproliferation expanding and displacing normal lung tissue.

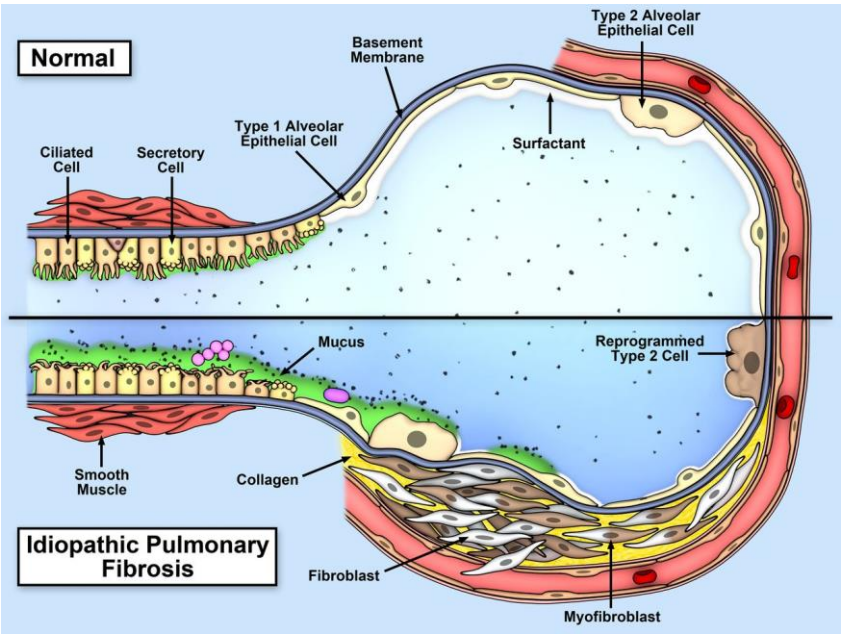


Figure 9. Second proposed model for MUC5B-induced pulmonary fibrosis. (Evans et al. 2016)

The third line of reasoning focuses on the relationship between MUC5B and motile cilia in the distal airway, in combination with the second hypothesis mentioned above. The hypothesis considers that honeycomb cyst could be generated because of aborted attempts to regenerate the terminal airways via activation of developmental pathways that signal in the primary cilium.

In summary, the origin of idiopathic pulmonary fibrosis is still unknown, but recent GWAS studies point to genetics as an important factor to understand the pathways involved. More specifically, interaction between common variants maybe key to understanding IPF. Furthermore, the *MUC5B* G to T variant has been shown to increase not only the IPF risk but also the development of interstitial lung abnormalities in healthy individuals. Nevertheless, little is known about *MUC5B* regulation or structure, although some reports have linked *MUC5B* dysregulation to different diseases such as asthma, cystic fibrosis or different types of cancer. To understand how MUC5B is overexpressed in IPF and the role that the rs35705950 variant plays may be a key to further asses the T polymorphism as a genetic predictor for IPF, allowing an early diagnosis and consequently improving therapeutic options

and effectivity.

1.8 Role of non-coding sequence variants in disease.

It has long been known that variations in regulatory regions of genes can impact gene expression. Due to improvements and decreasing cost of sequence technologies a vast number of studies using genome wide variants have identified and catalogued these noncoding variants associated with diseases (Mathelier et al. 2015). These studies are accumulating evidence of the role of noncoding variants in multiple human diseases. For example, a meta-analysis that included more than 1200 GWAS SNPs showed that more than a third of noncoding variants are likely to be causal for the phenotypic or disease traits observed (Visel et al. 2009). Some of the more promising examples includes the germline mutation in the TERT promoter, associated to melanoma, where a novel transcription factor binding site (TFBS) appears and allows the binding of an ETS family transcription factor (TF) (Horn et al. 2013, Khurana et al. 2016). The functional effects are more likely to be observed in the tissues where these TF family is expressed. Moreover, a SNP in the *MDM2* promoter increases the binding affinity of Sp1 TF, leading to accelerated tumor formation in many different cancer types through MDM2 overexpression and subsequent p53 pathway suppression (Bond et al. 2004, Bond et al. 2007). Variants in enhancers have been also related to various diseases, such as those described for *MYC*. Several SNPs in the *MYC* enhancer have been linked to increased risk for different cancer types. The MYC enhancer variants have also been linked to epigenetic changes, such as histone methylation and acetylation in enhancer regions in a tissue specific maner and these epigenetic signals have the potential to explain the specificity of these variants in each cancer type (Grisanzio et al. 2010).

Nevertheless, to directly understand the role of individual noncoding variants in human diseases has been notoriously difficult. The activity of regulatory regions is controlled through the interplay between many regulatory pathways such as epigenetic modifications, chromatin conformation and binding of transcription factors (TFs). Improvements in bioinformatic tools that compare sequences and predict *cis*-regulatory regions and TFBS have played a significant role in accelerating noncoding variants research, but the complexity of the mechanism underlying gene expression makes this endevour difficult (Mathelier et al. 2015).

1.9 Control of gene expression.

Even though all cells in an organism contains the same DNA, the specific subset of genes and mRNA expressed at each time point and cell type makes them present an exclusive morphology and function. Among the mechanism underlying this specificity, the most important are transcriptional and epigenetic regulation.

1.9.1 *Cis*-regulatory regions and transcriptional regulation.

The *cis*-regulatory sequences are binding sites for transcription regulators, whose presence on the DNA affects the rate of transcription initiation. These sequences can be located adjacent to the promoter, far upstream of it, or even within introns or enterly downstream of the gene (Alberts 2015). Many transcription regulators act through *mediator,* while some interact with the general transcription factors and RNA polymerase directly. Transcription regulators also act by recruiting proteins that alter the chromatin structure of the promoter. Moreover, whereas *mediator* and the general transcription factors are the same for all RNA polymerase II-transcribed genes, the transcription regulators and the locations of their binding sites relative to the promoter differ for each gene (Alberts 2015).

Furthermore, the mechanism of recognition of the *cis*-regulatory regions, is highly specific. Transcription regulators recognize short stretches of double-helical DNA of defined sequence called *cis*-regulatory sequence, and thereby determine which of the thousands of genes in a cell will be transcribed. Although each of these transcription regulators has unique features, most bind to DNA as homodimers or heterodimers and recognize DNA through structural motifs (Alberts 2015). These structural motifs make a series of contacts with the DNA by hydrogen bonds, ionic bonds, or hydrophobic interactions. Transcription regulators typically work in groups and bind to DNA cooperatively through approximately 20 of these contacts that are typically formed at the protein-DNA interface. Both, dimerization of transcription regulators and specifity of *cis*-regulatory regions recognition combine to makes transcriptional regulation dynamic and highly specific (Alberts 2015, Li et al. 2015).

Transcription regulators can act as activators or repressors of gene expression and they can assemble on other DNA-bound transcription regulators to form coactivators or co-repressors. In some cases, RNA polymerases can also act as scaffolds to hold a group of proteins together

(Alberts 2015). Once bound to the DNA, transcription regulators attract and position RNA polymerase II at the promoter (that transcribes all the protein-coding genes and many noncoding RNA genes) and furthermore induce the release of the polymerase from the promoter so that transcription can begin triggering changes in the chromatin structure of the promoters and making the underlying DNA more accessible. For *cis*-regulatory elements far away from the promoter, transcription regulators usually act inducing changes in the 3D chromatin conformation (Alberts 2015).

However, the presence of a transcription regulator binding site does not simply imply functionality (Li et al. 2015). There are subsets of epigenetic modifications linked to improved transcription regulators binding (active marks), and a subset linked to reduced binding (repressive marks) (Chen et al. 2013).

1.9.2 Epigenetics.

Epigenetic modifications are reversible and potentially heritable changes occurring in genomic DNA and chromatin that do not alter DNA sequence. The types of epigenetic modifications traditionally include DNA methylation and histone modifications (Wu et al. 2016).

1.9.2.1 CpG methylation and its role in transcriptional activation.

Methylation is one of the major epigenetic markers that regulate gene expression, and among the different methylation mechanisms, CpG methylation is one of the most important.

CpG methylation is a biochemical process in which a methyl group is added to a cytosine at the 5' position of a CpG dinucleotide, converting the cytosine to a 5-methyl cytosine (Bernstein et al. 2007, Alberts 2015). Methylation of cytosine provides a mechanism through which gene expression patterns can be passed onto progeny cells. The sequence CG is base-paired to exactly the same sequence in opposite orientation in the other DNA strand, allowing the existence of an inherited DNA methylation pattern by the daughter strand. An enzyme called *maintenance methyl transferase* acts preferentially on those CG sequence that are base-paired with a CG sequence that is already methylated (Alberts 2015). As a result, the CG methylation pattern is preserved through cell division and represents a major mechanism for gene repression, especially important in cell differentiation (Bernstein et al. 2007).

Nevertheless, methylated C nucleotides in vertebrate genome tend to be eliminated in the course of evolution, due to the way DNA repair enzymes work (Alberts 2015). Accidental deamination of an unmethylated C gives rise to U, easily recognized by DNA repair enzymes. But accidental deamination of a 5-methyl C can not be repaired in this way, because the deamination product is a T so is indistinguishable from the non-mutant T. Due to this, methylated C in the genome tends to disappear over time and is replaced by a T nucleotide. Consequently, more than three out of four CGs has been lost in this way and the remained CGs are clusteres in the so-called CpG islands (Alberts 2015).

The CpG islands are regions of around 1000 base pairs length, that usually include the promoters of genes. CpG islands also remain unmethylated in most somatic tissues whether or not the associated gene is expressed. The unmethylated state is maintained by sequence-specific DNA-binding proteins, many of whose *cis*-regulatory regions contain a CG. By their binding, CG sites are protected against methyl transferases. These proteins can also recruit DNA methylases, which conver 5-methyl C to hydroxyl-methyl C (Alberts 2015).

The presence of a methyl group is typically associated with gene silencing and is involved in many biological processes such as cell development, cell differentiation and immune responses (Bernstein et al. 2007). DNA methylation is regulated by methyltransferases that alters the methylation status depending on cell requirements (Bernstein et al. 2007). There are four different mechanisms where methylation can regulate TF function, the most common being where the CpG methylation interferes directly with TF binding (Fan et al. 2009). Many TFBS contain a CpG methylation site. When it is methylated the TFs fail to bind DNA and transcriptional activation is blocked. Also, DNA methylation can alter chromatin structure forming a co-repressor complex leading to decreased TF binding (Fan et al. 2009). Other mechanisms include the methylation of the TF promoter, leading to transcription repression, the recruitment of DNA methyltransferases induced by TFs or the promoted transcription of DNA methyltransferases by TFs (Wu et al. 2016).

2 Aims

Idiopathic pulmonary fibrosis is a fatal lung disease. The incidence of the disease is increasing, but the aetiology of IPF remains poorly understood. Recently, genetics has highlighted important pathways that are likely to be involved in disease generation and progression. The *MUC5B* polymorphism, the most important risk factor (genetic and otherwise) in IPF development, suggests a role for MUC5B. The rs35705950 polymorphism inside a cluster of mucins in chromosome 11 is located 3kb upstream of the *Mucin 5B* transcription starting site. The minor allele has been associated with higher expression of MUC5B but how it triggers IPF remains unknown. To understand the mechanism through which the minor allele upregulates MUC5B is important to determine its contribution in IPF and to find new pathways and possible targets contributing to the development of this fatal disease.

The aim of this thesis is to directly address the role of rs35705950 polymorphism in MUC5B overexpression as well as to determine possible mechanism underlying its effect.

Specific aims:

1. Study the direct effects of rs35705950 (G-to-T) polymorphism on MUC5B expression:

    a. To determine if the minor T allele directly affects the expression of MUC5B.

2. Mechanisms:

    a. To analyse if the MUC5B upregulation through the T allele is mediated by epigenetic mechansims.

    b. To analyse possible modifications in transcriptional regulation due to G-to-T switch and determine if it creates novel transcription factor binding sites leading to MUC5B upregulation.

3. Effect of G-to-T polymorphism on cellular phenotype:

    a. Conduct preliminary studies to evaluate the effect of the minor T-allele on the cellular phenotype in airway epithelial cells.

3 Materials and methods

In this chapter I will briefly go over the main methods used in this thesis. More detailed methods can be found in material and methods chapter in *Paper I* included in the thesis.

Cell culture

The BCi_NS1.1 cell line is a human bronchial epithelial cell line kindly provided by Dr. Matthew S. Walters, Weill Cornell Medical College, New York NY, USA (Walters et al. 2013). It was established by immortalization with retrovirus expressing human telomerase (hTERT). The bronchial epithelial cell line VA-10 was previously established by retroviral transduction of primary bronchial epithelial cells with E6 and E7 viral oncogenes (Halldorsson et al. 2007). Both cell lines were cultured in bronchial epithelial growth medium, BEGM (Lonza, Walkersville, MD) supplemented with 50 IU/ml penicillin and 50 µg/ml streptomycin (Gibco, Burlington, Canada).

The human lung adenocarcinoma derived alveolar epithelial cell line A549 (American Type Culture Collection, Rockville MA) was cultured in DMEM-Ham's-F12 basal medium supplemented with 10% fetal bovine serum (FBS), 50 IU/ml penicillin and 50 µg/ml streptomycin (Gibco). All cell lines were grow at 37°C, 5% CO_2.

Air-Liquid interface culture

To establish an air-liquid interface cultures (ALI), cells were seeded on the upper layer of Transwell cell culture filters (Corning®Costar®) pore size 0.4 µm, 12 mm diameter, polyester membrane) (Sigma-Aldrich, St. Louis, USA) at a density of 2×105 cells per well. The cultures were maintained on chemically defined bronchial epithelial cell medium (BEGM, Life Technologies/Sigma) for 5 days, 0,5 ml in the upper chamber and 1.5 ml in the lower chamber. After 5 days, medium was changed to DMEM/F-12 (Invitrogen), supplemented with 2% Ultroser G (Cergy-Saint-Christophe, France) for another 5 days. For ALI cultures, the medium was aspired from the apical side and the filter rinsed with PBS.

Goblet cell differentiation by IL-13 treatment

Cells were cultured for 5 days on BEGM and then for 5 days on DMEM/UG in a submerged culture. After 5 days of ALI culture, IL-13 (Peprotech, London, UK) was added to the basal side to a final concentration of 25 ng/ml and cultured for 14 days.

Proliferation Assay

For performing proliferation assays, one thousand (A549) and five thousand (BCi_NS1.1) cells per well were seed in a 48 well plate depending of the cell line. Confluence ratio was measured each two hours during a week. Four different pictures were taken by well.

Goblet cell differentiation by Phormol 12-myristate 13-acetate (PMA) treatment

Cells were cultured for 5 days on BEGM and then for 5 days on DMEM/UG in a submerged culture. After 5 days of ALI culture, PMA (Peprotech, London, UK) was added to the basal side to a final concentration of 10 µM and cultured for 10 days.

5-aza-2'-deoxycytidine DNA Methylation Inhibition

For ALI-cultures, cells were cultured for 5 days on BEGM and then for 5 days on DMEM/UG in a submerged culture. After 10 days of ALI culture, 5-aza-2'-deoxycytidine (Peprotech, London, UK) was added to the basal side to a final concentration of 10µM and cultured for 3 days.

Immunofluorescence staining

Cells were rinsed twice with chilled PBS. The fixation of cells was performed using 100% methanol at -20°C overnight. After removing methanol, cells were submerged in 100% acetone for one minute. Staining was performed using the immunofluorescence IMF buffer (0.1% TX-100, 0.15M NaCL, 5nM EDTA, 20mM HEPRES, pH 7.5, 0.02% NaN3 as preservative). Cells were incubated with primary antibody overnight at -4°C, and then rinsed three times, 15 min with IMF buffer. Cells were then incubated with secondary antibody and DAPI for two hours at room temperature, followed by four times washing with IMF

buffer. Cells were mounted using ProLong Antifade (Thermo Fisher Scientific). Antibodies used for this assay are listed in Table 2.

Immunofluorescence was visualized and captured using laser scanning Fluoview® FV1200 Confocal Microscope (Olympus Life Science).

Mitochondria quantification

MitoTracker™ Green FM (ThermoFisher) was added to fresh culture medium at final concentration of 50nM for 30 minutes. Cells were then washed with PBS and fixed.

Transient transfection

Cells were seeded at 70% confluence one day before transfection. FuGENE® HD Transfection Reagent (Promega) was used on BCi_NS1.1 and VA10 cells, while Lipofectamine (Thermo Fisher Scientific) was used on A549 cell line. All transfections were performed following manufacturer's instructions. The results were analysed 48h after transfection. Plasmids used for C/EBPβ overexpression was generously donated by Joan Massague: C/EBPβ LAP isoform (addgene#15738) and C/EBPβ LIP isoform (addgene#15737) [51]. Plasmids used for this assay are listed in Table 3.

Production of lentiviral and cell transduction

To produce the lentiviral cell lines containing pGreenFire1™ Pathway Reporter lentivector (Cat#TR010PA-N and Cat#TR000PA-1, System Bioscience) expressing the *MUC5B cis*-regulatory region and controls, we followed the general guideline provided by System Bioscience. Briefly, 70% confluent HEK-293T cells were cultured for 24 h w/o antibiotics and transfected (Lipofectamine, Thermo Fisher Scientific) with lentiviral transfection constructs and packaging plasmids (psPAX2 and pMD2.G) (Addgene plasmids #12260 and #12259, respectively). Culture medium containing the virus was harvested 24 and 48 hours post transfection and centrifuged at 1250 rpm at 4°C for 5 minutes and filtered through 0,45 µm filter. Lentiviral particle solution was added to the culture medium (containing 8 µg/ml polybrene) and then added to culture flasks of 70% confluent cells (BCi_NS1.1, VA10 and A549) at a low multiplicity of infection (MOI) and incubated for 20 hours. Cells were then cultured further for 24 hours in fresh culture media. Infected cells were then selected with puromycin or neomycin as appropriate for 48hs. Plasmids used for this assay are listed in Table 3.

Luciferase Assay

Each cell type was seeded at 70% confluence one day before transfection in a 96 well plate. To perform the luciferase assay, the Dual-Glo® Luciferase Assay System kit supplied by Promega was used, following the general guideline provided with the kit. Luminosity was measured in a microplate reader ModulusTM II (Turner BioSystem). Luciferase measurement were normalized using Renilla co-transfection. Plasmids used for this assay are listed in Table 3.

Real Time qPCR

RNA was isolated using Tri-Reagent® solution (Ambion) and cDNA preparation was carried out using RevertAid™ First strand cDNA Synthesis Kit (Fermentas) according to the manufacturer's instructions. Real-time PCR using Power SYBR Green PCR Master mix (Applied Biosystems) was used to detect the relative quantity of each cDNA. GAPDH was used as the endogenous reference gene. Data were analysed using 7500 Software v2.0 (Applied Biosystems). All primers used are listed in Table 1.

Western Blot

Protein lysates were acquired using RIPA buffer supplemented with phosphatase and protease inhibitor cocktails (Life Technologies). For western blots, 5µg of protein were loaded per lane, unless otherwise stated. Samples were denatured in Laemmli buffer, 10% β-mercaptoethanol by heating at 95°C for 5 min and then run on NuPage 10% Bis-Tris gels (Life Technologies) in 2-(N-morpholino) ethanesulfonic acid (MES) running buffer. Samples were then transferred to Immobilon FL PVDF membranes (Millipore). Membranes were blocked in Li-cor blocking buffer and primary antibodies were incubated overnight at 4 °C. Near-infrared fluorescence visualization was measured using the Odyssey CLx scanner (Li-Cor, Cambridge, UK). Antibodies used are detailed in Table 2.

In Vitro Methylation Assay

DNA fragments were cut out of the pGL3-*MUC5B*pr vector using XbaI-EcoRI restriction enzymes, generating 4.1Kb of the *MUC5B cis*-regulatory

region, with WT or T allele. Fragments were gel-purified using GeneJET PCR Purification Kit (ThermoFisher Scientific) following the manufacturer's instructions and subsequently methylated with the *M.SssI* methyltransferase (New England Biolabs) overnight at 37°C. The methylated fragments were then religated into the pGL3 basic vector. DNA concentration was measured at 260nm before the fragments were used in transfection on BCi_NS1.1, VA10 and A549 cell lines as described by Vincent et al (Vincent et al. 2007). Differential influence of methylation in the 4.1kb *MUC5B cis*-regulatory was measured by luciferase activity in three different experiments in a triplicate experiment of each transfected cell line. All plasmids used are listed in Table 3.

Bisulfite sequencing

$1x10^5$ cells were used to extract DNA from each cell line. Extraction was performed with PureLink Genomic DNA MiniKit (Invitrogen) following manufacture's instruction. Bisulfite conversion of DNA was done with EZ DNA Methylation-Gold™ Kit (ZymoPURE™, Germany). Amplification of the region of interest was done with the EpiMark® Hot Start Taq DNA Polymerase (New England Biolab, UK) and the resulting product was sequenced by Sanger sequencing (Beckman Coulters Genomics, GENEWIZ, UK.) DNA methylation analyses of bisulfite PCR amplicons were performed using Sequence scanner V1.0. DNA methylation level was scored as percentage methylation of individual CpG units in each sample. Primers used for bisulphate sequencing are listed in Table 1.

siRNA transfection

Small-interfering RNAs (siRNA) targeting human C/EBPβ were purchased from Sigma Aldrich (St Louis, MO, USA). siRNA ID: SASI_Hs02_00339146, SASI_Hs01_00236023, SASI_Hs02_00339148, SASI_Hs01_00339149, 10nM each. siRNA transfection was performed following manufactures instructions. We used MISSION siRNA Universal Negative Control (Sigma Aldrich) as a control. Cells were seeded in a 96 well plates for Luciferase Assays, or in a 12 well plate to be analysed by RT-qPCR. siRNAs were transfected using Lipofectamine 2000 reagent (Invitrogen) in OPTI-MEM medium (GIBCO). Twenty-four hours later, the transfected cells were transferred to complete medium. After 48h, the cells were harvested and used for luciferase measurements and RT-qPCR analysis.

CRISPR

Cells were seeded in a 12 well plate prior transfection with Cas-9WT vector, gRNA and homologous region (100bp single nucleotide sequence overlapping rs35705950 position carrying the minor allele used by Cas9WT enzyme to replace the original sequence), at 70% confluence. CRISPR efficiency was increased with non-homologous recombination inhibitor (Sigma-Aldrich). Selection was performed by Blasticidin. Single cell cloning assays were performed to select individual clones. Sequencing of individual clones was done to corroborate the genome. gRNA and repair template sequences are included in Table 1.

Chromatin immunoprecipitation (ChIP)

$1x10^{10}$ cells were used for cross-linking. Chromatin immunoprecipitation (ChIP) was done following the protocol previously described by Boyer et al. (Boyer et al. 2005) with the following modifications: Cross-linking of cells was done using 0.4% formaldehyde (ThermoFisher Scientific). Sonication was done using 300µL sample containing $1x10^{10}$ cells, with 30 cycles of 15'' sonication, followed by 30'' cooldown period. Primers and antibodies are listed in Table 1 and 2.

Statistical Analysis

Data are presented as means with standard deviations of measurements unless stated otherwise. Statistical differences between samples were assessed with Student two–tailed T-test. P-values below 0.05 were considered significant (***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq .0.05$).

Table 1. Summary of primer sequences.

ChIP	Forward	Reverse
ChIP1-M5Bpr	AGCTATTGAGACATCCCGGA	GCTGTGTCCCTTTCCTTCCT
ChIP2-M5Bpr	GAGTCGGGCCTGTTTGCT	AACCGGCTGTGTCCCTTT
ChIP3-M5Bpr	TGGGAGTCGGGCCTGTTT	CAGGAACCGGCTGTGTCC
ChIP4-M5Bpr	CTGCAGATGACGCTGTCTGT	GGGGCCCCAGCTTATGTAG
Negative Control - GAPDH	TCG AAC AGG AGG AGC AGA GAG CGA	TAC TAG CGG TTT TAC GGG CG
Positive Control - TNFAIP3	GCTGTTGCTCAATTGCTAGTC	CTTCTTGTGCTTACTTTCAGTTCTT
CRISPR	**Forward**	**Reverse**
Sequencing 1	ATGCTACTGGAAGCCTCGAA	CATCAGCTCCCAGGCACT
Sequencing 2	ATGCTACTGGAAGCCTCGAA	CATCAGCTCCCAGGCACT
Sequencing 3	GTTGGACCACAGGCACTGA	CTGCACAGCGACGTGAAC
Sequencing 4	CTGCAGATGACGCTGTCTGT	GGGGCCCCAGCTTATGTAG
Sequencing 5	GAATTCATGCTACTGGAAGCCTCGAA	TCTAGAGGGTCTGTCCCCAGAAGAAC
MUC5B gRNA1	AAAACTGGGAGTCGGGCCTGTTTGC	ACACCGCAAACAGGCCCGACTCCCAG
MUC5B gRNA2	AAAACCAAACACGCTGAGCAAACAGC	ACACCGCTGTTTGCTCAGCGTGTTTG
MUC5B gRNA4	ACACCGGGAGTCGGGCCTGTTTGCT	AAAACAGCAAACAGGCCCGACTCCC
MUC5B gRNA5	AAAACGATGCGGCCCTGGGAGTCGGC	ACACCGCCGACTCCCAGGGCCGCATC
Homologous region 1	TTATCTTCTGTTTTCAGCTCCTTCAACTGTGAAGAGGTGAACTCTTCAAACACGCTGAGCAAACAGGCCCGACTCCCA	
	GGGCCGCATCCGGGATGTCTCAATAGCTGTGGCCTTGACGTCCACCTCGGACCCCTGCCCCGGACCCAGCCCA	
Homologous region 2	AGGTGGACGTCAAGGCCACAGCTATTGAGACATCCCGGATGCGGCCCTGGGAGTCGGGCCTGTTTGCTCAGCGTG	
	TTTGAAGAGTTCACCTCTTCACAGTTGAAGGAGCTGAAAACAGAAGATAAAGGAAGGAAAGGGACACAGCCGGTTC	
Clonning primers	**Forward**	**Reverse**
SEQ1	ATGCTACTGGAAGCCTCGAA	CATCAGCTCCCAGGCACT
SEQ2	GGCTCTGAGCAGACCAAGAG	CTCAGGCAGCTCCTCTGTC
SEQ3	GTTGGACCACAGGCACTGA	CTGCACAGCGACGTGAAC
SEQ4	AGCCATGAGGGGTGACAG	TGACGAGCGTCATCTACAGG
SEQ5	CTGCAGATGACGCTGTCTGT	GGGGCCCCAGCTTATGTAG
MUC5Bpr-pGL3-MluI-NheI	ACGCGTATGCTACTGGAAGCCTCGAA	GCTAGCGGGTCTGTCCCCAGAAGAAC
MUC5Bpr-pGF-XbaI-EcoRI	GAATTCATGCTACTGGAAGCCTCGAA	TCTAGAGGGTCTGTCCCCAGAAGAAC
rs35705950_SEQ_587bp	CGGGTTCTGTGTGGTCTAGG	GCATCAGCGAGATAGCGTTT
Site Directed Mutagenesis [T]	CCTACGAAGGtGACAAAGAGTCCG	TGTGGCAGCGTCCTGGGA
Site Directed Mutagenesis [T]	CCTACGAAGGaGACAAAGAGTCCG	TGTGGCAGCGTCCTGGGA
Bisulfite sequencing	**Fordward**	**Reverse**
Bisulfite Sequencing	TTTGGTTAGAATGAGGGATAGTGAT	CAAAACCACAACTATTAAAACATCC
qRT-PCR	**IDT identification number**	
MUC5B	Hs.PT.5822513172g	
MUC5AC	Hs.PT.5115096441g	
GAPDH	Hs.PT.39a.22214836	
C/EBPβ	Hs.PT.5827185099g	

Table 2. Summary of antibodies.

Protein	Identification number
MUC5B	ab87376
MUC5AC	ab3649
C/EBPβ	ab32358
Acethylated α-tubulin	ab11323
Phalloidin	a22283
ChIP	**Identification number**
C/EBPβ	sc-150
Rabbit (DA1E) mAb IgG XP® Isotype Control	#3900

Table 3. Summary of plasmids.

MUC5B promoter Cloning vectors	Catalogue Number	Company
pGL3-LUC reporter vector	E1751	Promega
Renilla Reporter vector	E2231	Promega
pGF-LUC-GFP Lentivector	TR010PA-N	System Bioscience
pGF-LUC-GFP Negative Control	TR000PA-1	System Bioscience
psPAX2	#12260	Addgene
pMD2.G	#12259	Addgene
CRISPR Vectors	**Catalogue Number**	**Company**
CAS9-WT	#44758	Addgene
gRNA backbone (MLM3636)	#43860	Addgene
C/EBPβ overexpression	**Catalogue Number**	**Company**
C/EBPβ LAP	#15738	Addgene
C/EBPβ LIP	#15737	Addgene

4 Results

The common polymorphism rs35705950, where a G is replaced by a T nucleotide, has been shown in two GWAS studies and in seven subsequent replications to be correlated with a predisposition to develop IPF. Also, the strong eQTL association with MUC5B overexpression in lung suggests that the rs35705950 polymorphism acts by increasing MUC5B expression although the effects of increased MUC5B expression and how it may lead to IPF remains unknown. The molecular mechanism underlying the increased MUC5B expression effect remains unknown.

To corroborate the overexpression of MUC5B in IPF we compared the expression between normal and IPF human lung by immunohistochemistry (3,3'-Diaminobenzidine (DAB) staining) (Figure 10A) and immunofluorescense (Figure 10B) staining of paraffin-embedded samples. The results show that MUC5B (green) was expressed at a higher level in all the different analysed regions of lung tissue (bronchi, alveoli, and submucosal gland) of samples from

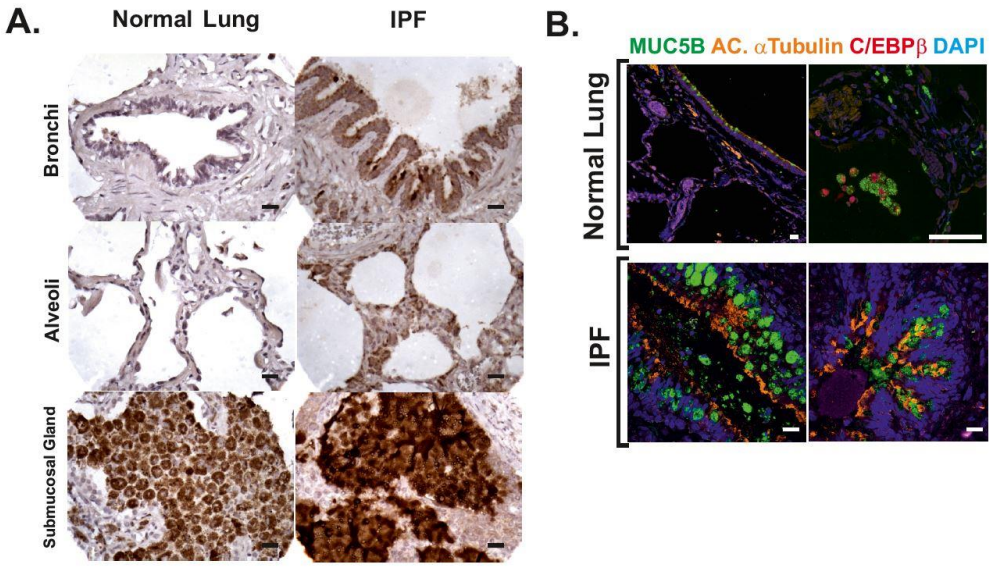


Figure 10. MUC5B is overexpress in IPF. a) Immunohistochemical staining of IPF and healthy human lung samples. b) Immunofluorescent staining in IPF and normal human lung samples. The antibodies against the different proteins are indicated with colors.

IPF patients whereas in normal lung, MUC5B was found in a small proportion in epithelia and in alveolar macrophages based on the morphology (Figure 10B, right). These results support the observation that MUC5B is expressed at a high level in IPF lungs compared to healthy lungs, pointing to MUC5B as an important factor in IPF development.

4.1 Rs35705950 risk allele for IPF is associated with higher expression of MUC5B.

Genetic studies need rigour in the use of comparable genotypes while analyzing the effects of a specific polymorphism. This means that the presence of other polymorphisms can affect the results of the variant being analyzed. Also, the signaling pathways expressed by each cell type vary under many circumstances (cell type, cancer related cell lines, etc.). Due to this, the *in vitro* experiments were done using three different cell lines, two immortalized epithelial cell lines, BCi_NS1.1 and VA-10, and the well-stablished adenocarcinoma cell line, A549. All these cell lines carry the WT allele at the rs35705950 locus. In each cell line the specific polymorphism, rs35705950, was studied in an isogenic background after genetic manipulation using different approaches (see below). BCi_NS1.1 and VA-10 were generated through immortalization using different approaches, but both have been shown to be valid models to mimic non-cancerous bronchial epithelium (Halldorsson et al. 2007, Walters et al. 2013). On the other hand, A549 is an alveolar cell line that even if is a cancerous cell line, responds well to genetic manipulations. Unfortunately, A549 is not able to differentiate in ALI cultures, restricting the variety of experiments for this line.

a. Transient transfection.

Even if the T allele has been associated with MUC5B overexpression by GTEx analysis (Figure 5) and in tissues, we can not guarantee that the effect is due to an enhanced activity of the *MUC5B cis*-regulatory region driven by T allele. To address directly the effect of the T allele on MUC5B overexpression, the 4.1kb of the *MUC5B cis*-regulatory region was cloned into a luciferase reporter vector, and the WT allele was replaced by site-directed mutagenesis with the T allele (Figure 11B). The region selected was amplified from BCi_NS1.1 cell line by PCR, previously shown by Sanger sequencing to carry the WT genotype. The site directed mutagenesis was done using specific primers that overlap

the rs35705950 region, and the plasmid obtained was sequenced to corroborate a specific mutation of the allele of interest.

Both vectors, carrying the T allele and the WT control allele were transiently transfected into BCi_NS1.1, VA10 and A549 cells and 48 hours after tranfection luciferase activity was assessed. Transfection was normalized based on Renilla luminiscense, transfected at the same time. As shown in Figure 11A there was a significant increase in the vector carrying the T allele compared to WT (~32% increase) in BCi_NS1.1 cells indicating a direct regulatory effect of the allele on downstream gene expression. This effect was not seen with VA10 and A549 cell lines, were the differences between WT and T allele were not significant. A possible explanation is the absence of epigenetic markers that trigger the difference between alleles. Other possible explanation may be the absence or maybe the low expression or inaccessibility of the transcription factor that modulates the expression through the T allele.

b. Lentiviral reporter plasmid.

In a transient transfection, epigenetics may be diminished in gene regulation analysis, due to the transitory expression of the plasmid in the cytoplasm. To create a more accurate model, the same 4.1kb *MUC5B cis*-regulatory region carrying each allele (T and WT) were cloned into a luciferase lentiviral vector and transduced into our three cell lines to generate stable cell lines (Figure 11D). Positive cells were selected by a resistance marker (puromycin in VA10 and A549 cell lines and neomycin in BCi_NS1.1) for three days after transduction. During the experiments, growth media was supplied with corresponding antibiotic to guarantee the exclusive growth of transduced cells.

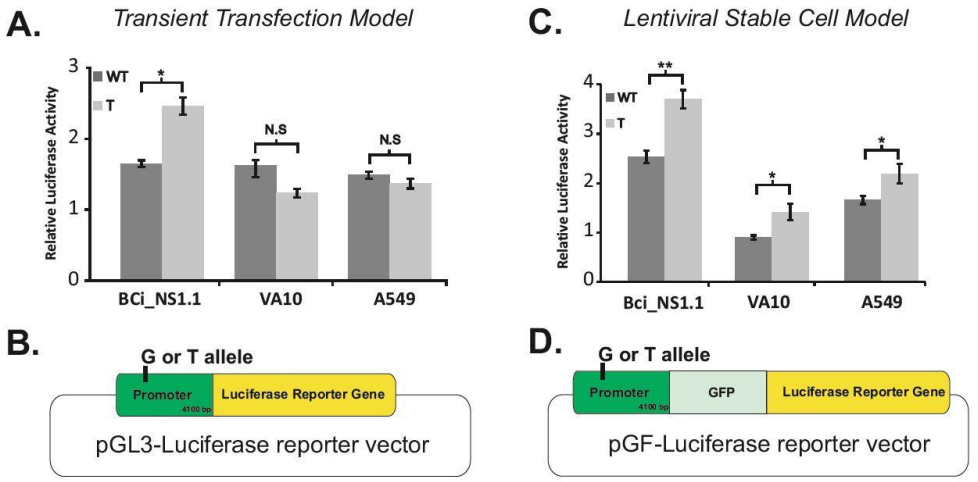


Figure 11. Risk allele (T) is associated with a higher MUC5B *cis*-regulatory region activity. a) Transient transfection on three lung epithelial cell lines (1µg) co-transfected with Renilla (0,5 µg). Luminescence was measured after 24h of transfection. RLA is obtained by dividing Luciferase activity by Renilla luminescence. Experiment was done in triplicate. b) Schematic view of pGL3-luciferase reporter vector with MUC5B 4,1Kb promoter region insert carrying each allele. c) Luciferase activity on stable cell lines transduced with luciferase reporter lentivector. Luminescence was measured after 24h in a monolayer culture. d) Schematic representation of pGF-Luciferase reporter lentivector with both alleles (*p<0.001 and **p <0.0001 with error bars representing SD).

As shown, an increased *cis*-regulatory activity on luciferase expression as measured by luciferase activity is seen when the T allele replaces the WT allele. This is most notorious in BCi_NS1.1 cell line (~34% increase). In contrast with the previous experiments using transient transfection, a significant increased activity was also seen in both VA10 (~24%) and A549 (~20%) cell lines. These results support a gain-of-function role of the T allele in regulating MUC5B expression.

c. CRISPR-Cas9 edited cell line.

To mimic more accurately the effect of rs35705950 T allele in MUC5B expression, the CRISPR-Cas9 editing technique was used in A549 and BCi_NS1.1 cell lines.

To generate the CRISPR edited cell lines, a guide RNA was specifically designed to recognize the region overlapping the rs35705950 polymorphism and 20 nucleotides region, called homologous region, was used to replace the WT allele with the T allele.

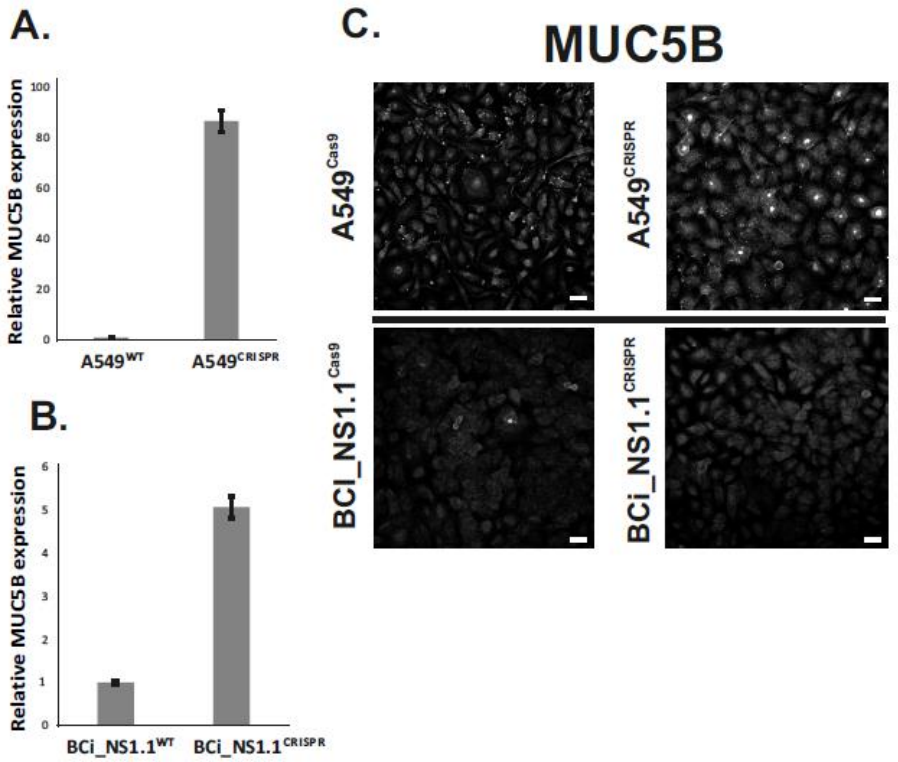


Figure 13. A549 and Bci_NS1.1 CRISPR cells show higher MUC5B expression compared to control. MUC5B expression was analyzed by RT-qPCR in wild type and CRIPSR A549 and BCi_NS1.1 cell lines (A and B) and by immunofluorescense staining of MUC5B protein (C). Both experiments show higher MUC5B expression in CRISPR cells. Scale bars = 50µm.

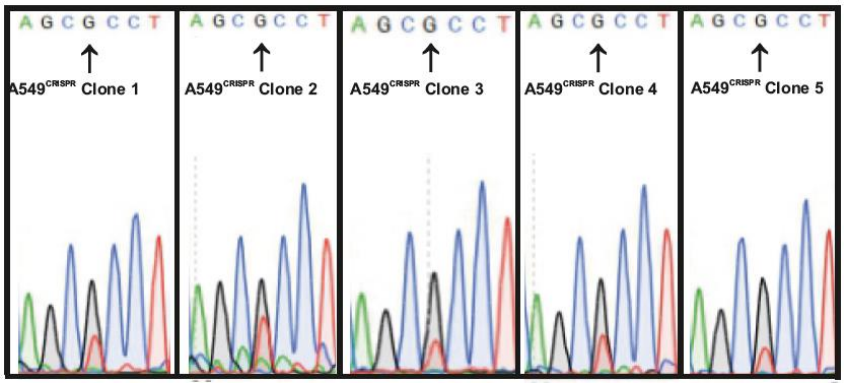


Figure 12. Sequence of A549^{CRISPR} clones (1-5). CRISPR-Cas9 editing technique was used to replace the WT allele by a T allele in A549 cell line. Five heterozygous clones [G/T] were obtained and the presence of the T allele was corroborated by DNA sequencing.

Both sequences were transiently transfected altogether with a Cas9 effector

vector. Positive clones were selected after 3 days using blasticidin selection. As shown in Figure 12, the resulting pool of cells after selection showed higher MUC5B expression at both RNA (Figure 12, A and B) and protein levels (Figure 12, C).

A single cell cloning step was done using A549^{CRISPR} pooled cells to select 123 clones, prior to sequencing. The 5 most stable heterozygous clones were used in the experiments, with continuous corroboration of the genotype. Figure 13 represents a summary of the sequence of the A549^{CRISPR} clones where the presence of the T allele is appreciable. It is noteworthy that none of the clones resulted in a T/T genotype. A possible explanation may be the characteristic hypotriploidity of A549 cell line for chromosome 11, which allows three different genotypes after CRISPR editing. Full sequence of A549^{CRISPR} clones can be found in the manuscript included with this thesis (*Paper I*).

The same single cell cloning step was carried out in BCi_NS1.1 cell line. Despite selection of several possitive colonies, none were able to survive the single cell cloning step. Due to this, a pool of BCi_NS1.1^{CRISPR} edited cells was used in the experiments with the BCi_NS1.1 cell line. The pool of cells was shown to carry the T-allele as confirmed with DNA sequencing (Figure 14).

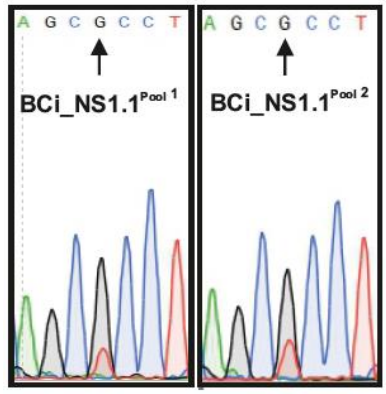


Figure 14. BCi_NS1.1 CRISPR cells were sequenced to quarantee the presence of the T allele.

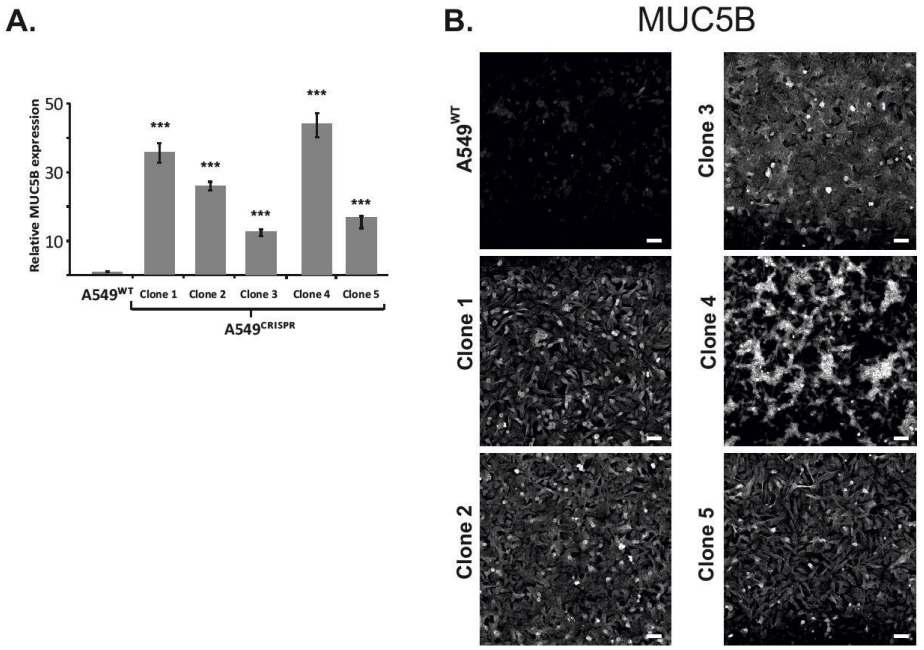


Figure 15. Risk allele (T) is associated with a higher expression of MUC5B. A) WT allele was replaced with the rs35705950 T variant in A549 cells by CRISPR to get five heterozygous cell lines [G/T]. RT-qPCR shows relative expression of MUC5B mRNA in different clones. Culture was performed on monolayer, 48h before RNA extraction. B) IF staining on wild type and A549^{CRISPR} clones (1-5) showing higher MUC5B expression levels compared to A549^{WT} (*p<0.05, **p <0.01 and ***p<0.001with error bars representing SD). Scale bars = 50µm.

A549^{CRISPR} edited clones were seed in monolayer and MUC5B expression analysed by RT-qPCR and IF staining. Our results show between 15 to 45-fold increased MUC5B RNA expression in A549^{CRISPR} edited clones when compared to A549^{WT} cell line (Figure 15A). Also, a marked increased was observed at the protein level in all the clones, analysed by IF-staining (Figure 15B). Even if the results are not homogeneous between the different clones, all of them support a gain-of-function role of the T allele when compared to A549^{WT}.

4.2 *MUC5B* is partially regulated by CpG-DNA methylation.

Methylation has been reported as an important mechanism regulating mucin expression. Furthermore, rs35705950 T allele disrups a predicted CpG site when present instead of the WT allele. The differences observed in Figure 11A and Figure 11C especially in the VA-10 and A549 cells where the only modification was the integration of the 4.1kb *cis*-regulatory region into the genome might also suggest epigenetic regulation. For this reason, we decided to further analysed the role of DNA methylation in the regulation of MUC5B expression, specifically analyzing if there is any difference between the WT and T alleles.

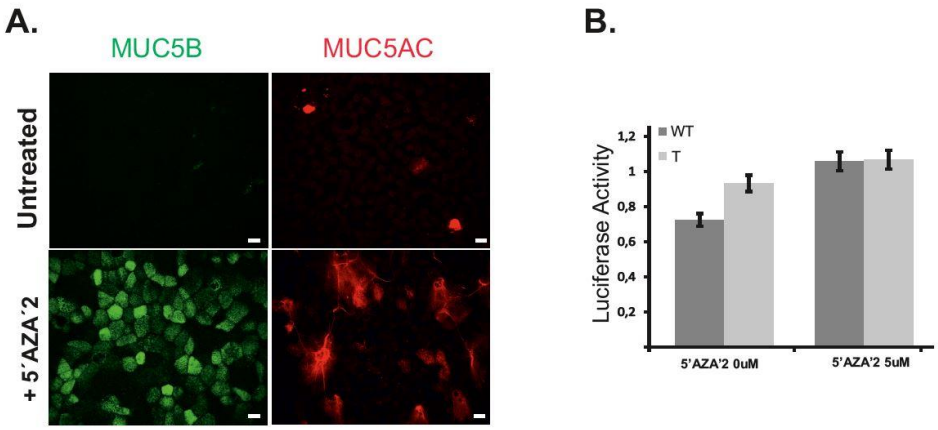


Figure 16. De-methylation increases MUC5B and MUC5AC expression. A) IF staining of BCi_NS1.1 cells with antibodies against MUC5B (left) and MUC5AC (right) after 48h with 5'AZA'2 treatment compared with untreated samples. Scale bar = 10µm. B) Results of transcription activation assays performed in the A549 cell line transduced with pGF-luciferase Lentivector and stimulated with 5‘AZA‘2. Increased luciferase activity was seen in both alleles after the treatment.

To corroborate the influence of methylation on mucin expression, an ALI culture was done using the BCi_NS1.1 cell line, while treating the cells with 5’AZA2’, a well-known DNA demethylation compound (Figure 16A). 5’AZA2’ concentration was optimized as 25µm in a 72hr experiment. In Figure 16A, MUC5B and MUC5AC expression was increased under 5’AZA2’ treatment, suggesting an inhibitory role of methylation on mucin expression. While the overexpression observed in MUC5B is seen in most of the cells, the increased expression of MUC5AC is only appreciable in some of the cells, suggesting a specifc effect of 5’AZA2’ on MUC5AC in a subpopulation of cells, e.g. the goblet cells. On the other hand, MUC5B overexpression under 5’AZA2’ treatment was seen in most cells although a different level of mucin expression

was seen between cells. The same experiment was carried on monolayer-seeded BCi_NS1.1 cells, where no significant differences were seen between 5'AZA2' treated and non-treated cells (data not shown) suggesting an exclusive effect of methylation in MUC5B expression after ephitelial cell differentiation. When we used the A549 stable cell line, transduced with the 4.1kb *MUC5B cis*-regulatory region under the stimulus of 5'AZA'2, an increased luciferase activity was seen in both alleles (Figure 16B). The untreated cell lines show the differential *MUC5B cis*-regulatory activity seen before (Figure 11C). However, when the de-methylation compound was added, the difference was not significant, mainly due to the increased luciferase activity in WT *cis*-regulatory domain. These results suggest an inhibitory role of methylation in the *MUC5B cis*-regulatory region and demonstrate higher sensitivity of WT *cis*-regulatory region to de-methylation when compared to T allele.

4.3 The rs35705950 T allele disrupts a repressive CpG DNA methylation site.

Whether or not the repressive effect of methylation is due to the presence of specific CpG islands in the MUC5B *cis*-regulatory region needs to be further analysed. To study the methylation pattern in our cell models, bisulfite sequencing was performed. DNA was extracted from each cell line and 800bp region surrounding the rs35705950 polymorphism was amplified. A bisulfite conversion was done to trace the methylated cytosines through CpG islands in the region amplified.

Ten different CpG sites were localized in 800bp surrounding rs35705950 polymorphism upstream of the MUC5B gene. Using *Sequence scanner V1.0*. DNA methylation level was scored as percentage methylation of individual CpG units in each sample. In our cell lines, previously sequenced as a WT genotype, rs35705950 remains mainly methylated, suggesting an important role for CpG methylation in this position. In VA10 and A549 the methylation

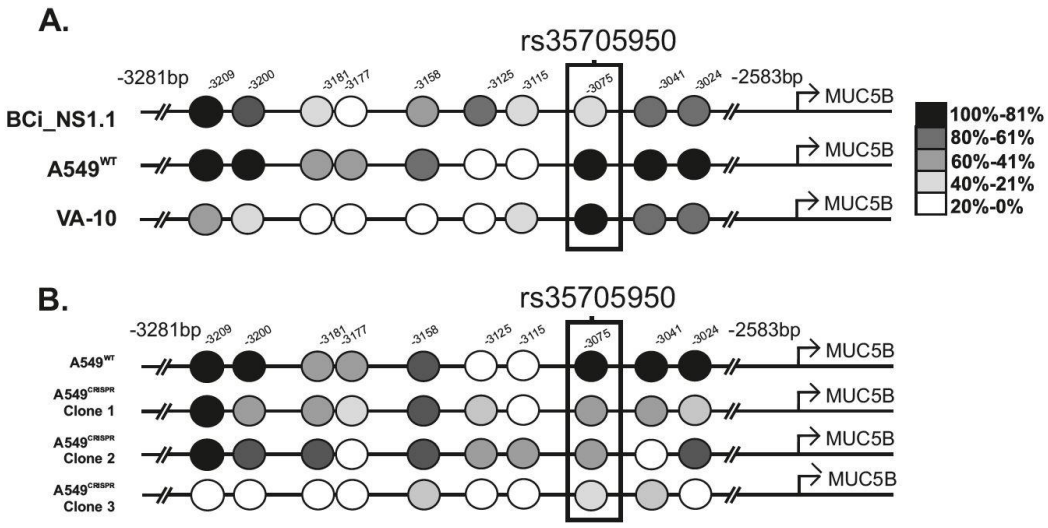


Figure 17. CpG methylation plays an important role in *MUC5B* regulation. A line drawing indicating the regulatory region of the MUC5B gene. Nucleotide positions are indicated with numbers. Circles indicate CpG islands where the level of shading indicates the level of methylation (indicated in inset). A) Results of bisulfite sequencing in BCi_NS1.1, A549 and VA10 cell lines. B) Results of bisulfite sequencing in A549 parental cells and in the different CRISPR clones. The *rs35705950* region is boxed.

ratio was around 100-81% while in BCi_NS1.1 it was 40-21% (Figure 17A). These results might explain the higher MUC5B basal expression shown in BCi_NS1.1 cell line compared to VA10 and A549. Also, downstream region of the polymorphism seems to be highly methylated in the three cell lines.

The same experiment was performed using the wild-type and three $A549^{CRISPR}$ clones (Figure 17B) Here we observed a partial demethylation in rs35705950 due to the presence of the T allele, when compared to $A549^{WT}$ cell line. The reason why we don't see a complete de-methylation of rs35705950 in $A549^{CRISPR}$ clones is probably the presence of a WT allele in the sequence, as the clones are heterozygous (G/T genotype). The downstream region of the polymorphism is also partially de-methylated, suggesting a significant impact of the T allele in the methylation pattern.

An *in vitro* methylation assay was done to directly test whether the disruption of a repressive CpG site in the rs35705950 position due to T allele would affect *MUC5B* expression. The 4.1kb *MUC5B cis*-regulatory insert was cut out from the transient luciferase reporter vector, and DNA CpG methylation was induced *in vitro* by *M.Sss1* methyltransferase. The methylated insert was

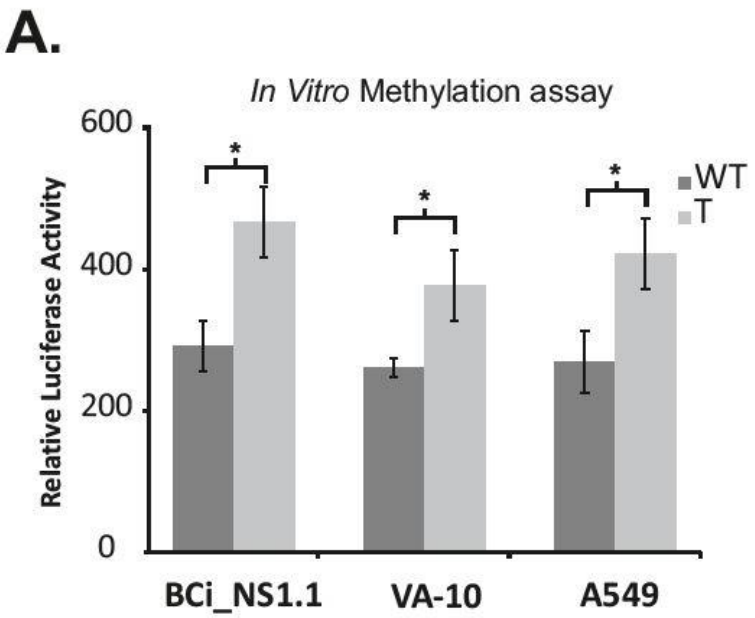


Figure 18. The diferential methylation on MUC5B *cis*-regulatory region increases luciferase activity in T carrier luciferase vector. A) 4.1kb MUC5B *cis*-regulatory region was methylated by a methyltransferase and ligated to a pGL3 luciferase reporter vector. Luciferase activity was measured comparing both alleles, WT and T. (*p<0.05 and **p <0.01 with error bars representing SD).

religated into the vector and directly transfected into the three cell lines. In Figure 18A, a significantly increased transcription activity was seen in all three cell lines in presence of the T allele after forced *in vitro* DNA methylation. This experiment is similar to Figure 11A, except for the direct *in vitro* methylation performed before the transcription activation. Comparing both experiments, the increased luciferase activity was only appreciable in BCi_NS1.1 in Figure 11A while in Figure 18A all the three cell lines shows an increased activity in presence of the T allele, thus suggesting a direct role of DNA-methylation in T mediated MUC5B overexpression.

4.4 The rs35705950 polymorphism modifies binding motifs in the *MUC5B cis*-regulatory domain.

Even though our data support a role of differential DNA methylation in explaining the effects of the T allele on MUC5B expression, the results seen in BCi_NS1.1 cells transiently transfected with the luciferase reporter vector (Figure 11A) would suggest additional effectors modulating the activity, such as a novel transcription factor binding site.

To analyse different binding motifs, the sequences with both alleles were compared using *Match*, a weight matrix-based program for predicting transcription factor binding sites in DNA sequences using the DNA flanking rs35705950 polymorphism. The program uses a library of positional weight matrices from *TRANSFAC® Public 6.0*. Our results indicate two main differences in binding motifs between alleles. When the T allele replaces the G allele, a novel C/EBP (*CCAAT/enhancer-binding protein*) binding motif is created in the rs35705950 position, and a PAX4 (*Paired Box 4*) binding motif disappears (Figure 19A).

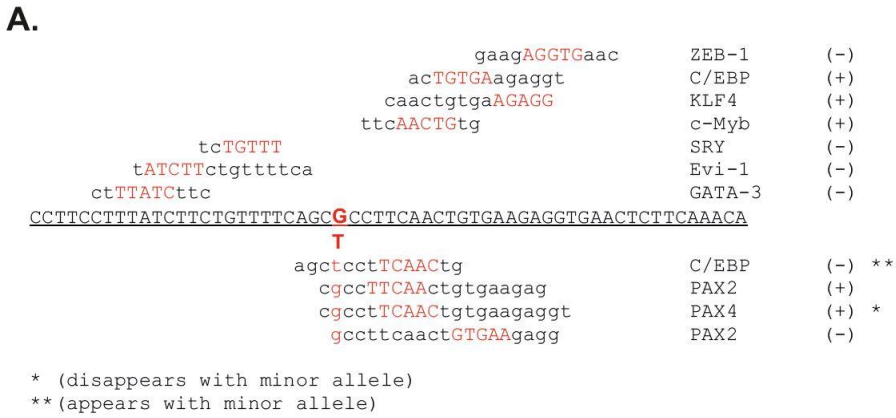


Figure 19. C/EBP is predicted to bind rs35705950 MUC5B *cis*-regulatory region only in presence of the T allele. A) Results of a weight matrix-based program (Match in TRANSFAC) for predicting transcription factor binding sites in the MUC5B *cis*-regulatory sequence using the DNA sequence flanking rs35705950 . (**) indicates a novel binding motif in the T-variant allele. (*) indicates the loss of a transcription factor binding motif, while (+) and (-) indicate the strand where the binding occurs.

4.5 Analysis of C/EBPβ as a mediator of MUC5B overexpression in presence of the T allele.

Comparing the TF candidates C/EBP and PAX4, several lines of evidence support C/EBP over PAX4 in potentially explaining the effect of the T-allele on MUC5B expression. PAX4 is absent from the lung (Figure 20A) as determined using the GTEx database. In contrast, C/EBPβ, a member of the C/EBP family is highly expressed in the lung (Figure 20B), based on GTEx analysis, and has been previously linked to inflammation and differentiation signalling pathways (Chinery et al. 1997, Vanoni et al. 2017), both related with IPF. C/EBP is

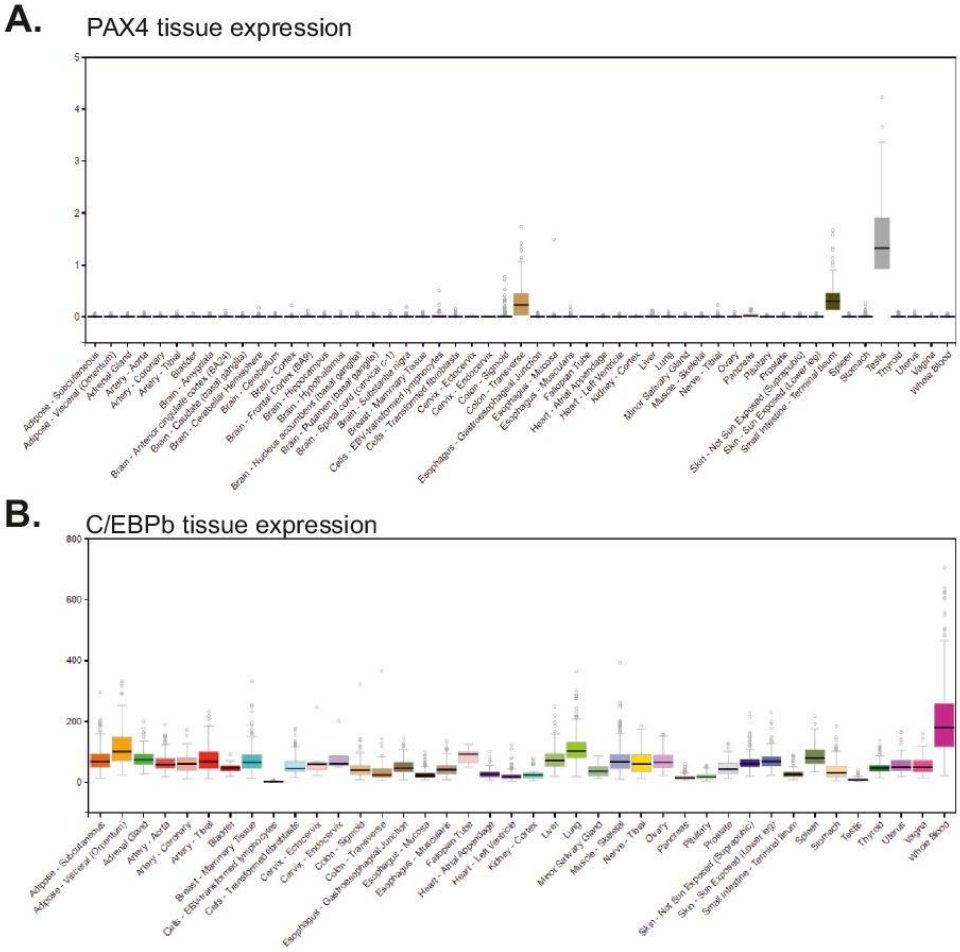


Figure 20. GTEx expression of PAX4 and C/EBPβ in human organs. A) PAX4 is mainly express in testis, small intestine and colon. B) C/EBPβ is express in almost all cell types and organs but its expression is higher in blood, adipose tissue and lung.

therefore a good candidate for a transcription factor that differentially regulates MUC5B based on the rs35705950 variant.

4.6 Co-expression of MUC5B and C/EBPβ in healthy and fibrotic lung.

To further analyse C/EBPβ as a candidate transcription factor mediating MUC5B overexpression through the T allele, we analysed the expression pattern of MUC5B and C/EBPβ in IPF lung samples by imunofluorescense of

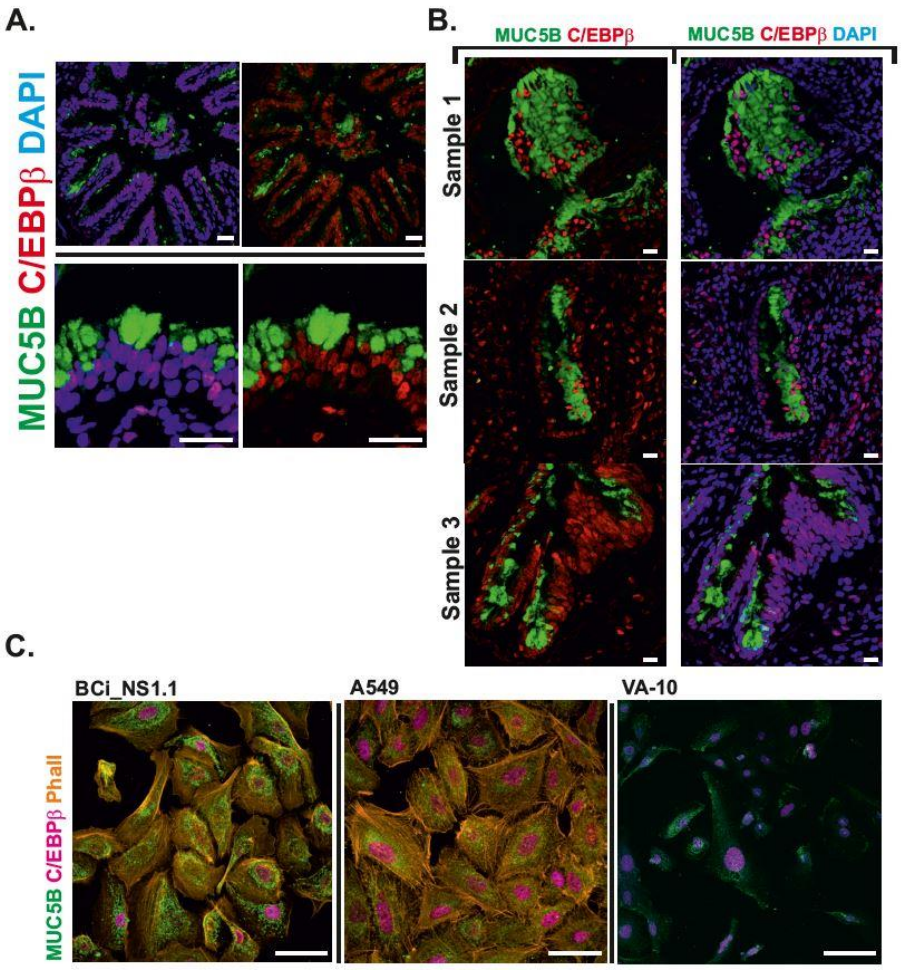


Figure 21. C/EBPβ and MUC5B are co-expressed in IPF samples and in BCi_NS1.1, A549 and VA10 cell lines. A) An IF-P was done in IPF human samples. Four different patientes samples were analysed (A-B) showing a co-expression of C/EBPβ and MUC5B in human lung ephitelial cells. C) C/EBPβ is co-expressed with MUC5B in the three cell lines used as an *in vitro* model (BCi_NS1.1, A549 and VA10). Scale bars=50μm.

paraffin-embedded IPF lung tissue samples (Figure 21A, 21B). In the figures, we can see that both proteins are expressed in lung epithelia. While MUC5B is mainly expressed in the pseudostratified epithelium, C/EBPβ is not only present in the substratified epithelium but also in the interstitial tissue. As expected, C/EBPβ is present in the nucleus while MUC5B is in the cytoplasm.

We corroborated the co-expression of both MUC5B and C/EBPβ in our cell lines *in vitro* (Figure 21C). Our three cell lines were seeded in monolayer at 90% confluence and MUC5B and C/EBPβ was detected by inmunofluorescense. The results show higher expression of C/EBPβ in A549 as compared to BCi_NS1.1 and VA10, but MUC5B expression seems to be higher in BCi_NS1.1 cell line. In summary, both in lung tissue and in culture, cells that express MUC5B also co-express the transcription C/EBPβ.

4.7 A shared signalling pathway through IL-13 signaling.

In addition to the co-expression, we wanted to test whether C/EBPβ might be regulated through a common signalling pathway driving MUC5B expression. Interleukin-13 (IL-13) has been shown to induce goblet cell differentiation and hyperplasia *in vitro*, and it has been used before in our lab to overexpress MUC5AC (Arason et al. 2014). We cultured both, VA10 and BCi_NS1.1 cell lines under ALI culture conditions with or without IL-13 stimulus for 15 days. In both cell lines, IL-13 induced C/EBPβ as well as MUC5B expression at the mRNA (Figure 22A, 22B) and protein (Figure 22C) levels, suggesting a co-regulatory pathway. Compared to 5'AZA'2 stimulus, IL-13 seems to affect MUC5B expression in a homogeneous pattern in all cells in the ephitelium, suggesting a high sensitivity of all ephitilial cells to IL-13. As IL-13 increases C/EBPβ in ephitelial cells, and IL-13 is overexpressed under inflammatory conditions it may connect MUC5B overexpression to IPF processes and link injury-repair processes to IPF.

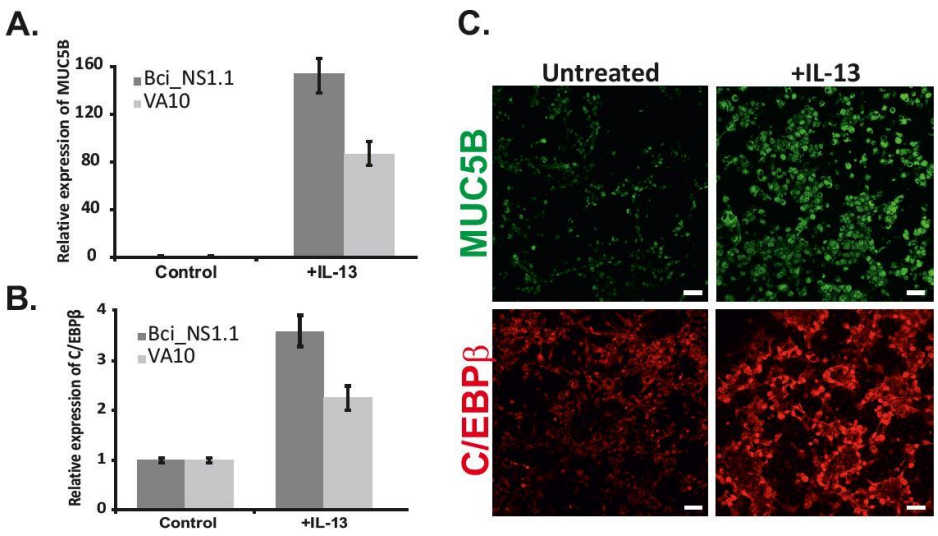


Figure 22. MUC5B and C/EBPβ share a common signaling pathway through IL-13 stimulus. BCi_NS1.1 and VA10 cell lines cultured under ALI conditions induce MUC5B (A) and C/EBPβ (B) expression under IL-13 stimulus at the mRNA level compared to untreated samples. C) MUC5B (green) and C/EBPβ (red) expression in the absence and presence of IL-13 as determined by IF staining. Scale bar=50µm.

4.8 Overexpression of C/EBPβ increases MUC5B overexpression only through T allele.

C/EBPβ is a bZIP transcription factor, encoded from a gene without introns on chromosome 20. C/EBPβ has three co-expressed isoforms (Figure 23) (Yeh et al. 1995, Zhu et al. 2002). Two of them, named LAP and LAP* (Liver-Enriched Transcriptional Activator Protein) are transcriptionally active due to the presence of the transactivatory domain, while the LIP isoform (Liver-Enriched Transcriptional Inhibitor Protein) is a shorter version of LAP and LAP* isoform, lacking the transactivatory domain, explaining why LIP isoform is inhibitory with respect to transcription. It has been previously reported that C/EBPβ isoforms can act as homo- or heterodimers, interacting with the differents isoforms of C/EBPβ or other members of C/EBP family (e.g. C/EBPα, C/EBPζ). The dimers can

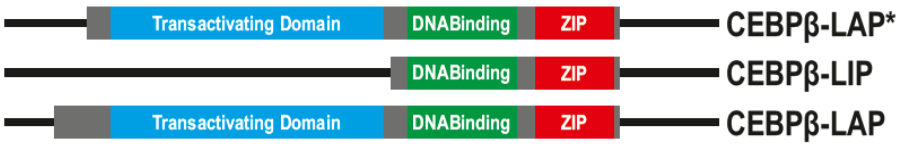


Figure 23. Schematic representation of C/EBPβ isoforms. There are three major isoforms of C/EBPβ transcription factor generated by alternative splicing. Liver-Enriched Transcriptional Activator Protein or LAP isoforms (LAP and LAP*) are related to an activatory role due to the presence of a transactivating domain. The Liver-Enriched Transcriptional Inhibitor Protein or LIP is related to an inhibitory role because of the lack of transactivating domain.

bind to promoters and enhancer regions, inducing gene expression. Inflammatory cytokines such as interleukins 1 (IL-1) (Armstrong et al. 2009) and 6 (IL-6) (Xiao et al. 2004) are among the genes induced by C/EBPβ.

To further corroborate a possible role of C/EBPβ in MUC5B overexpression through the T allele, the two major C/EBPβ isoforms were transiently transfected into our luciferase stable cell lines carrying the 4.1kb *MUC5B* promoter insert. Luciferase measurements were done 48 hrs after transfection. In our results (Figure 24A, 24B), a significant increase of luciferase activity was seen (~56% in BCi_NS1.1 and ~16% in A549) when the activating isoform was transfected (LAP*) in the T-allele carrying luciferase construct. Similarly, inhibition of luciferase activity was observed in the T-allele luciferase construct when the inhibitory (LIP) isoform was overexpressed (~21% in BCi_NS1.1 and ~14% in A549). When both isoforms were overexpressed in equimolar concentration, an intermediate differential luciferase activity was seen in the T-allele constructs. These results suggest that both isoforms can specifically affect the T-allele *cis*-regulatory region in our cellular models. These data were corroborated in BCi_NS1.1 and A549 cell lines, while in VA10 cells (Figure 24C) the tendency was similar but was not statistically significant. The effect of the LAP isoform was furthermore dose-dependent, as shown in Figure 24D.

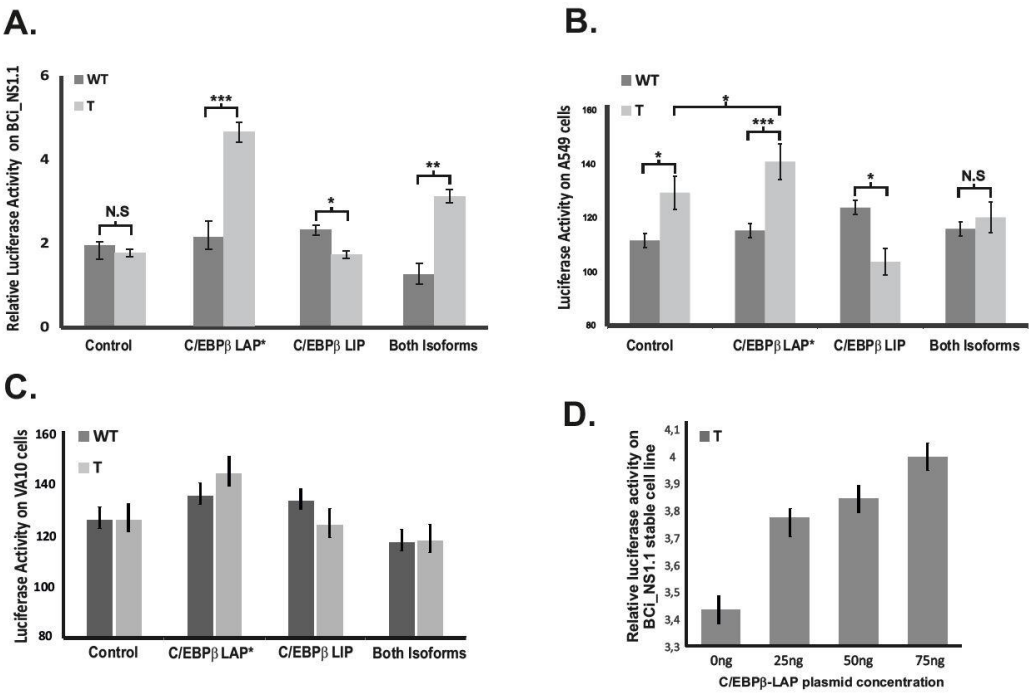


Figure 24. Overexpression of C/EBPβ LAP isoform induces a higher MUC5B *cis*-regulatory domain activity. The two major C/EBPβ isoform were transiently transfected in BCi_NS1.1 (A), A549 (B), and VA10 (C) luciferase stable cell lines together and separately. An increased luciferase activity was seen when LAP isoform was transfected while an inhibition of luciferase activity is seen when LIP isoform is transfected. An intermediate effect is appreciable when both isoform are transfected in equimolar concentration. D) C/EBPβ-LAP transient vector was transfect in Bci_NS1.1 cell line in increasing concentration to show a positive dose-dependent correlation of LAP isoform with T allele in BCi_NS1.1 luciferase stable cell line (*p<0.05, **p <0.01 and ***p<0.001 with error bars representing SD). N.S=Non-significant.

4.9 Inhibition of C/EBPβ restores basal MUC5B expression in T allele models.

Our results above show that C/EBPβ overexpression increases MUC5B expression, specifically through the T allele. Also, we have seen that this effect is seen with the activating isoform, while the inhibitory isoform specifically reduces MUC5B expression. To further corroborate the possible role of C/EBPβ and its different isoforms in MUC5B overexpression, we knocked-down C/EBPβ using siRNA. The transactivating domain can be used to specifically inhibit the LAP isoform but the LIP isoform lacks specific sequence to target and thus the LIP isoform can not be specifically inhibited without inhibition of LAP isoform at the same time. For this reason, a knock-down of C/EBPβ was done with 4 different siRNAs, targeting only LAP* isoform, or both isoforms using $A549^{CRISPR}$ clones compared to $A549^{WT}$ (Figure 25A) and the A549 luciferase stable cell line carrying the 4.1kb *cis*-regulatory insert (Figure 25C). Both mRNA and luciferase activity were measured 48hs after transfection. In both experiments, knockdown of C/EBPβ inhibited specifically MUC5B expression in cells carrying the T-allele, restoring the MUC5B levels to the same levels as seen in the WT G-allele. Furthermore, MUC5B expression was slightly higher when both isoforms (LAP and LIP) were knocked down when compared to exclusive knock down of the LAP isoform. Specificity of the different C/EBPβ siRNAs was measured by RT-qPCR (Figure 25B) and western blot (Figure 25D) showing efficient C/EBPβ knockdown.

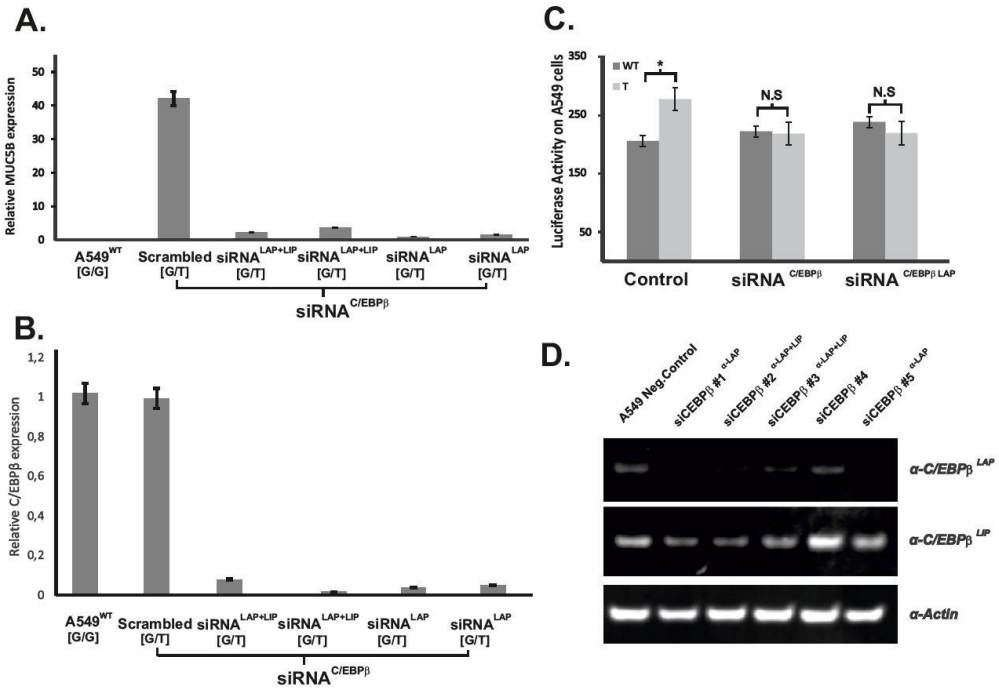


Figure 25. Inhibition of C/EBPβ isoforms by siRNA restores basal MUC5B expression and luciferase activity. A) A549^{CRISPR} clone was used to inhibitit C/EBPβ expression in comparison to A549^{WT} cell line. A reduced MUC5B expression was detected by RT-qPCR. B) C/EBPβ knock-down was analysed by RT-qPCR. C) A549 luciferase stable cell lines were used to analysed the effect of C/EBPβ knock-down in luciferase expression. The inhibition of C/EBPβ restores luciferase activity to basal expression in the cell line carrying the T allele. D) Inhibition of C/EBPβ was analized by western blot. siRNA#4 was not used in any experiment (*p<0.05 and **p <0.01 with error bars representing SD).

4.10 C/EBPβ regulates MUC5B overexpression in T genotype cells through direct binding to the MUC5B *cis*-regulatory domain.

Even though our experiments corroborate an important role of C/EBPβ in mediating MUC5B overexpression through the T allele, whether or not the transcriptional regulation is carried out by direct binding or indirect signaling pathway activation has not been shown. To further analyze the mechanism through which C/EBPβ mediates MUC5B overexpression, a chromatin immunoprecipitation (ChIP) assay focusing on the *cis*-regulatory region containing the T-allele was performed. We compared direct binding of C/EBPβ for the A549^{WT} (G/G genotype) and A549^{CRISPR} clone (G/T genotype). For this experiment, a *TNFAIP3* promoter region was used as a positive control, focusing on a region where C/EBPβ is known to bind. As a negative control, a region of the *Actin* promoter was used which is predicted not to bind C/EBPβ. The negative control for the antibody was a non-specific IgG Isotype.

As shown in Figure 26 there is a marked difference in the binding of C/EBPβ to the region of the rs35705950 polymorhism in WT vs T allele. No difference was seen in the positive control region (Figure 26). These results indicate that C/EBPβ specifically binds to the T allele in the rs35705950 region. The results of one of three ChIP-experiments are shown in Figure 26 while the results of other two experiments can be found in the supplement of the manuscript included with the thesis (*Paper I*).

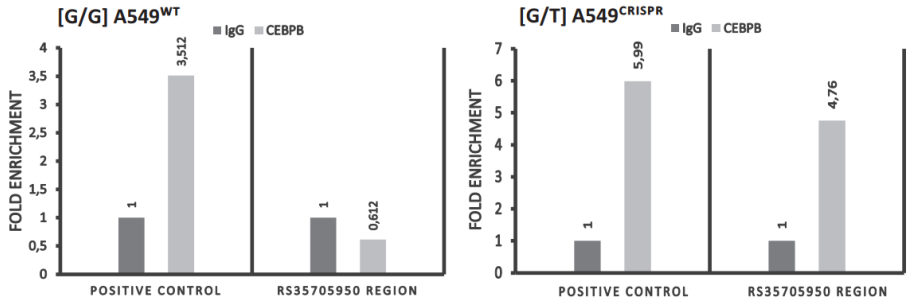


Figure 26. C/EBPβ regulates MUC5B overexpression in T genotype cells through direct binding to MUC5B *cis*-regulatory domain. A) Chromatin Immunoprecipitation assay shows partial binding of C/EBPβ to the MUC5B cis-regulatory region to [G/T] genotype A549^{CRIPSR} cell line but not to [G/G] A549^{WT} genotype. Dark grey represents the immunoprecipitation with NS*IgG Isotype, used as an antibody control. Light grey represents inmunoprecipitacion with C/EBPβ. *TNFAIP3* was used as a positive control.

4.11 Rs35705950 T allele induces morphological changes in A549^{CRISPR} cells.

The results shown above support the gain-of-function role of the rs35705950 T allele and that the overexpression of MUC5B can be contributed to a combined effect of a lost of a CpG-site and gain of a positive transcription site for C/EBPβ. However, little is known about the cellular effects of MUC5B overexpression in airway epithelial cells. The data presented in this part of the thesis is not included in *Paper I*.

To analyze the effects of MUC5B overexpression, the A549^{CRISPR} cell line (G/T genotype) was compared to A549^{WT} (G/G genotype) cell line, transfected only with Cas9 expression vector, in absence of gRNA, used as control for transfection-induced cellular stress. Four different transfected cell lines per each genotype were seeded on coverslips and fixed for transmission electronic microscopy. Representative images of A549$^{WT\text{-}Cas9}$ control cells compared to A549^{CRISPR}, are shown in Figure 27. When compared to the control cells, the A549^{CRISPR} cell line shows major changes in morphology. In a high proportion of A549^{CRISPR} cells (G/T genotype) there are many (>100) small and round vesicles in the cytoplasm that are absent in the A549$^{WT\text{-}Cas9}$ control cell line. Based on these differences and the high expression of MUC5B in the A549^{CRISPR} cells we hypothesized that these vesicles are MUC5B containing.

Another feature representative of the morphology observed in the A549^{CRISPR} cells is an increased number of mitochondria. Figure 27 shows that the vesicles are often surrounded by numerous mitochondria. Interestingly, it has been previously reported that AECIIs from human IPF lungs have accumulation of dysmorphic and dysfunctional mitochondria associated with upregulation of ER stress markers (Mora et al. 2017). Considering the large size of MUC5B and its complex posttranslational modifications, including extensive glycosylation, we hypothesize that the induction of MUC5B overexpression could lead to ER stress caused by the complex post-translational modifications of MUC5B and the lack of proper machinery to secrete MUC5B that is consequently retained in the cytoplasm.

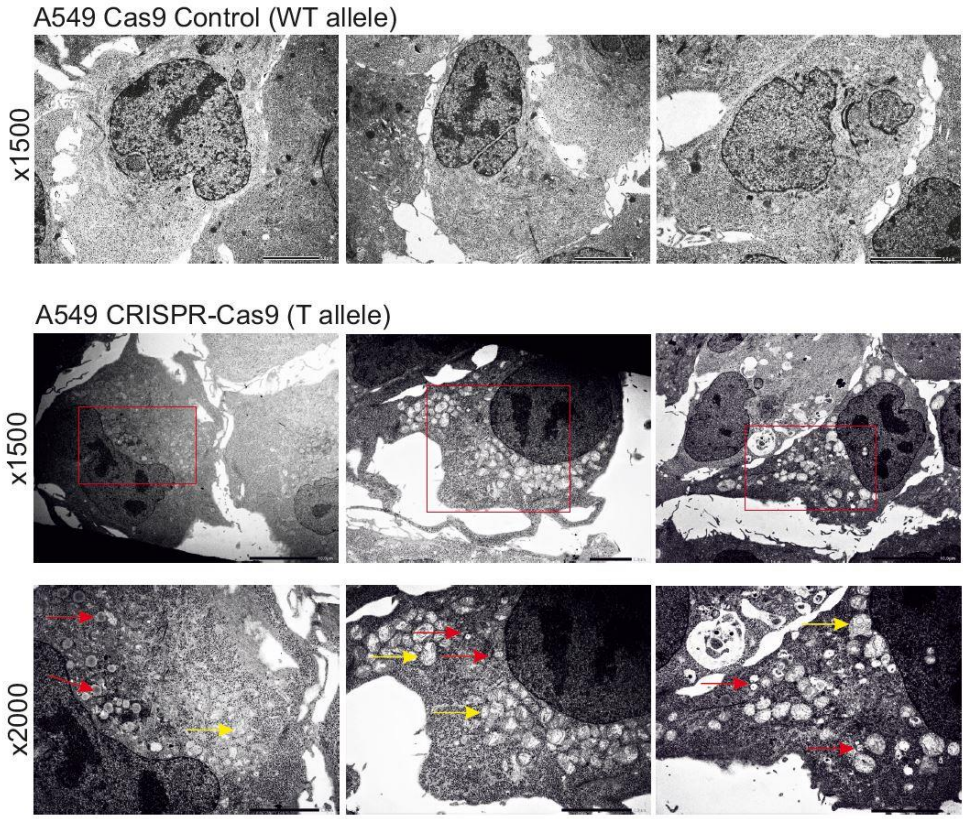


Figure 27. A549^{CRISPR} cell line presents an altered morphology under EM, related to ER stress. A549 cell line was transfected with Cas9 transient expression vector alone (A549^{Cas9} Control) or together with the gRNA to mediates the genotype editing. As a results, A549^{CRISPR} cells were generated, with a [G/T] genotype compared to [G/G] genotype in control. A triplicate was seed and fixed for electronic microscopy, where numerous vesicles was visible as well as increased number of mitochondria compared to the control. In A549^{CRISPR} cell line, the upper row represents images at x1500 amplification is shown. At the bottom row, a higher magnification (x2000) of the region inside the red square above are present for a better visualization of the vesicles and mitochondria described. Red arrow points to vesicles, while yellow arrow indicates mitocondria position.

4.12 The rs35705950 T allele increases the number of mitochondria in A549^{CRISPR} cells.

To corroborate the finding from the TEM showing increased number of mitochondria in T-allele cells, a specific mitochondrial stain, the so called MitoTracker reagent, was used to quantify mitochondria in the cells. A549 and BCi_NS1.1 WT-Cas9 controls and CRISPR cell pools were seed in a monolayer and stained with 50nM of MitoTracker for 30 minutes prior to fixation.

By imunofluorescense imaging, a marked increase in the staining intensity was seen indicating an increase in the number of mitochondria in the T carrier cell lines. A representative picture is shown in Figure 28.

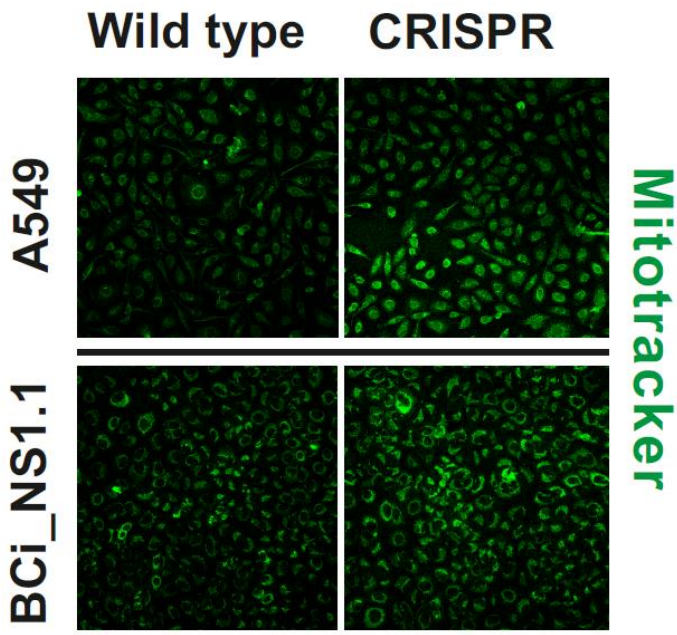


Figure 28. The presence of rs35705950 T allele increases the mitochondria number in A549 and BCi_NS1.1 cell lines. A z-stack IF picture was taken to quantify the mitochondria number in A549 and BCi_NS1.1 Cas9 control and CRISPR cells.

4.13 The rs35705950 T allele reduces proliferation in A549 and BCi_NS1.1 CRISPR cells.

To test the effects of the rs35705950 T allele on cell viability, a proliferation assay was done comparing A549 and BCi_NS1.1 WT-Cas9 control cells to the CRISPR cells carrying the IPF associated T allele. For this experiment, cells were seeded in equal numbers and confluence was tracked using IncuCyte® S3 Live-Cell Analysis System over 3 to 4 days. Our results show that in both cell lines the wild type (G/G genotype) cell lines had significantly higher confluency over time compared to the T carrier cell lines (G/T genotype) indicating faster proliferation and/or higher survival.

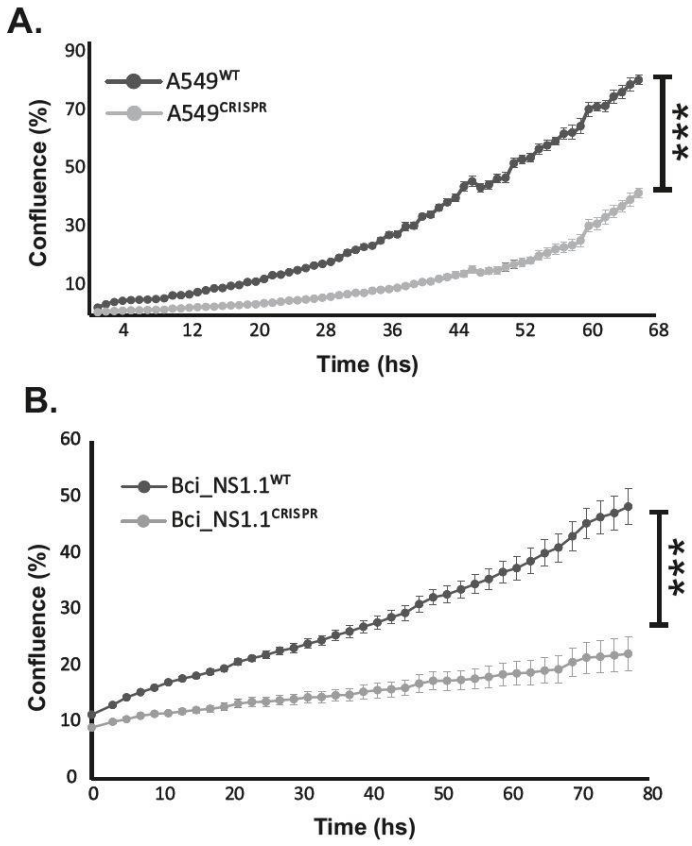


Figure 29. Cell lines carrying the wild type allele proliferates faster than T allele carriers. A549 (A) and Bci_NS1.1 (B) cell lines carrying the WT or the T allele were compared in a proliferation assay. In both cell lines, the wild type cell line shows an increased proliferation ratio when compared to the T allele carriers. The data are represented in percentage of confluence over time. Error bars represent SD.

4.14 The rs35705950 T allele increases ER stress markers in A549 and BCi_NS1.1 CRISPR cells.

The described characteristics of the A549^{CRISPR} cell line, i.e. the increased mitochondrial number, and the complex post-translational nature of the highly expressed MUC5B shares similarities with the effects observed in individuals carrying mutations in the surfactant genes *SFTPC* and *SFTPA2*, that has been shown to lead to ER stress in the airway epithelium. To address whether MUC5B overexpression might lead to ER stress we quantified one of the major ER stress-related markers, the C/EBPζ transcription factor, usually known as C/EBP homologous protein (CHOP). CHOP is a member of the C/EBP family, known to be increased in expression and to accumulate in the nucleus during apoptosis induced by ER stress. This nuclear accumulation furthermore leads to growth arrest (Li et al. 2014).

In Figure 30B, we show that the presence of the T allele increases CHOP mRNA levels (7-fold increase) in A549^{CRISPR} cells compared to A549$^{WT\text{-}Cas9}$ control, suggesting an increase in the pro-apoptotic signaling pathway associated with ER stress.

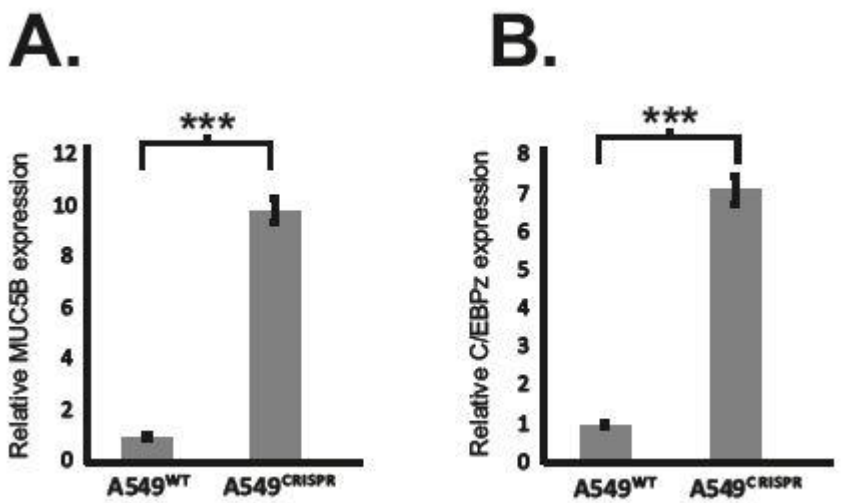


Figure 30. CHOP is overexpress in A549^{CRISPR} cells compared to A549^{WT}. A) The A549^{CRISPR} cell line, carrying the rs35705950 T allele shows increased MUC5B expression compared to A549^{WT}. B) The A549^{CRISPR} cell line shows higher expression of the ER stress related pro-apoptotic transcription factor CHOP, when compared to A549^{WT}. Error bars represent SD.

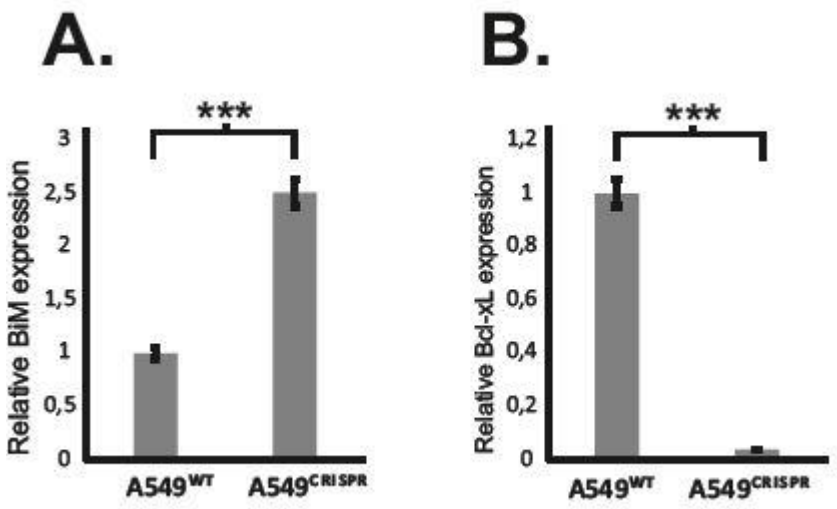


Figure 31. BiM pro-apoptotic protein and Bcl-xL anti-apoptotic protein are dysregulated in A549^{CRISPR} cell line. A) BiM expression is upregulated in A549^{CRISPR} cell line compared to control (A549^{WT}), while Bcl-xL (B) is downregulated. Error bars represent SD.

For further understanding of the possible role of ER stress in the A549^{CRISPR} cell line, two other ER-stress related factors were analyzed, namely the pro-apoptotic BiM (Bcl-2-like protein 11, commonly called BiM) protein (O'Connor et al. 1998) and the anti-apoptotic B-cell lymphoma-extra large (Bcl-xL) (Gaudette et al. 2014) a mitochondrial transmembrane protein. As we can see in Figure 31, the presence of the T allele induced the pro-apoptotic protein BiM (Figure 31A) and inhibits Bcl-xL (Figure 31B), related to cell survival.

Althogether, these findings point to the rs35705950 T allele as a potential inducer of ER stress, likely through MUC5B overexpression that can lead to apoptosis through an ER stress cascade mediated by CHOP and known apoptosis regulators (BiM and Bcl-xL).

Nevertheless, further analysis is required to directly address the role of rs35705950 T allele and MUC5B as inducers of ER stress and subsequent apoptosis.

5 Aspects on methodology and experimental approach

I will shortly address problems I encountered related to methodological or experimental approaches in my Ph.D. work.

CRISPR-Cas9

The novel CRISPR-Cas9 editing technique was used in this project. However, different challenges had to be solved to use this technique in the available cell models I used. At the beginning of this project, BCi_NS1.1 cell line was intended to be used as the main cell model and VA-10 as backup. Nevertheless, BCi_NS1.1 and VA-10 are highly sensitive to the chemical stress associated with lipid transfection, as well as to single cell colony selection, necessary to guarantee a unique genotype cell line after CRISPR-Cas9 editing. Due to this, the A549 cell line was used to replace the original cell models in most part of our experiments, as it is more easily manipulated than the other cell lines.

A549 was successfully used for CRISPR-Cas9 and survived the single cell cloning step. However, one characteristic of this cell line is the hypotriploidity for chromosome 11 which contains the *MUC5B* gene.

In order to overcome the problem with single cell cloning, a pool of BCi_NS1.1 CRISPR cells selected by blasticidin was used to corroborate the effects of the T allele in A549 cells.

ALI culture

The BCi_NS1.1 cell line as well as VA-10 has been used before in ALI cultures. However, A549 is an alveolar derived cell line and is unable to differentiate in ALI culture. For that reason, the effect of the minor T allele was not tested in the ALI culture conditions with $A549^{CRISPR}$ cells.

Another technical limit of the monolayer culture was the lack of effect of IL-13. IL-13 induces goblet cell differentiation of the epithelial cells, which in turn, increase MUC5B expression. In monolayer, the MUC5B overexpression after IL-13 addition was not efficient in all three cell lines.

Western Blot of Mucins

As was described before, mucins are large and complex glycoproteins. Due to this, western blotting of mucin proteins is difficult, and we were not able to get high quality MUC5B Western blots. For this reason, all measurement of MUC5B expression had to be done by immunoflorescense or RT-qPCR.

6 Discussion

IPF is a severe pulmonary disease of unknown aetiology. Among the genetic factors that trigger IPF development is the MUC5B *cis*-regulatory rs35705950 polymorphism, where a G is replaced by a T nucleotide. It is the strongest known IPF risk factor, but the molecular effects of the polymorphism remain poorly understood.

In this thesis we have shown how the presence of the minor T allele located in the rs35705950 position directly increases *MUC5B* expression in our cell models. Furthermore, we show that MUC5B upregulation is mediated through a combination of two different mechanisms, epigenetic and transcriptional regulation. First, the presence of the minor T allele disrupts a CpG methylation site, but DNA methylation is known to repress MUC5B expression. Furthermore, we show that changes in methylation pattern of the *MUC5B cis*-regulatory region is sufficient to increases its expression and that the presence of the minor T allele not only affects rs35705950 methylation site but also the surrounding methylation pattern. Secondly, the results presented show how a new C/EBPβ binding site is created when the minor T allele replaces the wild type allele. Furthermore, our results indicate that C/EBPβ is a regulator of MUC5B and we show that it binds to the rs35705950 region only in presence of the minor allele, increasing MUC5B expression in our cell models.

Interestingly, preliminary results presented in this thesis also suggest an effect of the T allele on cell morphology and viability. First, we have shown the presence of numerous cytoplasmic vesicles in our CRISPR cell line, and an increased number of mitochondria along with a decreased proliferation and or survival. We also show overexpression of one of the major transcription factors associated with cellular stress (CHOP), and the pro-apoptotic protein BiM and the down-regulation of the anti-apoptotic transcription factor Bcl-xL, all possibly due to the excessive MUC5B production. Nevertheless, it should be emphasized that further studies are required to better understand the cellular consequences of the T-allele and its effects on MUC5B overexpression.

6.1 Epigenetic regulation and *MUC5B*.

Mucins *MUC4, MUC5AC, MUC5B* and *MUC6* are clustered on chromosome 11p15. Their promoter shows a GC-rich structure previously characterized by Vincent et al. (2007). They report a specific repressive role of the hypermethylated *MUC5B* promoter at -2615/-2168 position from the

transcription start site (TSS), whereas MUC5AC expression was rarely influenced by epigenetic mechanisms. Furthermore, histone methylation was also shown to inhibit MUC5B expression resulting in complete silencing of its expression. In this thesis we have analized in our cell models by bisulfite sequencing the CpG methylation status of *MUC5B cis*-regulatory region in the 800bp surrounding rs35705950 position, 3kb upstream of the TSS, an area further away from the TSS than previously analyzed by other groups. Our results corroborate the repressive effects of CpG methylation in MUC5B expression based on the effect of 5'-aza-2′-deoxycytidine demethylation compound in our cell models. In accordance with previous results, less effect was seen in MUC5AC expression under similar experimental conditions.

Even if previous studies have indicated CpG methylation to be important in *MUC5B* regulation, the role of rs35705950 polymorphism in the methylation pattern hasn´t been directly analyzed previously. Our results show that the minor T allele destroys a CpG methylation site, and based on the *in vitro* methylation assay, we directly show that the loss of this single CpG site increases MUC5B expression in the luciferase-based promoter assay.

Nevertheless, MUC5B overexpression in T genotype samples may not be explained exclusively by methylation, as we show in Figure 11, where even with no methylation of the promoter, a differential MUC5B expression between alleles is appreciable in the BCi_NS1.1 cell line. This would suggest additional mechanisms to be at play.

6.2 Transcriptional regulation of *MUC5B*.

The rs35705950 polymorphism is located in a DNA region characterized by high DNAse hypersensitivity, that may indicate the binding of transcription factors that regulate *MUC5B*. Our data based on *in silico* binding analysis predicts a C/EBPβ binding site overlapping the rs35705950 position when the IPF associated T allele is present. Furthermore, we shown the positive regulation of *MUC5B* by C/EBPβ specifically through T allele, and its direct binding to the *MUC5B cis*-regulatory region only in the presence of the minor T allele. These data strongly support a role for C/EBPβ in mediating the overexpression of MUC5B in carriers of the minor T-allele. Recently, Helling et al (2017) reported the presence of a highly conserved binding site for a FOXA2 transcription factor, only 32bp upstream of the rs35705950 polymorphism. This finding may suggest an important role of this region in MUC5B expression. The reported FOXA2 binding site however, doesn't overlap the polymorphism, and

no specific effect of the T allele was reported on FOXA2 binding or activity. Due to the short distance between FOXA2 and C/EBPβ binding sites, it could be hypothesized that both act together as a binding complex, promoting *MUC5B* expression. C/EBPβ has been reported to dimerize with other members of the C/EBP family and certain members of the ATF family (Tsukada et al. 2011) and the complexity of C/EBP family regulation opens the possibility of a mechanism involving other transcription factors.

MUC5B expression is complex and the signaling pathways that mediate the effect of environmental factors and disease conditions are poorly understood. It has been shown that activation of MEK1/2, PI3K, SPhk1, and MAPK14 (p38α-MAPK) are implicated in IL-13–induced mucus production (Atherton et al. 2003, Kono et al. 2010). Furthermore, through PMA signalling PKC, EGF/TGF-a, Ras/Raf, Mek, ERK and Sp-1 signalling pathways have also been associated with MUC5AC and MUC5B expression (Hewson et al. 2004). Interestingly, C/EBPβ is also overexpressed after PMA stimulus (Zhu et al. 2002) and, in our results, we show a common signalling pathway between MUC5B and C/EBPβ involving IL-13, suggesting common signalling pathways. As mentioned before, C/EBPβ has been previously associated with inflammation and immune responses (Kinoshita et al. 1992, Chinery et al. 1997, Roy et al. 2002, Pless et al. 2008). Similarly, MUC5B has been implicated in innate immune response in the lung (Roy et al. 2014). These findings point to a similar expression pattern between MUC5B and C/EBPβ, that may connect the IPF development to the T allele carriers under specific environmental conditions.

6.2.1 Characteristics of the C/EBP family of transcription factors.

CCAAT/enhancer (C/EBP) family is a basic leucine zipper (bZIP) transcription factor family tightly regulated and involved in many different processes. There are six different members designated in chronological order of their discovery: C/EBPα, C/EBPβ, C/EBPγ, C/EBPδ, C/EBPε and C/EBPζ (Tsukada et al. 2011). All of them share a highly conserved bZIP domain and are widely expressed. They are linked to metabolism, innate and adaptative immunity, inflammation, hematopoiesis, adipogenesis, osteoclastogenesis, cell cycle, cellular proliferation, differentiation, and exertion of important functions in many physiological processes (Darlington et al. 1998, Cassel et al. 2003, Centrella et al. 2004, Friedman 2007, Smink et al. 2010).

The C/EBP family is tightly regulated. A combination of activator/repressor binding sites on genes, interactions among activators, repressors and their functional modifiers further generate considerable diversity (Tsukada et al. 2011). In addition, post-translational modification provides further mechanisms to regulate transcriptional activity. Phosphorylation of the C/EBP family and different interactions between other transcription factor are involved in C/EBP function. C/EBPβ has been linked to the following genes: C reactive protein; alpha1-acid glycoprotein; a2-macroglobuin; hemopexin; haptoglobin; cytokines such as TNFa, IL-1b, IL-6, IL-8, IL-12, G-CSF, MIP-1a, MIP-1b; and receptors for G-CSF, GM-CSF and M-CSF (Tsukada et al. 2011). Of the different C/EBP members, one is specially interesting. It has been reported that C/EBPζ (also known as CHOP) acts as a negative regulator of other C/EBP family proteins by generating non-functional dimers, thereby inhibiting C/EBP function (Ron et al. 1992). CHOP has been shown to enhance the IL-6 promoter by sequestering the inhibitory isoform of C/EBPβ (LIP) away from activating C/EBPβ isoforms (LAP* and LAP). CHOP preferentially associates with LIP rather than LAP*/LAP (Hattori et al. 2003). Furthermore, CHOP expression is induced in the acute-phase response, adipocytic differentiation and cellular stress, being cellular stress a characteristic observed in our cell models. In our results, we have seen increased expression of CHOP in the $A549^{CRISPR}$ cell line, and a differential response of T genotype cells to the inhibitory (LIP) and the activatory (LAP*/LAP) isoforms of C/EBPβ. The increased expression of CHOP in T-allele carrying cells could further accentuate the positive effects of C/EBPβ-LAP*/LAP expression on MUC5B expression. Based of what has been previously reported, we could thus hypothesize that C/EBPβ LAP*/LAP isoforms are preferentially binding to the T-allele under cellular stress conditions when CHOP expression is increased, due to sequestering of C/EBPβ LIP isoform.

Given the complexity of the mechanism, it is difficult to formulate a simple pathway involved in C/EBPβ expression or to determine the specific environmental condition that triggers MUC5B overexpression in T allele carriers. This will need further studies.

6.3 MUC5B and IPF.

An important hypothesis on the cellular events leading to IPF involves ER stress in the alveolar region with subsequent aberrant wound healing responses to the ongoing lung injury, eventually leading to excesive

extracellular matrix deposition and fibrosis (Zhang et al. 2017). ER stress has been shown to lead to expression of mediators such as TGF-β and a variety of other chemokines (Selman et al. 2001, Wilson et al. 2009, Wolters et al. 2014, Knipe et al. 2015). ER stress also leads to epithelial cell dysfunction or apoptosis, that activates the resident fibroblast to proliferate as part of the injury repair response (Zhang et al. 2017). Epithelial cells also recruit inflammatory cells such as macrophages, that produce cytokines or chemokines that create an environment of fibrosis to induce lung repair (Wynn et al. 2012, Ho et al. 2014). Collectively, persistent injury to lung epithelium induces a cascade of abnormal regulatory mechanisms to cause an airway epithelial apoptosis, continuous fibroblast activation and increased myofibroblast differentiation, which then could lead to excessive ECM deposition and distort lung tissue architecture, resulting in IPF (Harding et al. 1999, Selman et al. 2001, Tashiro et al. 2007, Wilson et al. 2009, Wolters et al. 2014, Knipe et al. 2015). ER stress markers include PERK, ATF6α and IRE1α. When ER occurs, ATF6α is relased from the inactivate state where it is kept through binding to BiP chaperone (Isler et al. 2005, Schroder et al. 2005, Wang et al. 2016). A cascade of ER stress is then activated. Interestingly, C/EBP proteins can form homotypic heterodimers, especially with members of the CREB/ATF family, via the leucine zipper dimerization domain (Newman et al. 2003) greatly expanding the repertoire of DNA-binding specificities and altering functional activities and trans-activation potential. Furthermore, and as stated above, the pro-apoptotic transcription factor CHOP is overexpressed under prolonged ER stress (Metallo et al. 1997, Kohler et al. 1999, Craig et al. 2001, Listman et al. 2005), being activated by all three UPR pathways (Hollenbeck et al. 2000, Newman et al. 2003), resulting in the reduction of anti-apoptotic mitochondrial protein Bcl-xL, activating the pro-apoptotic intention at mitochondria associated with mitochondrial damage (Ron et al. 1992).

Much of the literature on ER stress comes from studies focusing on the mechanism explaining how surfactant mutations lead to IPF (Lawson et al. 2008, Wang et al. 2009). Surfactant proteins are a highly glycosylated protein going through complex post-translational modifications. Interestingly, the role of epithelial ER stress and apoptosis in IPF agree with the phenotypic changes observed in our preliminary investigation of our $A549^{CRISPR}$ cell line. Our results not only show an increased number of mitochondria, but also markedly reduced cell proliferation and/or survival and increased expression of the pro-apoptotic CHOP and BiM and the reduced Bcl-xL related to cell survival.

We can thus speculate that the large size of MUC5B and its glycan-based structure could lead to ER stress when the glycoprotein is markedly overexpressed. This is in accordance with the hypothesis connecting the surfactant protein mutations (SFTPC and SFTPC2) to IPF through ER-stress (Vinson et al. 1993, Tsukada et al. 1994, Hollenbeck et al. 2000, Podust et al. 2001, Gombart et al. 2007) that also involves a similar glycan-based structure. The alveolar cells (AEC) with these mutations are more prone to apoptosis after bleomycin treatment. Interestingly, ER stress has been found to induce EMT, another key feature of IPF (Plevy et al. 1997). While surfactant is expressed in alveoli, MUC5B is expressed in larger airways. If MUC5B leads to epithelial ER stress it remains to be seen where in the airway this occurs. We have recently shown that larger airway epithelium shows hallmarks of injury and partial EMT especially over fibroblastic foci in IPF (Jonsdottir et al. 2015). MUC5B is also expressed in smaller airways and could furthermore be aberrently expressed in distal airways under inflammatory conditions in subjects carrying the T-allele. The observation that MUC5B overexpression due to the T allele induces cellular stress and the role of C/EBP family on the rs35705950 polymorphism provides novel insights into of how the T allele might lead to IPF development.

6.4 Characteristics of the rs35705950 polymorphism.

While thousands of SNPs have been associated with common diseases only a few have led to a molecular understanding of how the variant affects the regulatory networks of gene regulation. This project casts light on the molecular consequences of the most important genetic variant associated with idiopathic pulmonary fibrosis (IPF). Deciphering the molecular effects of non-coding SNP´s has proven to be notoriously difficult, but the *MUC5B* variant presents unique characteristics making it an ideal polymporphic candidate to study. The polymorphism is from a genomic standpoint relatively simple, i.e. it is not in complex linkage equilibrium with other SNPs. Furthermore, it has a high OR (4.5-6.6) suggesting that the effect is fairly strong and thus easier to elucidate. The rs35705950 polymorphism also has a strong eQTL and this eQTL is specific to the lung. These facts highlight the importance of the rs35705950 polymorphism as a risk factor for IPF. Many other polymorphisms are in complex networks of many other polymorphic variants and often the association with a specific gene is not straightforward.

Understanding how non-coding disease associated variants mediate

molecular responses is critical in translating all the plethora of genetic variants associated with most complex diseases. It is likely that epigenetic and transcriptional regulatory regions will play an important role.

7 Conclusions

In this thesis we have used two main models to study the connection between the rs35705950 minor T allele and the correlated MUC5B overexpression; a luciferase reporter vector and CRISPR edited cell lines. In this thesis we show that the T-allele directly mediates increased MUC5B expression. We furthermore decipher the mechanism of how the T allele induces MUC5B overexpression. Interestingly, the overexpression is a result of two mechanisms. First, we show how the presence of the T allele disrupts a CpG methylation site, a natural mechanism to repress MUC5B expression, leading to increased expression. Secondly, when the T allele replaces the wild type allele, a novel C/EBPβ binding site is generated. Our results show a positive correlation between MUC5B expression and C/EBPβ LAP expression that is specific to the T allele carriers. Furthermore, we show how C/EBPβ directly binds to the rs35705950 position specifically in the presence of the minor T allele in our cell models. To demonstrate that C/EBPβ can acts as candidate transcription factor mediating this effect, we show the co-expression of MUC5B and C/EBPβ through lung epithelium in human samples, both normal lungs and in IPF. We also show a common signaling pathway where both are overexpressed through IL-13 stimulation.

Additionally, we show preliminary findings that the presence of the minor T allele induces phenotypic changes in our cell models. Traditional markers of ER stress were identified such as mitochondria proliferation, growth arrest, and the activation of several members of the ER stress-related apoptosis signaling pathway such as CHOP, BiM and Bcl-xL. Furthermore, the C/EBPβ transcription factor was identified as a potential target to study in fibrosis associated with IPF. Nevertheless, further studies are needed to decipher how C/EBPβ and MUC5B can result in fibrosis and the pathological consequences associated with IPF.

References

(2000). "American Thoracic Society. Idiopathic pulmonary fibrosis: diagnosis and treatment. International consensus statement. American Thoracic Society (ATS), and the European Respiratory Society (ERS)." Am J Respir Crit Care Med **161**(2 Pt 1): 646-664.

Adriaensen, D., I. Brouns, J. Van Genechten and J. P. Timmermans (2003). "Functional morphology of pulmonary neuroepithelial bodies: extremely complex airway receptors." Anat Rec A Discov Mol Cell Evol Biol **270**(1): 25-40.

Alberts, B. (2015). Molecular biology of the cell. New York, NY, Garland Science, Taylor and Francis Group.

American Thoracic, S. and S. European Respiratory (2002). "American Thoracic Society/European Respiratory Society International Multidisciplinary Consensus Classification of the Idiopathic Interstitial Pneumonias. This joint statement of the American Thoracic Society (ATS), and the European Respiratory Society (ERS) was adopted by the ATS board of directors, June 2001 and by the ERS Executive Committee, June 2001." Am J Respir Crit Care Med **165**(2): 277-304.

Arason, A. J., H. R. Jonsdottir, S. Halldorsson, B. E. Benediktsdottir, J. T. Bergthorsson, S. Ingthorsson, O. Baldursson, S. Sinha, T. Gudjonsson and M. K. Magnusson (2014). "deltaNp63 has a role in maintaining epithelial integrity in airway epithelium." PLoS One **9**(2): e88683.

Armanios, M. Y., J. J. Chen, J. D. Cogan, J. K. Alder, R. G. Ingersoll, C. Markin, W. E. Lawson, M. Xie, I. Vulto, J. A. Phillips, 3rd, P. M. Lansdorp, C. W. Greider and J. E. Loyd (2007). "Telomerase mutations in families with idiopathic pulmonary fibrosis." N Engl J Med **356**(13): 1317-1326.

Armstrong, D. A., L. N. Phelps and M. P. Vincenti (2009). "CCAAT enhancer binding protein-beta regulates matrix metalloproteinase-1 expression in interleukin-1beta-stimulated A549 lung carcinoma cells." Mol Cancer Res **7**(9): 1517-1524.

Atherton, H. C., G. Jones and H. Danahay (2003). "IL-13-induced changes in the goblet cell density of human bronchial epithelial cell cultures: MAP kinase and phosphatidylinositol 3-kinase regulation." Am J Physiol Lung Cell Mol Physiol **285**(3): L730-739.

Barkauskas, C. E., M. J. Cronce, C. R. Rackley, E. J. Bowie, D. R. Keene, B. R. Stripp, S. H. Randell, P. W. Noble and B. L. Hogan (2013).

"Type 2 alveolar cells are stem cells in adult lung." J Clin Invest **123**(7): 3025-3036.

Baumgartner, K. B., J. M. Samet, D. B. Coultas, C. A. Stidley, W. C. Hunt, T. V. Colby and J. A. Waldron (2000). "Occupational and environmental risk factors for idiopathic pulmonary fibrosis: a multicenter case-control study. Collaborating Centers." Am J Epidemiol **152**(4): 307-315.

Bernstein, B. E., A. Meissner and E. S. Lander (2007). "The mammalian epigenome." Cell **128**(4): 669-681.

Boers, J. E., A. W. Ambergen and F. B. Thunnissen (1998). "Number and proliferation of basal and parabasal cells in normal human airway epithelium." Am J Respir Crit Care Med **157**(6 Pt 1): 2000-2006.

Bond, G. L., W. Hu, E. E. Bond, H. Robins, S. G. Lutzker, N. C. Arva, J. Bargonetti, F. Bartel, H. Taubert, P. Wuerl, K. Onel, L. Yip, S. J. Hwang, L. C. Strong, G. Lozano and A. J. Levine (2004). "A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans." Cell **119**(5): 591-602.

Bond, G. L. and A. J. Levine (2007). "A single nucleotide polymorphism in the p53 pathway interacts with gender, environmental stresses and tumor genetics to influence cancer in humans." Oncogene **26**(9): 1317-1323.

Bonella, F., S. Stowasser and L. Wollin (2015). "Idiopathic pulmonary fibrosis: current treatment options and critical appraisal of nintedanib." Drug Des Devel Ther **9**: 6407-6419.

Borie, R., B. Crestani, P. Dieude, H. Nunes, Y. Allanore, C. Kannengiesser, P. Airo, M. Matucci-Cerinic, B. Wallaert, D. Israel-Biet, J. Cadranel, V. Cottin, S. Gazal, A. L. Peljto, J. Varga, D. A. Schwartz, D. Valeyre and B. Grandchamp (2013). "The MUC5B variant is associated with idiopathic pulmonary fibrosis but not with systemic sclerosis interstitial lung disease in the European Caucasian population." PLoS One **8**(8): e70621.

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch and R. A. Young (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." Cell **122**(6): 947-956.

Bradley, B., H. M. Branley, J. J. Egan, M. S. Greaves, D. M. Hansell, N. K. Harrison, N. Hirani, R. Hubbard, F. Lake, A. B. Millar, W. A. Wallace, A. U. Wells, M. K. Whyte, M. L. Wilsher, B. T. S. S. o. C. C. British Thoracic Society Interstitial Lung Disease Guideline Group, A. Thoracic Society of, S. New Zealand Thoracic and S. Irish Thoracic

(2008). "Interstitial lung disease guideline: the British Thoracic Society in collaboration with the Thoracic Society of Australia and New Zealand and the Irish Thoracic Society." Thorax **63 Suppl 5**: v1-58.

Brody, A. R. (1993). "Asbestos-induced lung disease." Environ Health Perspect **100**: 21-30.

Burger-van Paassen, N., M. van der Sluis, J. Bouma, A. M. Korteland-van Male, P. Lu, I. Van Seuningen, G. Boehm, J. B. van Goudoever and I. B. Renes (2011). "Colitis development during the suckling-weaning transition in mucin Muc2-deficient mice." Am J Physiol Gastrointest Liver Physiol **301**(4): G667-678.

Cassel, T. N. and M. Nord (2003). "C/EBP transcription factors in the lung epithelium." Am J Physiol Lung Cell Mol Physiol **285**(4): L773-781.

Centrella, M., S. Christakos and T. L. McCarthy (2004). "Skeletal hormones and the C/EBP and Runx transcription factors: interactions that integrate and redefine gene expression." Gene **342**(1): 13-24.

Chen, C. C., S. Xiao, D. Xie, X. Cao, C. X. Song, T. Wang, C. He and S. Zhong (2013). "Understanding variation in transcription factor binding by modeling transcription factor genome-epigenome interactions." PLoS Comput Biol **9**(12): e1003367.

Chinery, R., J. A. Brockman, D. T. Dransfield and R. J. Coffey (1997). "Antioxidant-induced nuclear translocation of CCAAT/enhancer-binding protein beta. A critical role for protein kinase A-mediated phosphorylation of Ser299." J Biol Chem **272**(48): 30356-30361.

Cogan, J. D., J. A. Kropski, M. Zhao, D. B. Mitchell, L. Rives, C. Markin, E. T. Garnett, K. H. Montgomery, W. R. Mason, D. F. McKean, *et al.* (2015). "Rare variants in RTEL1 are associated with familial interstitial pneumonia." Am J Respir Crit Care Med **191**(6): 646-655.

Collard, H. R., K. J. Anstrom, M. I. Schwarz and D. A. Zisman (2007). "Sildenafil improves walk distance in idiopathic pulmonary fibrosis." Chest **131**(3): 897-899.

Craig, J. C., M. A. Schumacher, S. E. Mansoor, D. L. Farrens, R. G. Brennan and R. H. Goodman (2001). "Consensus and variant cAMP-regulated enhancers have distinct CREB-binding properties." J Biol Chem **276**(15): 11719-11728.

Darlington, G. J., S. E. Ross and O. A. MacDougald (1998). "The role of C/EBP genes in adipocyte differentiation." J Biol Chem **273**(46): 30057-30060.

Desai, T. J., D. G. Brownfield and M. A. Krasnow (2014). "Alveolar progenitor and stem cells in lung development, renewal and cancer." Nature **507**(7491): 190-194.

Dirksen, E. R. (1991). "Centriole and basal body formation during ciliogenesis revisited." Biol Cell **72**(1-2): 31-38.

Erlich, H., A. M. Valdes, J. Noble, J. A. Carlson, M. Varney, P. Concannon, J. C. Mychaleckyj, J. A. Todd, P. Bonella, A. L. Fear, E. Lavant, A. Louey, P. Moonsamy and C. Type 1 Diabetes Genetics (2008). "HLA DR-DQ haplotypes and genotypes and type 1 diabetes risk: analysis of the type 1 diabetes genetics consortium families." Diabetes **57**(4): 1084-1092.

Erwteman, T. M., M. C. Braat and W. G. van Aken (1977). "Interstitial pulmonary fibrosis: a new side effect of practolol." Br Med J **2**(6082): 297-298.

Evans, C. M., T. E. Fingerlin, M. I. Schwarz, D. Lynch, J. Kurche, L. Warg, I. V. Yang and D. A. Schwartz (2016). "Idiopathic Pulmonary Fibrosis: A Genetic Disease That Involves Mucociliary Dysfunction of the Peripheral Airways." Physiol Rev **96**(4): 1567-1591.

Fahy, J. V. and B. F. Dickey (2010). "Airway mucus function and dysfunction." N Engl J Med **363**(23): 2233-2247.

Fan, S. and X. Zhang (2009). "CpG island methylation pattern in different human tissues and its correlation with gene expression." Biochem Biophys Res Commun **383**(4): 421-425.

Fingerlin, T. E., E. Murphy, W. Zhang, A. L. Peljto, K. K. Brown, M. P. Steele, J. E. Loyd, G. P. Cosgrove, D. Lynch, S. Groshong, *et al.* (2013). "Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis." Nat Genet **45**(6): 613-620.

Friedman, A. D. (2007). "Transcriptional control of granulocyte and monocyte development." Oncogene **26**(47): 6816-6828.

Gaudette, B. T., N. N. Iwakoshi and L. H. Boise (2014). "Bcl-xL protein protects from C/EBP homologous protein (CHOP)-dependent apoptosis during plasma cell differentiation." J Biol Chem **289**(34): 23629-23640.

Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean and G. R. Abecasis (2015). "A global reference for human genetic variation." Nature **526**(7571): 68-74.

Gombart, A. F., J. Grewal and H. P. Koeffler (2007). "ATF4 differentially regulates transcriptional activation of myeloid-specific genes by C/EBPepsilon and C/EBPalpha." J Leukoc Biol **81**(6): 1535-1547.

Grisanzio, C. and M. L. Freedman (2010). "Chromosome 8q24-Associated Cancers and MYC." Genes Cancer **1**(6): 555-559.

Halldorsson, S., V. Asgrimsson, I. Axelsson, G. H. Gudmundsson, M. Steinarsdottir, O. Baldursson and T. Gudjonsson (2007). "Differentiation potential of a basal epithelial cell line established from human bronchial explant." In Vitro Cell Dev Biol Anim **43**(8-9): 283-289.

Harding, H. P., Y. Zhang and D. Ron (1999). "Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase." Nature **397**(6716): 271-274.

Hattori, T., N. Ohoka, H. Hayashi and K. Onozaki (2003). "C/EBP homologous protein (CHOP) up-regulates IL-6 transcription by trapping negative regulating NF-IL6 isoform." FEBS Lett **541**(1-3): 33-39.

Helling, B. A., A. N. Gerber, V. Kadiyala, S. K. Sasse, B. S. Pedersen, L. Sparks, Y. Nakano, T. Okamoto, C. M. Evans, I. V. Yang and D. A. Schwartz (2017). "Regulation of MUC5B Expression in Idiopathic Pulmonary Fibrosis." Am J Respir Cell Mol Biol **57**(1): 91-99.

Hewson, C. A., M. R. Edbrooke and S. L. Johnston (2004). "PMA induces the MUC5AC respiratory mucin in human bronchial epithelial cells, via PKC, EGF/TGF-alpha, Ras/Raf, MEK, ERK and Sp1-dependent mechanisms." J Mol Biol **344**(3): 683-695.

Ho, Y. Y., D. Lagares, A. M. Tager and M. Kapoor (2014). "Fibrosis--a lethal component of systemic sclerosis." Nat Rev Rheumatol **10**(7): 390-402.

Hollenbeck, J. J. and M. G. Oakley (2000). "GCN4 binds with high affinity to DNA sequences containing a single consensus half-site." Biochemistry **39**(21): 6380-6389.

Horimasu, Y., S. Ohshimo, F. Bonella, S. Tanaka, N. Ishikawa, N. Hattori, N. Kohno, J. Guzman and U. Costabel (2015). "MUC5B promoter polymorphism in Japanese patients with idiopathic pulmonary fibrosis." Respirology **20**(3): 439-444.

Horn, S., A. Figl, P. S. Rachakonda, C. Fischer, A. Sucker, A. Gast, S. Kadel, I. Moll, E. Nagore, K. Hemminki, D. Schadendorf and R. Kumar (2013). "TERT promoter mutations in familial and sporadic melanoma." Science **339**(6122): 959-961.

Hubbard, R., M. Cooper, M. Antoniak, A. Venn, S. Khan, I. Johnston, S. Lewis and J. Britton (2000). "Risk of cryptogenic fibrosing alveolitis in metal workers." Lancet **355**(9202): 466-467.

Hubbard, R., S. Lewis, K. Richards, I. Johnston and J. Britton (1996). "Occupational exposure to metal or wood dust and aetiology of cryptogenic fibrosing alveolitis." Lancet **347**(8997): 284-289.

Hubbard, R., A. Venn, C. Smith, M. Cooper, I. Johnston and J. Britton (1998). "Exposure to commonly prescribed drugs and the etiology of

cryptogenic fibrosing alveolitis: a case-control study." Am J Respir Crit Care Med **157**(3 Pt 1): 743-747.

Hunninghake, G. M., H. Hatabu, Y. Okajima, W. Gao, J. Dupuis, J. C. Latourelle, M. Nishino, T. Araki, O. E. Zazueta, S. Kurugol, *et al.* (2013). "MUC5B promoter polymorphism and interstitial lung abnormalities." N Engl J Med **368**(23): 2192-2200.

Hutchinson, J., A. Fogarty, R. Hubbard and T. McKeever (2015). "Global incidence and mortality of idiopathic pulmonary fibrosis: a systematic review." Eur Respir J **46**(3): 795-806.

Hutchinson, J. P., T. M. McKeever, A. W. Fogarty, V. Navaratnam and R. B. Hubbard (2014). "Increasing global mortality from idiopathic pulmonary fibrosis in the twenty-first century." Ann Am Thorac Soc **11**(8): 1176-1185.

Idiopathic Pulmonary Fibrosis Clinical Research, N., D. A. Zisman, M. Schwarz, K. J. Anstrom, H. R. Collard, K. R. Flaherty and G. W. Hunninghake (2010). "A controlled trial of sildenafil in advanced idiopathic pulmonary fibrosis." N Engl J Med **363**(7): 620-628.

Isler, J. A., A. H. Skalet and J. C. Alwine (2005). "Human cytomegalovirus infection activates and regulates the unfolded protein response." J Virol **79**(11): 6890-6899.

Iwai, K., T. Mori, N. Yamada, M. Yamaguchi and Y. Hosoda (1994). "Idiopathic pulmonary fibrosis. Epidemiologic approaches to occupational exposure." Am J Respir Crit Care Med **150**(3): 670-675.

Jonsdottir, H. R., A. J. Arason, R. Palsson, S. R. Franzdottir, T. Gudbjartsson, H. J. Isaksson, G. Gudmundsson, T. Gudjonsson and M. K. Magnusson (2015). "Basal cells of the human airways acquire mesenchymal traits in idiopathic pulmonary fibrosis and in culture." Lab Invest **95**(12): 1418-1428.

Jung, S. H., W. C. Chung, K. M. Lee, C. N. Paik, J. H. Jung, M. K. Lee, Y. K. Lee and I. S. Chung (2010). "Risk factors in malignant transformation of gastric epithelial neoplasia categorized by the revised Vienna classification: endoscopic, pathological, and immunophenotypic features." Gastric Cancer **13**(2): 123-130.

Kaur, A., S. K. Mathai and D. A. Schwartz (2017). "Genetics in Idiopathic Pulmonary Fibrosis Pathogenesis, Prognosis, and Treatment." Front Med (Lausanne) **4**: 154.

Khurana, E., Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin and M. Gerstein (2016). "Role of non-coding sequence variants in cancer." Nat Rev Genet **17**(2): 93-108.

Kim, C. F., E. L. Jackson, A. E. Woolfenden, S. Lawrence, I. Babar, S. Vogel, D. Crowley, R. T. Bronson and T. Jacks (2005). "Identification

of bronchioalveolar stem cells in normal lung and lung cancer." Cell **121**(6): 823-835.

Kinoshita, S., S. Akira and T. Kishimoto (1992). "A member of the C/EBP family, NF-IL6 beta, forms a heterodimer and transcriptionally synergizes with NF-IL6." Proc Natl Acad Sci U S A **89**(4): 1473-1476.

Kirkham, S., J. K. Sheehan, D. Knight, P. S. Richardson and D. J. Thornton (2002). "Heterogeneity of airways mucus: variations in the amounts and glycoforms of the major oligomeric mucins MUC5AC and MUC5B." Biochem J **361**(Pt 3): 537-546.

Knight, D. A. and S. T. Holgate (2003). "The airway epithelium: structural and functional properties in health and disease." Respirology **8**(4): 432-446.

Knipe, R. S., A. M. Tager and J. K. Liao (2015). "The Rho kinases: critical mediators of multiple profibrotic processes and rational targets for new therapies for pulmonary fibrosis." Pharmacol Rev **67**(1): 103-117.

Kohler, J. J., S. J. Metallo, T. L. Schneider and A. Schepartz (1999). "DNA specificity enhanced by sequential binding of protein monomers." Proc Natl Acad Sci U S A **96**(21): 11735-11739.

Kono, Y., T. Nishiuma, T. Okada, K. Kobayashi, Y. Funada, Y. Kotani, S. Jahangeer, S. Nakamura and Y. Nishimura (2010). "Sphingosine kinase 1 regulates mucin production via ERK phosphorylation." Pulm Pharmacol Ther **23**(1): 36-42.

Lawson, W. E., P. F. Crossno, V. V. Polosukhin, J. Roldan, D. S. Cheng, K. B. Lane, T. R. Blackwell, C. Xu, C. Markin, L. B. Ware, G. G. Miller, J. E. Loyd and T. S. Blackwell (2008). "Endoplasmic reticulum stress in alveolar epithelial cells is prominent in IPF: association with altered surfactant protein processing and herpesvirus infection." Am J Physiol Lung Cell Mol Physiol **294**(6): L1119-1126.

Li, M. J., B. Yan, P. C. Sham and J. Wang (2015). "Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression." Brief Bioinform **16**(3): 393-412.

Li, Y., Y. Guo, J. Tang, J. Jiang and Z. Chen (2014). "New insights into the roles of CHOP-induced apoptosis in ER stress." Acta Biochim Biophys Sin (Shanghai) **46**(8): 629-640.

Listman, J. A., N. Wara-aswapati, J. E. Race, L. W. Blystone, N. Walker-Kopp, Z. Yang and P. E. Auron (2005). "Conserved ETS domain arginines mediate DNA binding, nuclear localization, and a novel mode of bZIP interaction." J Biol Chem **280**(50): 41421-41428.

Lumsden, A. B., A. McLean and D. Lamb (1984). "Goblet and Clara cells of human distal airways: evidence for smoking induced changes in their numbers." Thorax **39**(11): 844-849.

Lynch, T. J. and J. F. Engelhardt (2014). "Progenitor cells in proximal airway epithelial development and regeneration." J Cell Biochem **115**(10): 1637-1645.

Margaritopoulos, G. A., E. Vasarmidi and K. M. Antoniou (2016). "Pirfenidone in the treatment of idiopathic pulmonary fibrosis: an evidence-based review of its place in therapy." Core Evid **11**: 11-22.

Mason, D. P., M. E. Brizzio, J. M. Alster, A. M. McNeill, S. C. Murthy, M. M. Budev, A. C. Mehta, O. A. Minai, G. B. Pettersson and E. H. Blackstone (2007). "Lung transplantation for idiopathic pulmonary fibrosis." Ann Thorac Surg **84**(4): 1121-1128.

Mathai, S. K., C. A. Newton, D. A. Schwartz and C. K. Garcia (2016). "Pulmonary fibrosis in the era of stratified medicine." Thorax **71**(12): 1154-1160.

Mathelier, A., W. Shi and W. W. Wasserman (2015). "Identification of altered cis-regulatory elements in human disease." Trends Genet **31**(2): 67-76.

Mathieson, J. R., J. R. Mayo, C. A. Staples and N. L. Muller (1989). "Chronic diffuse infiltrative lung disease: comparison of diagnostic accuracy of CT and chest radiography." Radiology **171**(1): 111-116.

Mayo, J. R. (2009). "CT evaluation of diffuse infiltrative lung disease: dose considerations and optimal technique." J Thorac Imaging **24**(4): 252-259.

Mermigkis, C., J. Chapman, J. Golish, D. Mermigkis, K. Budur, A. Kopanakis, V. Polychronopoulos, R. Burgess and N. Foldvary-Schaefer (2007). "Sleep-related breathing disorders in patients with idiopathic pulmonary fibrosis." Lung **185**(3): 173-178.

Metallo, S. J. and A. Schepartz (1997). "Certain bZIP peptides bind DNA sequentially as monomers and dimerize on the DNA." Nat Struct Biol **4**(2): 115-117.

Molyneaux, P. L., M. J. Cox, S. A. Willis-Owen, P. Mallia, K. E. Russell, A. M. Russell, E. Murphy, S. L. Johnston, D. A. Schwartz, A. U. Wells, W. O. Cookson, T. M. Maher and M. F. Moffatt (2014). "The role of bacteria in the pathogenesis and progression of idiopathic pulmonary fibrosis." Am J Respir Crit Care Med **190**(8): 906-913.

Mora, A. L., M. Bueno and M. Rojas (2017). "Mitochondria in the spotlight of aging and idiopathic pulmonary fibrosis." J Clin Invest **127**(2): 405-414.

Musk, A. W. and J. A. Pollard (1979). "Pindolol and pulmonary fibrosis." Br Med J **2**(6190): 581-582.

Navaratnam, V., K. M. Fleming, J. West, C. J. Smith, R. G. Jenkins, A. Fogarty and R. B. Hubbard (2011). "The rising incidence of idiopathic pulmonary fibrosis in the U.K." Thorax **66**(6): 462-467.

Neurohr, C., P. Huppmann, D. Thum, W. Leuschner, W. von Wulffen, T. Meis, H. Leuchte, R. Baumgartner, G. Zimmermann, R. Hatz, S. Czerner, L. Frey, P. Ueberfuhr, I. Bittmann, J. Behr and G. Munich Lung Transplant (2010). "Potential functional and survival benefit of double over single lung transplantation for selected patients with idiopathic pulmonary fibrosis." Transpl Int **23**(9): 887-896.

Newman, J. R. and A. E. Keating (2003). "Comprehensive identification of human bZIP interactions with coiled-coil arrays." Science **300**(5628): 2097-2101.

Noth, I., Y. Zhang, S. F. Ma, C. Flores, M. Barber, Y. Huang, S. M. Broderick, M. S. Wade, P. Hysi, J. Scuirba, *et al.* (2013). "Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study." Lancet Respir Med **1**(4): 309-317.

O'Connor, L., A. Strasser, L. A. O'Reilly, G. Hausmann, J. M. Adams, S. Cory and D. C. Huang (1998). "Bim: a novel member of the Bcl-2 family that promotes apoptosis." EMBO J **17**(2): 384-395.

Olson, A. L., J. J. Swigris, D. C. Lezotte, J. M. Norris, C. G. Wilson and K. K. Brown (2007). "Mortality from pulmonary fibrosis increased in the United States from 1992 to 2003." Am J Respir Crit Care Med **176**(3): 277-284.

Park, J. S., J. S. Yeom, J. H. Seo, J. Y. Lim, C. H. Park, H. O. Woo, H. S. Youn, J. S. Jun, J. H. Park, G. H. Ko, S. C. Baik, W. K. Lee, M. J. Cho and K. H. Rhee (2015). "Immunohistochemical Expressions of MUC2, MUC5AC, and MUC6 in Normal, Helicobacter pylori Infected and Metaplastic Gastric Mucosa of Children and Adolescents." Helicobacter **20**(4): 260-268.

Peljto, A. L., M. Selman, D. S. Kim, E. Murphy, L. Tucker, A. Pardo, J. S. Lee, W. Ji, M. I. Schwarz, I. V. Yang, D. A. Schwartz and T. E. Fingerlin (2015). "The MUC5B promoter polymorphism is associated with idiopathic pulmonary fibrosis in a Mexican cohort but is rare among Asian ancestries." Chest **147**(2): 460-464.

Peljto, A. L., M. P. Steele, T. E. Fingerlin, M. E. Hinchcliff, E. Murphy, S. Podlusky, M. Carns, M. Schwarz, J. Varga and D. A. Schwartz (2012). "The pulmonary fibrosis-associated MUC5B promoter polymorphism does not influence the development of interstitial pneumonia in systemic sclerosis." Chest **142**(6): 1584-1588.

Peljto, A. L., Y. Zhang, T. E. Fingerlin, S. F. Ma, J. G. Garcia, T. J. Richards, L. J. Silveira, K. O. Lindell, M. P. Steele, J. E. Loyd, *et al.* (2013). "Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis." JAMA **309**(21): 2232-2239.

Plantier, L., B. Crestani, S. E. Wert, M. Dehoux, B. Zweytick, A. Guenther and J. A. Whitsett (2011). "Ectopic respiratory epithelial cell differentiation in bronchiolised distal airspaces in idiopathic pulmonary fibrosis." Thorax **66**(8): 651-657.

Pless, O., E. Kowenz-Leutz, M. Knoblich, J. Lausen, M. Beyermann, M. J. Walsh and A. Leutz (2008). "G9a-mediated lysine methylation alters the function of CCAAT/enhancer-binding protein-beta." J Biol Chem **283**(39): 26357-26363.

Plevy, S. E., J. H. Gemberling, S. Hsu, A. J. Dorner and S. T. Smale (1997). "Multiple control elements mediate activation of the murine and human interleukin 12 p40 promoters: evidence of functional synergy between C/EBP and Rel proteins." Mol Cell Biol **17**(8): 4572-4588.

Podust, L. M., A. M. Krezel and Y. Kim (2001). "Crystal structure of the CCAAT box/enhancer-binding protein beta activating transcription factor-4 basic leucine zipper heterodimer in the absence of DNA." J Biol Chem **276**(1): 505-513.

Ponnusamy, M. P., P. Seshacharyulu, A. Vaz, P. Dey and S. K. Batra (2011). "MUC4 stabilizes HER2 expression and maintains the cancer stem cell population in ovarian cancer cells." J Ovarian Res **4**(1): 7.

Portal, C., V. Gouyer, M. Magnien, S. Plet, F. Gottrand and J. L. Desseyn (2017). "In vivo imaging of the Muc5b gel-forming mucin." Sci Rep **7**: 44591.

Raghu, G., H. R. Collard, J. J. Egan, F. J. Martinez, J. Behr, K. K. Brown, T. V. Colby, J. F. Cordier, K. R. Flaherty, J. A. Lasky, *et al.* (2011). "An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management." Am J Respir Crit Care Med **183**(6): 788-824.

Rawlins, E. L., T. Okubo, Y. Xue, D. M. Brass, R. L. Auten, H. Hasegawa, F. Wang and B. L. Hogan (2009). "The role of Scgb1a1+ Clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium." Cell Stem Cell **4**(6): 525-534.

Reis, C. A., L. David, P. Correa, F. Carneiro, C. de Bolos, E. Garcia, U. Mandel, H. Clausen and M. Sobrinho-Simoes (1999). "Intestinal metaplasia of human stomach displays distinct patterns of mucin (MUC1, MUC2, MUC5AC, and MUC6) expression." Cancer Res **59**(5): 1003-1007.

Rock, J. R., M. W. Onaitis, E. L. Rawlins, Y. Lu, C. P. Clark, Y. Xue, S. H. Randell and B. L. Hogan (2009). "Basal cells as stem cells of the mouse trachea and human airway epithelium." Proc Natl Acad Sci U S A **106**(31): 12771-12775.

Ron, D. and J. F. Habener (1992). "CHOP, a novel developmentally regulated nuclear protein that dimerizes with transcription factors C/EBP and LAP and functions as a dominant-negative inhibitor of gene transcription." Genes Dev **6**(3): 439-453.

Ross, M. H. and W. Pawlina (2016). Histology : a text and atlas : with correlated cell and molecular biology. Philadelphia, Wolters Kluwer Health.

Roy, M. G., A. Livraghi-Butrico, A. A. Fletcher, M. M. McElwee, S. E. Evans, R. M. Boerner, S. N. Alexander, L. K. Bellinghausen, A. S. Song, Y. M. Petrova, *et al.* (2014). "Muc5b is required for airway defence." Nature **505**(7483): 412-416.

Roy, S. K., J. Hu, Q. Meng, Y. Xia, P. S. Shapiro, S. P. Reddy, L. C. Platanias, D. J. Lindner, P. F. Johnson, C. Pritchard, G. Pages, J. Pouyssegur and D. V. Kalvakolanu (2002). "MEKK1 plays a critical role in activating the transcription factor C/EBP-beta-dependent gene expression in response to IFN-gamma." Proc Natl Acad Sci U S A **99**(12): 7945-7950.

Schroder, M. and R. J. Kaufman (2005). "The mammalian unfolded protein response." Annu Rev Biochem **74**: 739-789.

Seibold, M. A., R. W. Smith, C. Urbanek, S. D. Groshong, G. P. Cosgrove, K. K. Brown, M. I. Schwarz, D. A. Schwartz and S. D. Reynolds (2013). "The idiopathic pulmonary fibrosis honeycomb cyst contains a mucocilary pseudostratified epithelium." PLoS One **8**(3): e58658.

Seibold, M. A., A. L. Wise, M. C. Speer, M. P. Steele, K. K. Brown, J. E. Loyd, T. E. Fingerlin, W. Zhang, G. Gudmundsson, S. D. Groshong, *et al.* (2011). "A common MUC5B promoter polymorphism and pulmonary fibrosis." N Engl J Med **364**(16): 1503-1512.

Selman, M., T. E. King, A. Pardo, S. American Thoracic, S. European Respiratory and P. American College of Chest (2001). "Idiopathic pulmonary fibrosis: prevailing and evolving hypotheses about its pathogenesis and implications for therapy." Ann Intern Med **134**(2): 136-151.

Selman, M. and A. Pardo (2006). "Role of epithelial cells in idiopathic pulmonary fibrosis: from innocent targets to serial killers." Proc Am Thorac Soc **3**(4): 364-372.

Selman, M., A. Pardo and N. Kaminski (2008). "Idiopathic pulmonary fibrosis: aberrant recapitulation of developmental programs?" PLoS Med **5**(3): e62.

Smink, J. J. and A. Leutz (2010). "Rapamycin and the transcription factor C/EBPbeta as a switch in osteoclast differentiation: implications for lytic bone diseases." J Mol Med (Berl) **88**(3): 227-233.

Spira, A., J. Beane, V. Shah, G. Liu, F. Schembri, X. Yang, J. Palma and J. S. Brody (2004). "Effects of cigarette smoke on the human airway epithelial cell transcriptome." Proc Natl Acad Sci U S A **101**(27): 10143-10148.

Stewart, J. P., J. J. Egan, A. J. Ross, B. G. Kelly, S. S. Lok, P. S. Hasleton and A. A. Woodcock (1999). "The detection of Epstein-Barr virus DNA in lung tissue from patients with idiopathic pulmonary fibrosis." Am J Respir Crit Care Med **159**(4 Pt 1): 1336-1341.

Stock, C. J., H. Sato, C. Fonseca, W. A. Banya, P. L. Molyneaux, H. Adamali, A. M. Russell, C. P. Denton, D. J. Abraham, D. M. Hansell, A. G. Nicholson, T. M. Maher, A. U. Wells, G. E. Lindahl and E. A. Renzoni (2013). "Mucin 5B promoter polymorphism is associated with idiopathic pulmonary fibrosis but not with development of lung fibrosis in systemic sclerosis or sarcoidosis." Thorax **68**(5): 436-441.

Stone, K. C., R. R. Mercer, B. A. Freeman, L. Y. Chang and J. D. Crapo (1992). "Distribution of lung cell numbers and volumes between alveolar and nonalveolar tissue." Am Rev Respir Dis **146**(2): 454-456.

Stuart, B. D., J. Choi, S. Zaidi, C. Xing, B. Holohan, R. Chen, M. Choi, P. Dharwadkar, F. Torres, C. E. Girod, *et al.* (2015). "Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening." Nat Genet **47**(5): 512-517.

Sylvester, P. A., N. Myerscough, B. F. Warren, I. Carlstedt, A. P. Corfield, P. Durdey and M. G. Thomas (2001). "Differential expression of the chromosome 11 mucin genes in colorectal cancer." J Pathol **195**(3): 327-335.

Tamura, Y., M. Higashi, S. Kitamoto, S. Yokoyama, M. Osako, M. Horinouchi, T. Shimizu, M. Tabata, S. K. Batra, M. Goto and S. Yonezawa (2012). "MUC4 and MUC1 expression in adenocarcinoma of the stomach correlates with vessel invasion and lymph node metastasis: an immunohistochemical study of early gastric cancer." PLoS One **7**(11): e49251.

Tang, Y. W., J. E. Johnson, P. J. Browning, R. A. Cruz-Gervis, A. Davis, B. S. Graham, K. L. Brigham, J. A. Oates, Jr., J. E. Loyd and A. A. Stecenko (2003). "Herpesvirus DNA is consistently detected in lungs of patients with idiopathic pulmonary fibrosis." J Clin Microbiol **41**(6): 2633-2640.

Tashiro, E., N. Hironiwa, M. Kitagawa, Y. Futamura, S. Suzuki, M. Nishio and M. Imoto (2007). "Trierixin, a novel Inhibitor of ER stress-induced XBP1 activation from Streptomyces sp. 1. Taxonomy, fermentation, isolation and biological activities." J Antibiot (Tokyo) **60**(9): 547-553.

Thomas, A. Q., K. Lane, J. Phillips, 3rd, M. Prince, C. Markin, M. Speer, D. A. Schwartz, R. Gaddipati, A. Marney, J. Johnson, R. Roberts, J. Haines, M. Stahlman and J. E. Loyd (2002). "Heterozygosity for a surfactant protein C gene mutation associated with usual interstitial pneumonitis and cellular nonspecific interstitial pneumonitis in one kindred." Am J Respir Crit Care Med **165**(9): 1322-1328.

Thornton, D. J., K. Rousseau and M. A. McGuckin (2008). "Structure and function of the polymeric mucins in airways mucus." Annu Rev Physiol **70**: 459-486.

Toskala, E., S. M. Smiley-Jewell, V. J. Wong, D. King and C. G. Plopper (2005). "Temporal and spatial distribution of ciliogenesis in the tracheobronchial airways of mice." Am J Physiol Lung Cell Mol Physiol **289**(3): L454-459.

Travis, W. D., U. Costabel, D. M. Hansell, T. E. King, Jr., D. A. Lynch, A. G. Nicholson, C. J. Ryerson, J. H. Ryu, M. Selman, A. U. Wells, *et al.* (2013). "An official American Thoracic Society/European Respiratory Society statement: Update of the international multidisciplinary classification of the idiopathic interstitial pneumonias." Am J Respir Crit Care Med **188**(6): 733-748.

Tsakiri, K. D., J. T. Cronkhite, P. J. Kuan, C. Xing, G. Raghu, J. C. Weissler, R. L. Rosenblatt, J. W. Shay and C. K. Garcia (2007). "Adult-onset pulmonary fibrosis caused by mutations in telomerase." Proc Natl Acad Sci U S A **104**(18): 7552-7557.

Tsukada, J., K. Saito, W. R. Waterman, A. C. Webb and P. E. Auron (1994). "Transcription factors NF-IL6 and CREB recognize a common essential site in the human prointerleukin 1 beta gene." Mol Cell Biol **14**(11): 7285-7297.

Tsukada, J., Y. Yoshida, Y. Kominato and P. E. Auron (2011). "The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation." Cytokine **54**(1): 6-19.

van Moorsel, C. H., M. F. van Oosterhout, N. P. Barlo, P. A. de Jong, J. J. van der Vis, H. J. Ruven, H. W. van Es, J. M. van den Bosch and J. C. Grutters (2010). "Surfactant protein C mutations are the basis of a significant portion of adult familial pulmonary fibrosis in a dutch cohort." Am J Respir Crit Care Med **182**(11): 1419-1425.

Vanoni, S., Y. T. Tsai, A. Waddell, L. Waggoner, J. Klarquist, S. Divanovic, K. Hoebe, K. A. Steinbrecher and S. P. Hogan (2017). "Myeloid-derived NF-kappaB negative regulation of PU.1 and c/EBP-beta-driven pro-inflammatory cytokine production restrains LPS-induced shock." Innate Immun **23**(2): 175-187.

Vincent, A., M. Perrais, J. L. Desseyn, J. P. Aubert, P. Pigny and I. Van Seuningen (2007). "Epigenetic regulation (DNA methylation, histone modifications) of the 11p15 mucin genes (MUC2, MUC5AC, MUC5B, MUC6) in epithelial cancer cells." Oncogene **26**(45): 6566-6576.

Vinson, C. R., T. Hai and S. M. Boyd (1993). "Dimerization specificity of the leucine zipper-containing bZIP motif on DNA binding: prediction and rational design." Genes Dev **7**(6): 1047-1058.

Visel, A., E. M. Rubin and L. A. Pennacchio (2009). "Genomic views of distant-acting enhancers." Nature **461**(7261): 199-205.

Wakatsuki, K., Y. Yamada, M. Narikiyo, M. Ueno, T. Takayama, H. Tamaki, K. Miki, S. Matsumoto, K. Enomoto, T. Yokotani and Y. Nakajima (2008). "Clinicopathological and prognostic significance of mucin phenotype in gastric cancer." J Surg Oncol **98**(2): 124-129.

Walsh, M. D., M. Clendenning, E. Williamson, S. A. Pearson, R. J. Walters, B. Nagler, D. Packenas, A. K. Win, J. L. Hopper, M. A. Jenkins, A. M. Haydon, C. Rosty, D. R. English, G. G. Giles, M. A. McGuckin, J. P. Young and D. D. Buchanan (2013). "Expression of MUC2, MUC5AC, MUC5B, and MUC6 mucins in colorectal cancers and their association with the CpG island methylator phenotype." Mod Pathol **26**(12): 1642-1656.

Walters, M. S., K. Gomi, B. Ashbridge, M. A. Moore, V. Arbelaez, J. Heldrich, B. S. Ding, S. Rafii, M. R. Staudt and R. G. Crystal (2013). "Generation of a human airway epithelium derived basal cell line with multipotent differentiation capacity." Respir Res **14**: 135.

Wang, C., Y. Zhuang, W. Guo, L. Cao, H. Zhang, L. Xu, Y. Fan, D. Zhang and Y. Wang (2014). "Mucin 5B promoter polymorphism is associated with susceptibility to interstitial lung diseases in Chinese males." PLoS One **9**(8): e104919.

Wang, M. and R. J. Kaufman (2016). "Protein misfolding in the endoplasmic reticulum as a conduit to human disease." Nature **529**(7586): 326-335.

Wang, Y., P. J. Kuan, C. Xing, J. T. Cronkhite, F. Torres, R. L. Rosenblatt, J. M. DiMaio, L. N. Kinch, N. V. Grishin and C. K. Garcia (2009). "Genetic defects in surfactant protein A2 are associated with pulmonary fibrosis and lung cancer." Am J Hum Genet **84**(1): 52-59.

Wei, R., C. Li, M. Zhang, Y. L. Jones-Hall, J. L. Myers, I. Noth and W. Liu (2014). "Association between MUC5B and TERT polymorphisms and different interstitial lung disease phenotypes." Transl Res **163**(5): 494-502.

Willis, B. C., J. M. Liebler, K. Luby-Phelps, A. G. Nicholson, E. D. Crandall, R. M. du Bois and Z. Borok (2005). "Induction of epithelial-mesenchymal transition in alveolar epithelial cells by transforming growth factor-beta1: potential role in idiopathic pulmonary fibrosis." Am J Pathol **166**(5): 1321-1332.

Wilson, M. S. and T. A. Wynn (2009). "Pulmonary fibrosis: pathogenesis, etiology and regulation." Mucosal Immunol **2**(2): 103-121.

Wolff, R. K. (1986). "Effects of airborne pollutants on mucociliary clearance." Environ Health Perspect **66**: 223-237.

Wolters, P. J., H. R. Collard and K. D. Jones (2014). "Pathogenesis of idiopathic pulmonary fibrosis." Annu Rev Pathol **9**: 157-179.

Wu, H., M. Zhao, A. Yoshimura, C. Chang and Q. Lu (2016). "Critical Link Between Epigenetics and Transcription Factors in the Induction of Autoimmunity: a Comprehensive Review." Clin Rev Allergy Immunol **50**(3): 333-344.

Wynn, T. A. and T. R. Ramalingam (2012). "Mechanisms of fibrosis: therapeutic translation for fibrotic disease." Nat Med **18**(7): 1028-1040.

Xiao, W., D. R. Hodge, L. Wang, X. Yang, X. Zhang and W. L. Farrar (2004). "Co-operative functions between nuclear factors NFkappaB and CCAT/enhancer-binding protein-beta (C/EBP-beta) regulate the IL-6 promoter in autocrine human prostate cancer cells." Prostate **61**(4): 354-370.

Yeh, W. C., Z. Cao, M. Classon and S. L. McKnight (1995). "Cascade regulation of terminal adipocyte differentiation by three members of the C/EBP family of leucine zipper proteins." Genes Dev **9**(2): 168-181.

Young, H. W., O. W. Williams, D. Chandra, L. K. Bellinghausen, G. Perez, A. Suarez, M. J. Tuvim, M. G. Roy, S. N. Alexander, S. J. Moghaddam, R. Adachi, M. R. Blackburn, B. F. Dickey and C. M. Evans (2007). "Central role of Muc5ac expression in mucous metaplasia and its regulation by conserved 5' elements." Am J Respir Cell Mol Biol **37**(3): 273-290.

Zhang, L., Y. Wang, N. S. Pandupuspitasari, G. Wu, X. Xiang, Q. Gong, W. Xiong, C. Y. Wang, P. Yang and B. Ren (2017). "Endoplasmic reticulum stress, a new wrestler, in the pathogenesis of idiopathic pulmonary fibrosis." Am J Transl Res **9**(2): 722-735.

Zhang, Y., I. Noth, J. G. Garcia and N. Kaminski (2011). "A variant in the promoter of MUC5B and idiopathic pulmonary fibrosis." N Engl J Med **364**(16): 1576-1577.

Zhu, Y., M. A. Saunders, H. Yeh, W. G. Deng and K. K. Wu (2002). "Dynamic regulation of cyclooxygenase-2 promoter activity by isoforms of CCAAT/enhancer-binding proteins." J Biol Chem **277**(9): 6923-6928.

Original publications

Paper I

A MUC5B polymorphism associated with Idiopathic Pulmonary Fibrosis mediates overexpression through decreased CpG methylation and C/EBPβ recruitment driving transcriptional activation

Amaranta U. Armesto[1], Erna Magnusdottir[2,3], Ari J. Arason[1], Olafur A. Stefansson[7], Gunnar Gudmundsson[4,5], Thorarinn Gudjonsson[1,6], Magnus K. Magnusson[4,6*]

[1] Stem Cell Research Unit, Biomedical Centre, University of Iceland and Departments of [2]Anatomy, [3]Biomedical Science and [4]Pharmacology and Toxicology, Faculty of Medicine, University of Iceland, and [5]Department of Pulmonary Medicine, [6]Department of Laboratory Hematology, Landspitalinn University Hospital, Reykjavik, Iceland, [7]deCODE Genetics/AMGEN, Inc., Reykjavik, Iceland.

***** corresponding author: Magnus K. Magnusson magnuskm@hi.is (MM)

Short title: CEBPβ and CpG methylation regulate MUC5B in idiopathic pulmonary fibrosis

Abstract

Idiopathic pulmonary fibrosis is a progressive and fatal lung disease of unknown aetiology. The strongest genetic risk factor associated with IPF development is a *MUC5B* promoter polymorphism (*rs35705950*). However, the mechanism underlying its effects remains unknown. In this study we have focused on the molecular consequences of the polymorphism on the regulation of MUC5B expression. We have identified a combined mechanism involving both methylation and direct transcriptional regulation mediated by the polymorphic variant on MUC5B overexpression. Our results demonstrate that the minor allele (T) associated with *rs35705950* disturbs a DNA methylation site, directly increasing MUC5B expression. Furthermore, this same variant also creates a novel binding site for the transcription factor C/EBPβ leading to transcriptional activation of *MUC5B*. Our findings provide a novel insight into the regulatory effects of the IPF risk allele, rs35705950 and identifies C/EBPβ as an important regulatory factor in the development of IPF.

Introduction

Idiopathic pulmonary fibrosis (IPF) is an irreversible interstitial lung disease characterized by a progressive scarring of lung parenchyma often leading to a fatally declining lung function. IPF incidence has been estimated to be around 75/100,000, affecting 5 million people world-wide [1-3]. Furthermore, IPF is believed to be an under diagnosed disease [4-6].

Due to the complexity and progressive nature of the disease, available treatments are limited and have only a modest impact on IPF progression. Until recently, only lung transplant has been proved to increase survival [1]. In recent years, improvement in development of new therapies has met some success, as two novel drugs have been introduced, pirfenidone and nintedanib. Both drugs are believed to target profibrotic signalling pathways in IPF. Pirfenidone inhibits TGF-β1 production possibly by inhibiting the upregulation of HSP47 and Col1 RNA in fibroblasts [7]. Nintedanib is a potent small-molecule receptor tyrosine kinase inhibitor targeting platelet-derived growth factor receptors *(PDGF-R)*, fibroblast growth factor receptors *(FGFR)* and vascular endothelial growth factor *(VEGF)* -family tyrosine kinase receptors [8].

Both genetic and environmental factors are believed to contribute to the onset and progression of the disease. Several environmental factors have been associated with IPF, including exposure to metal and wood dust [9-12], viruses [13-15], drugs [16-18], and cigarette smoke [19-

[21]. Recent findings point to genetic factors as major triggers of IPF. Rare mutations found in 6 genes (*TERT* [22,23], *TERC* [22,23], *RTEL1* [24,25], *PARN* [25], *STPC* [26,27] and *SFTPA2* [28]), with associated variants in 11 different loci [29] indicate that the telomerase pathway and surfactant protein genes play a key role in IPF pathogenesis. However, these mutations only explain a small proportion of IPF cases.

Two large genome-wide association studies (GWAS) have been conducted on pulmonary fibrosis (familial and sporadic) [30,31]. Both studies showed the most important genetic risk to be conferred by a common G-to-T risk variant in the upstream region of *MUC5B* (*rs35705950*). The frequency of the risk allele (T) is ~35% among European ancestry cases, compared with ~9% of European ancestry controls [32,33]. These studies furthermore identified other common variants, including three polymorphisms in the *TOLLIP* gene and a desmoplakin (*DSP*) intron variant [31]. These studies concluded that the genome-wide variants account for 30-35% of the IPF risk, suggesting an important role for these common genetic variations in the disease aetiology [33].

Deciphering the molecular and cellular effects of non-coding SNP´s has proven to be notoriously difficult. With the thousands of common SNPs associated with many common diseases only a few have led to a clear molecular understanding [34-36]. There are many possible explanations for this. Most often there is a complex network of SNPs in varying degrees of linkage disequilibrium all associated with the phenotype or disease [36]. Resolving the true

informative SNP within this complex network of variants can be really challenging. Another explanation for the difficulty is connecting a non-coding variant to a target gene or regulatory region mediating the effects of the genetic association. Even though the closest gene is often a good candidate, the effects can be on genes further away. This can sometimes be resolved through expression-quantitative trait loci (eQTL), i.e. if the sequence variant influences the expression level of one or more genes.

The *MUC5B* promoter variant *rs35705950* is the strongest known risk factor (genetic and otherwise) for the development of IPF [32,33]. This variant is unusual as it is relatively simple, with no other variants in strong linkage disequilibrium and carries a strong eQTL association with MUC5B expression in lung. The odds ratio (OR) associated with carrying one allele of the variant is 4.5-6.6, while homozygosity leads to an OR of 9.6-20.2 [30,31,37-41]. Furthermore, carriers of the *MUC5B* variant are also at risk of getting subclinical interstitial lung disease based on screening with high-resolution CT [4]. Interestingly, recent studies have examined the relevance of *MUC5B* variation in other ethnic groups, showing that the frequency of *rs35705950* minor allele (A G-to-T SNP) is 11, 8 and 1% among European, South Asian and East Asian populations, respectively [38,42,43], and it is almost non-existent in Africans [44]. Despite the different prevalence, the evidence indicates that the risk of developing IPF associated with the variant is comparable to the risk observed in European population [4,45].

MUC5B (Mucin 5B) is a highly glycosylated protein, expressed throughout the upper and lower respiratory tract. In human upper airways MUC5B is predominantly expressed in nasal and oral gland secretions, while in tracheobronchial conducting airways it is expressed in surface epithelium and submucosal glands, where it is less predominant than the related MUC5AC in surface epithelia. In bronchioles MUC5B predominates, while MUC5AC is less expressed. Under disease conditions, the MUC5B expression pattern changes dramatically, it becomes overexpressed, showing significantly higher expression than MUC5AC [1,46,47]. The pattern suggests an important role for MUC5B as a host defence barrier. The G-to-T *rs35705950* variant is in a presumed regulatory region 3 kb upstream of the transcription start site in the *MUC5B* gene on chromosome 11p15. A DNase I hypersensitivity site overlaps the variant location suggesting an important regulatory role. Furthermore, the *MUC5B* variant lies within a cluster of CpG islands and CpG DNA methylation has been shown to affect *MUC5B* expression [48].

Even though the *MUC5B* variant has a strong and consistent genetic association with IPF and the variant seems to positively regulate MUC5B expression, the mechanism underlying its role in IPF is poorly understood. It has been hypothesized that, due to the large size of the MUC5B protein, its production may carry a significant metabolic stress, which can interfere with differentiation of airway stem cells [33]. It should also be emphasized that due to the increased

expression level there could also be secondary problems with post-translational modifications, such as glycosylation. In a mouse model of intestinal inflammation aberrant mucin assembly has been shown to cause ER-stress through activation of the unfolded protein response leading to inflammation, apoptosis, and wound repair [49]. ER-stress has been shown to be involved in IPF, specifically in cases caused by mutations in surfactant proteins, another protein that is highly glycosylated [50]. Other proposed mechanisms consider the possibility that IPF is a mucociliary disease caused by recurrent injury/inflammation/repair at the bronchoalveolar junction, as MUC5B overexpression might cause a reduction in mucociliary function and retention of particles leading to lung injury [33].

In this study, we aim to define the molecular mechanisms involved in mediating the effects of the *MUC5B* variant (T allele) on MUC5B expression. We use two immortalized human bronchial epithelial cell lines (BCi_NS1.1 and VA-10) and the well-established adenocarcinoma cell line A549 to corroborate the positive effects of the variant on MUC5B expression. With our models, we show that the T allele is associated with a twofold mechanism affecting MUC5B expression, leading to both the disruption of a DNA methylation site that directly increases MUC5B expression and the creation of a novel C/EBPβ transcription factor binding site that also positively regulates MUC5B expression.

Results

Rs35705950 risk allele for IPF is associated with higher expression of MUC5B.

The common polymorphism *rs35705950*, where a G is replaced by a T nucleotide, has been shown in two GWAS studies to be correlated with a predisposition to develop IPF [30,31]. The polymorphism is located in a mucin gene cluster, three kilobases upstream of the transcription start site of *MUC5B* on chromosome 11 (S1 Fig). The molecular mechanism explaining how this polymorphism affects the aetiology of IPF has not been elucidated, although both hetero- and homozygote carriers have been shown to express higher levels of MUC5B in the lung epithelium [32,51]. To study the direct effects of this polymorphism on *MUC5B* gene regulation, a 4.1 kb of the *MUC5B* cis-regulatory region was cloned into a luciferase reporter vector and the G of the wild type allele was replaced with the risk associated T allele using site directed mutagenesis (S3 Fig).

The resulting lentiviral vectors containing the luciferase gene under the control of 4.1kb *MUC5B* upstream region containing the promoter and either the wild type or the T variant (*rs35705950*), were used to stably transduce two human bronchial-derived basal epithelial cell lines, Bci_NS1.1 and VA10, and a human lung adenocarcinoma derived alveolar epithelial cell line A549. After stable selection the cell lines were cultured at confluency and luciferase reporter activity measured. As shown in Fig 1a the IPF risk

associated T allele consistently increased the luciferase reporter signal in all three cell lines, with the effect being most pronounced in Bci_NS1.1 cells.

To directly asses the effect of the T allele in the endogenous *rs35705950* site, we used CRISPR/Cas9 genome editing. The bronchial basal cell lines did not survive single cell cloning and thus we were only able to generate A549 edited cell lines. Five heterozygous A549^{CRISPR} clones (1-5) were generated, carrying the human [G/T] alleles (S4 Fig). The T allele had a very significant effect on the expression levels of both MUC5B mRNA (Fig 1b) and protein (Fig 1c), showing a direct effect of the T allele on MUC5B expression. The above data support the idea that the T allele confers an increased promoter/enhancer activity to the *MUC5B* upstream cis-regulatory sequence.

The rs35705950 T allele disrupts a repressive CpG DNA methylation site

It has been previously reported that the methylation of several CpG islands in the 4.1kb promoter region of *MUC5B* affects the expression of the MUC5B protein [48]. Interestingly, the presence of the T allele disrupts a CpG island that was previously characterized as methylated in the wild type. We hypothesized that the effects of the T allele could be mediated, at least in part, through epigenetic regulation. To analyse the effect of DNA methylation on MUC5B

expression we used an air-liquid interphase (ALI) culture of Bci_NS1.1 cells. Under these conditions the cells differentiate into both goblet cells producing mucins and ciliated cells [52]. We treated these cultures with 5'aza2'-deoxycytidine (5'AZA2'), a DNA methylation inhibitor. 5'AZA2' treatment resulted in increased expression of MUC5B and MUC5AC indicating that DNA methylation either directly or indirectly impacts MUC5B expression (Fig 2a). To directly address the potential effects of DNA-methylation at the T allele, we used M.SssI CpG methyltransferase to *in vitro* methylate the 4.1 kb MUC5B cis-regulatory region containing either the wild type or the T allele before ligating it into the luciferase reporter vector for transient transfection into our cell lines. After *in vitro* DNA methylation, the luciferase signal was stronger in all cell lines carrying the T allele compared to the WT allele (Fig 2b). To directly address the potential effects of DNA-methylation at the T allele, we compared the effects of the T-allele vs WT in a transient luciferase assay with or without direct *in vitro* methylation of the 4.1 kb MUC5B cis-regulatory prior to transient transfection. The cis regulatory region (with T-allele or WT) was restriction cut from the plasmid, and the fragments was *in vitro* methylated using M.SssI CpG methyltransferase prior to ligating back into a luciferase promoter plasmid and transiently transfected into the three cell lines. In Fig 2b without *in vitro* methylation no difference is seen in luciferase activity in VA10 or A549 cells between T-allele and WT while after in vitro methylation a significant increase is seen in the T-allele carrying luciferase plasmid compared to WT in all three cell

lines. This supports the idea that differential methylation of the T allele and wild type allele is at least partially involved in regulating MUC5B expression.

Targeted bisulfite sequencing of the region in our cell lines further revealed that methylation was present in all cell lines at the WT-genotype although the Bci_NS1.1 cells, is less methylated (21-40%), than in both VA10 and A549 cells (81-100%) under normal cell culture conditions (Fig 2d). All these cell lines are WT at the *rs35705950* SNP, as stated above. The CRISPR/Cas9 edited A549 cell lines harbouring the heterozygous [G/T] genotype showed less methylation (41-60%, clones 1-2, 21-40% clone 3) at the risk allele (Fig 2e) compared to $A549^{WT}$ (81-100%), confirming that the presence of the T allele reduces methylation in the polymorphic site. This suggests that direct DNA methylation might at least partially explain the increased MUC5B expression in individuals carrying IPF risk associated T allele.

C/EBPβ mediates MUC5B overexpression through rs35705950 T allele.

Even though our data support a role of differential DNA methylation to explain the effects of T allele on MUC5B expression, additional mechanisms might also be at play. To further study the potential effects of the T allele of *rs35705950* we used *Match,* a weight matrix-based program for predicting transcription factor binding

sites in DNA sequences using the DNA sequence flanking *rs35705950* polymorphism. The program uses a library of positional weight matrices from TRANSFAC® Public 6.0. Fig 3a shows the predicted transcription factor binding motif. Binding motifs for PAX2 and PAX4 overlap the polymorphism in the wild-type sequence while the T allele leads to the loss of the PAX4 motif. The T allele furthermore leads to a gain of a novel CCAAT/enhancer-binding proteins (C/EBPs) motif. PAX2 and PAX4 have very low expression levels in lung tissue (S2 Fig). In the C/EBP family, C/EBPβ has a documented role in inflammation [53,54] and C/EBPβ is highly expressed in lung tissue (S2 Fig), making it an ideal candidate to study further. As shown in Fig S2d the eQTL effects of the *rs35705950* polymorphism is primarily seen in the lung, suggesting lung specific effects. To further address whether the T-allele might lead to a novel C/EBPβ binding site we carried out a direct comparison of a predicted C/EBPβ binding using motifbreakR [55]. Fig S5A shows that there is a predicted consensus binding site for C/EBPβ on the negative strand in the minor allele. On the negative strand the T-allele is now A and the WT allele a C. The PWM (position weight matrix) score for the A-allele is 89% (P = 0.0041) compared to a PWM score of 71% (P = 0.103) for the C-allele (the major allele), novel binding of C/EBPβ to the negative strand of the IPF-associated *rs35705950* polymorphism.

C/EBPβ has three isoforms generated by alternative splicing (S1 Fig). Two of these isoforms (LAP and LAP*) mediate transcriptional

activation through the transactivation domain, while the third isoform (LIP) has an inhibitory role due to the lack of the transactivation domain [56].

Immunofluorescence staining on four IPF lung samples show that C/EBPβ and MUC5B are co-expressed in airway epithelial cells (Fig 3b, S3b Fig). Furthermore, in healthy lung samples C/EBPβ is highly expressed in alveolar cells resembling alveolar macrophages and they also express high levels of MUC5B (S3e Fig).

In addition to the aforementioned co-expression, C/EBPβ might act through a common signalling pathway potentially driving MUC5B overexpression through the T allele. IL-13 has been previously used to induce goblet cell hyperplasia in asthma models and has also been used to induce MUC5AC expression in cell culture models [52]. We cultured both VA10 and Bci_NS1.1 cell lines, under ALI culture conditions with or without IL-13. In both cell lines, IL-13 induced C/EBPβ as well as MUC5B expression at both the protein (Fig 3c) and mRNA level (Fig 3d), suggesting a co-regulatory pathway.

To analyse specifically the differential effects of C/EBPβ on *MUC5B* cis-regulatory domain on the wild type and T alleles, we overexpressed the C/EBPβ isoforms in the BCi_NS1.1 (Fig 4a), A549 (Fig 4b) and VA10 (S5a Fig) cell lines harbouring the stably integrated luciferase reporter associated with the 4.1kb *MUC5B* promoter region. The C/EBPβ LAP* isoform increased *MUC5B* promoter regulated luciferase activity 2.5-fold when the T allele was present,

compared to the wild type in both BC1_NS1.1 and A549. This relationship is dose-dependent, as shown in S5b Fig. Overexpression of the C/EBPβ LIP isoform reduced *MUC5B* reporter activity when the T allele was present in both cell lines (Fig 4 a, b). When both isoforms were co-expressed in equimolar concentrations an increased luciferase signal was seen in the BCi_NS1.1 cell line.

To further corroborate the role of C/EBPβ in regulating MUC5B expression through the risk allele, we knocked down C/EBPβ expression using siRNAs targeting either the LAP/LAP* or all isoforms. Stable cells lines carrying the luciferase reporter vector and $A549^{CRISPR}$ clones were used to analyse the effect of C/EBPβ knock-down on MUC5B expression by RT-qPCR (Fig 4c, S5c Fig) and luciferase reporter assay (Fig 4d, S5d Fig). The data show that siRNAs targeting C/EBPβ restores MUC5B expression in T allele genotype to a similar level as wild type.

To verify that the *rs35705950* polymorphism effects are through direct binding of C/EBPβ to the T allele, three separate chromatin immunoprecipitation experiments were performed on $A549^{WT}$ cells and $A549^{CRISPR}$ clones. As shown in a representative ChIP experiment in Fig 4e, there is a marked increase in C/EBPβ binding seen in the MUC5B cis-regulatory domain in CRISPR-Cas edited $A549^{CRISPR}$ [G/T] cells compared to $A549^{WT}$ (4.76-fold enrichment for G/T cells compared to 0.61-fold enrichment in WT cells). Results for other two ChIP experiments is shown in Fig S6A. To rule out that the C/EBPβ antibody was binding non-specifically to a highly expressed

genomic region, a so-called "hyperchippable" artefact, we also compared the binding of the C/EBPβ antibody to a known hyperchippable region (the H3F3a promoter region[57]) to the binding of the IgG control. As shown in figure S6B there was no difference in binding of the C/EBPβ antibody to the hyperchippable region compared to the IgG control. Our data confirm that C/EBPβ binds to the *rs35705950* T allele and given the strong T allele specific positive expression effects it suggests that this transcription factor is a candidate key regulator of MUC5B expression in carriers of the T allele in IPF.

Discussion

The results presented in this study suggest that the IPF associated polymorphism, *rs35705950* (T allele), has a direct and major effect on MUC5B expression. We have shown that it is mediated through a combination of effects on both epigenetic and transcriptional regulatory mechanisms. Specifically, we have shown that the T allele directly reduces DNA methylation and that this increases the MUC5B expression in our cell culture models. Furthermore, we have shown that the T allele creates a novel C/EBPβ binding site that strongly positively regulates MUC5B expression.

CpG DNA methylation has previously been identified as an important regulatory mechanism for mucins. The human genes *MUC2*, *MUC5AC*, *MUC5B* and *MUC6* are clustered on chromosome 11p15 and their promoters show a GC-rich structure characterized by Vincent *et al*.[48]. Hypermethylation of *MUC5B* promoter is the major mechanism responsible for its silencing, combined with histone acetylation while the expression of *MUC5AC* was rarely influenced by epigenetic marks. In this study, we confirm a general role of methylation in *MUC5B* repression. Furthermore, we show that the *rs35705950* T allele disrupts a CpG methylation site and this disruption relieves a DNA methylation-dependent negative regulation site. Furthermore, the *in vitro* methylation assay corroborates the differential methylation pattern between alleles.

Data from the ENCODE project suggest that the area surrounding the *MUC5B* polymorphism is located within a cis-regulatory element, based both on the DNAse hypersensitivity cluster and the high number of transcription factors binding in the area. Our data based on *in silico* binding analysis predicts a C/EBPβ binding site overlapping *rs35705950* when the IPF associated T allele is present. Here, we directly show that *rs35705950* polymorphism creates a novel C/EBPβ binding site and that this binding positively regulates MUC5B expression. C/EBPβ would thus be implicated as an important regulator of MUC5B expression in carriers of the *rs35705950* T allele.

Mucin regulating signalling pathways are complex and poorly understood. It has been shown that activation of MEK1/2, PI3K, SPhk1, and MAPK14 (p38α-MAPK) are implicated in IL-13–induced mucus production [58,59]. Furthermore, through PMA signalling PKC, EGF/TGF-a, Ras/Raf, Mek, ERK and Sp-1 signalling pathways have also been associated with MUC5AC and MUC5B expression [60]. Interestingly, C/EBPβ is also overexpressed after PMA stimulus [61] and, in our results, we show a common signalling pathway between MUC5B and C/EBPβ involving IL-13, suggesting common signalling pathways. C/EBPβ has been previously associated with inflammation and immune responses [54,62-64]. Similarly, MUC5B has been implicated in innate immune response in the lung [65], although C/EBPβ has not been implicated so far. C/EBPβ is also expressed in macrophages [66,67]. Recently, Satoh *et al*.[68] identified a monocyte

derived cell, termed segregated-nucleus-containing atypical monocyte (SatM). These SatM cells were shown to be crucial for bleomycin induced fibrosis in a mouse model and C/EBPβ was shown to be a key transcriptional regulator in their differentiation. Interestingly, the SatM-termed monocytes, induce a pro-fibrotic signalling pathway linking them directly to the fibrosis. Furthermore, macrophages have been shown to be related to goblet cell hyperplasia and in that context to induce MUC5B but not MUC5AC in human bronchial epithelial cells [69]. Whether macrophages play any role in dysregulation of MUC5B in IPF remains to be tested.

Other potential transcriptional binding sites are present in the region of the *rs35705950* polymorphic site. Highly preserved HOX9 and a FOXA2 binding domains have been described [70,71]. The binding of FOXA2 transcription factor was shown to be 32bp upstream of *rs35705950* polymorphism. The proximity to the polymorphism may suggest a co-regulatory site that could through, e.g. interaction with the novel C/EBPβ binding site affect MUC5B expression. Whether the overexpression is caused by a combination of FOXA2 and C/EBPβ still needs to be clarified. However, our results are the first to directly show a discriminatory binding of transcriptional regulators at the polymorphic site. The eQTL data showing that the positive expression effects of the *rs35705950* polymorphic site is primarily seen in the lung (Fig S2d) could be related to a lung specific pattern of expression of transcription factors, e.g. C/EBPβ with or without other coregulators such as FOXA2. Other tissues that express high

levels of MUC5B, such as salivary glands, stomach, small and large intestine are not shown to have a positive eQTL with the polymorphic site.

In our results, we have also shown that C/EBPβ is co-expressed with MUC5B. Its presence is predominantly in basal cells in the pseudostratified epithelium under normal conditions, while in fibrotic lungs, it is also expressed in ciliated and goblet cells. This could be related with the role of C/EBPβ previously associated with fibrosis development [72,73].

To summarize, we have shown that T allele associated with the *rs35705950* polymorphism strongly induces MUC5B overexpression. Interestingly, this appears to be through two independent mechanisms. Firstly, it leads to disruption of a CpG methylation site that naturally represses MUC5B, resulting in an overexpression. Secondly, the same T allele creates a novel C/EBPβ binding site that positively regulates MUC5B expression. These results identify the C/EBPβ transcription factor as a potential target to study in fibrosis associated with IPF. Further studies are needed to decipher how C/EBPβ and MUC5B can result in fibrosis.

Materials and Methods

Cell culture

The BCi_NS1.1 cell line is a human bronchial epithelial cell line kindly provided by Dr. Matthew S. Walters, Weill Cornell Medical College, New York NY, USA [74]. It was established by immortalization with retrovirus expressing human telomerase (hTERT). The bronchial epithelial cell line VA-10 was previously established by retroviral transduction of primary bronchial epithelial cells with E6 and E7 viral oncogenes [75]. Both cell lines were cultured in bronchial epithelial growth medium, BEGM (Lonza, Walkersville, MD) supplemented with 50 IU/ml penicillin and 50 μg/ml streptomycin (Gibco, Burlington, Canada).

The human lung adenocarcinoma derived alveolar epithelial cell line A549 (American Type Culture Collection, Rockville MA) was cultured in DMEM-Ham's-F12 basal medium supplemented with 10% fetal bovine serum (FBS), 50 IU/ml penicillin and 50 μg/ml streptomycin (Gibco). All cell lines were cultured at 37°C, 5% CO_2. All three cell lines are WT (homozygous G-allele) at *rs35705950*.

Air-Liquid interface culture

To establish an air-liquid interface cultures (ALI), cells were seeded on the upper layer of Transwell cell culture inserts (Corning®Costar®) pore size 0.4 μm, 12 mm diameter, polyester membrane) (Sigma-Aldrich, St. Louis, USA) at density of 2×10^5 cells per well. The cultures were maintained on chemically defined

bronchial epithelial cell medium (BEGM, Cell Applications, San Diego) for 5 days, 0,5 ml in the upper chamber and 1.5 ml in the lower chamber. After 5 days, medium was changed to DMEM/F-12 (Invitrogen), supplemented with 2% UltroserG (Cergy-Saint-Christophe, France) for additional 5 days. For ALI culture, the medium was aspired from the apical side and the cell layer rinsed 1x with PBS.

Goblet cell differentiation by IL-13 treatment

Cells were cultured for 5 days on BEGM and then for 5 days on DMEM/UG in a submerged culture. After 5 days of ALI culture, IL-13 (Peprotech, London, UK) was added to the basal side to a final concentration of 25 ng/ml and cultured for 14 days.

Immunofluorescence staining

Cells were rinsed twice with chilled PBS. The fixation of cells was performed using 100% methanol at -20°C overnight. Subsequently, cells were submerged in 100% acetone for one minute. Staining was performed using immunofluorescence buffer, (IMF) (0.1% TX-100, 0.15M NaCL, 5nM EDTA, 20mM HEPRES, pH 7.5, 0.02% NaN3 as preservative). Cells were incubated with primary antibody overnight at +4°C, and then rinsed three times for 15 min with IMF buffer. Cells were then incubated with a secondary antibody and DAPI for two hours at room temperature, followed by four times washing with IMF buffer. Cells were mounted using ProLong Antifade (Thermo Fisher Scientific). Antibodies used for this assay are listed in S2

Table. Immunofluorescence was visualized and captured using laser scanning Fluoview® FV1200 Confocal Microscope (Olympus Life Science).

Transient transfection

Cells were grown at 70% confluence one day before transfection. FuGENE® HD Transfection Reagent (Promega) was used on BCi_NS1.1 and VA10 cells, while Lipofectamine (Thermo Fisher Scientific) was used on A549 cells. All transfection was performed following manufacturer's instructions. The results were analysed 48h after transfection. Plasmids used for C/EBPβ overexpression was generously donated by Joan Massague: C/EBPβ LAP isoform (addgene#15738) and C/EBPβ LIP (addgene#15737) [76]. Plasmids used for this assay are listed in S3 Table.

Production of lentiviral and cell transduction

To produce lentiviral cell lines containing pGreenFire1™ Pathway Reporter lentivector (Cat#TR010PA-N and Cat#TR000PA-1, System Bioscience) expressing the *MUC5B* promoter region and controls, we followed the general guideline provided by System Bioscience. Briefly, 70% confluent HEK-293T cells were cultured for 24 h w/o antibiotics and transfected (Lipofectamine, Thermo Fisher Scientific) with lentiviral transfection constructs and packaging plasmids (psPAX2 and pMD2.G) (Addgene plasmids #12260 and #12259, respectively). Culture medium containing the virus was harvested 24 and 48 hours post transfection and centrifuged at 1250 rpm at 4°C

for 5 minutes and filtered through 0,45 μm filter. Lentiviral particle solution was added to culture medium (containing 8 μg/ml polybrene) and then added to culture flasks of 70% confluent cells (BCi_NS1.1, VA10 and A549) at a low multiplicity of infection (MOI) and incubated for 20 hours. cells were then cultured further for 24 hours in fresh culture media. Infected cells were then selected with puromycin or neomycin as appropriate for 48hs. Plasmids used for this assay are listed in S3 Table.

Luciferase Assay

Each cell type was seeded at 70% confluence one day before transfection in a 96 well plate. To perform the luciferase assay, the Dual-Glo® Luciferase Assay System kit supplied by Promega was used, following the general guidelines provided with the kit. Luminosity was measured in a microplate reader Modulus™ II (Turner BioSystem). Luciferase measurement was normalized using Renilla co-transfection. Plasmids used for this assay are listed in S3 Table.

Real Time qPCR

RNA was isolated using Tri-Reagent® solution (Ambion) and cDNA preparation was carried out using RevertAid™ First strand cDNA Synthesis Kit (Fermentas) according to the manufacturer's instructions. Real-time PCR using Power SYBR Green PCR Master mix (Applied Biosystems) was used to detect the relative quantity of each cDNA. *GAPDH* was used as the endogenous reference gene.

Data were analysed using 7500 Software v2.0 (Applied Biosystems). All primers used are listed in S1 Table.

Western Blot

Protein lysates were acquired using RIPA buffer supplemented with phosphatase and protease inhibitor cocktails (Life Technologies). For western blots, 5μg of protein was used per lane, unless otherwise stated. Samples were denatured using Laemli buffer, 10% β-mercaptoethanol at 95°C for 5 min and run on NuPage 10% Bis-Tris gels (Life Technologies) in 2-(N-morpholino) ethanesulfonic acid (MES) running buffer. Samples were then transferred to Immobilon FL PVDF membranes (Millipore). Membranes were blocked in Li-cor blocking buffer and primary antibodies were incubated overnight at 4 °C. Near-infrared fluorescence visualization was measured using Odyssey CLx scanner (Li-Cor, Cambridge, UK). Antibodies used are detailed in S2 Table.

In Vitro Methylation Assay

DNA fragments were cut out of the pGL3-*MUC5B*pr vector using XbaI-EcoRI restriction enzymes, generating 4.1Kb of the *MUC5B* promoter region, with WT or T allele. Fragments were gel-purified using GeneJET PCR Purification Kit (ThermoFisher Scientific) following the manufacturer's instructions and subsequently methylated with M.SssI methyltransferase (New England Biolabs) overnight at 37°C. The methylated fragments were then ligated into the pGL3 basic vector. DNA concentration were measured at 260nm

before being used in transfection as described by Vincent et al [48]. Differential influence of methylation in the 4.1kb *MUC5B* promoter was measured by luciferase activity in three individual experiment performed in triplicates for each transfected cell line. All plasmids used are listed in S3 Table.

Bisulfite sequencing

$1x10^5$ cells were used to extract DNA from each cell line. Extraction was performed with PureLink Genomic DNA MiniKit (Invitrogen) following the manufactures instruction. Bisulfite conversion of DNA was performed with EZ DNA Methylation-Gold™ Kit (ZymoPURE™, Germany). Amplification of the region of interest was done with the EpiMark® Hot Start Taq DNA Polymerase (New England Biolab, UK) and the resulting product was sequenced by Sanger sequencing (Beckman Coulters Genomics, GENEWIZ, UK.) DNA methylation analyses of bisulfite PCR amplicons were performed using Sequence scanner V1.0. DNA methylation level was scored as percentage methylation of individual CpG units in each sample. Primers used for bisulphate sequencing are listed in S1 Table.

5-aza-2'-deoxycytidine DNA Methylation Inhibition

After 10 days of ALI culture, 5-aza-2'-deoxycytidine (Peprotech, London, UK) was added to basal side to a final concentration of 10μM and the cells further cultured for 3 days before being analysed.

siRNA transfection

Small-interfering RNAs (siRNA) targeting human *C/EBPβ* were purchased from Sigma Aldrich (St Louis, MO, USA). siRNA ID: SASI_Hs02_00339146,SASI_Hs01_00236023, SASI_Hs02_00339148, SASI_Hs01_00339149, 10nM each. siRNA transfection was performed following manufactures instructions. MISSION siRNA Universal Negative Control (Sigma Aldrich) was used as a control. Cells were seeded in 96 well plates for Luciferase Assay, or in a 12 well plate to be analysed by a RT-qPCR. siRNA was transfected using Lipofectamine 2000 reagent (Invitrogen) in OPTI-MEM medium (GIBCO). Twenty-four hours later, the transfected cells were transferred to complete medium. After 48h, the cells were harvested and used for luciferase measurement and RT-qPCR.

CRISPR

Cells were seeded in a 12 well plate prior transfection with Cas-9^{WT} vector, gRNA and homologous region and a non-homologous recombination inhibitor (Sigma-Aldrich), at 70% confluence. Selection was performed with Blasticidin. Single cell cloning assays were performed to select individual clones. Sequencing of individual clones was performed to corroborate the genotype (S3c Fig). gRNA and repair template sequences are included in S1 Table.

Chromatin immunoprecipitation (ChIP)

$1x10^{10}$ cells were used to cross-linking. Chromatin immunoprecipitation (ChIP) was done following the protocol

previously described by Boyer et al. [77] with the following modifications: Cross-linking of cells was done at 0.4% formaldehyde (ThermoFisher Scientific). Sonication was done using 300µL sample containing $1x10^{10}$ cells, with 30 cycles of 15'' sonication, 30''cooldown. Primers and antibodies are listed in S1 and S2 Table.

Statistical Analysis

Data are presented as means with standard deviations of measurements unless stated otherwise. Statistical differences between samples were assessed with Student two –tailed T-test. P-values below 0.05 were considered significant (***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq .0.05$).

References

1. Hutchinson, J.P., McKeever, T.M., Fogarty, A.W., Navaratnam, V. & Hubbard, R.B. Increasing global mortality from idiopathic pulmonary fibrosis in the twenty-first century. *Ann Am Thorac Soc* **11**, 1176-85 (2014).
2. Coultas, D.B., Zumwalt, R.E., Black, W.C. & Sobonya, R.E. The epidemiology of interstitial lung diseases. *Am J Respir Crit Care Med* **150**, 967-72 (1994).
3. Hirakawa, H. *et al.* Cathepsin S deficiency confers protection from neonatal hyperoxia-induced lung injury. *Am J Respir Crit Care Med* **176**, 778-85 (2007).
4. Hunninghake, G.M. *et al.* MUC5B promoter polymorphism and interstitial lung abnormalities. *N Engl J Med* **368**, 2192-200 (2013).
5. Olson, A.L. *et al.* Mortality from pulmonary fibrosis increased in the United States from 1992 to 2003. *Am J Respir Crit Care Med* **176**, 277-84 (2007).
6. Raghu, G. *et al.* An official ATS/ERS/JRS/ALAT statement: idiopathic pulmonary fibrosis: evidence-based guidelines for diagnosis and management. *Am J Respir Crit Care Med* **183**, 788-824 (2011).
7. Margaritopoulos, G.A., Vasarmidi, E. & Antoniou, K.M. Pirfenidone in the treatment of idiopathic pulmonary fibrosis: an evidence-based review of its place in therapy. *Core Evid* **11**, 11-22 (2016).
8. Bonella, F., Stowasser, S. & Wollin, L. Idiopathic pulmonary fibrosis: current treatment options and critical appraisal of nintedanib. *Drug Des Devel Ther* **9**, 6407-19 (2015).
9. Hubbard, R. *et al.* Risk of cryptogenic fibrosing alveolitis in metal workers. *Lancet* **355**, 466-7 (2000).
10. Hubbard, R., Lewis, S., Richards, K., Johnston, I. & Britton, J. Occupational exposure to metal or wood dust and aetiology of cryptogenic fibrosing alveolitis. *Lancet* **347**, 284-9 (1996).
11. Iwai, K., Mori, T., Yamada, N., Yamaguchi, M. & Hosoda, Y. Idiopathic pulmonary fibrosis. Epidemiologic approaches to occupational exposure. *Am J Respir Crit Care Med* **150**, 670-5 (1994).
12. Baumgartner, K.B. *et al.* Occupational and environmental risk factors for idiopathic pulmonary fibrosis: a multicenter case-control study. Collaborating Centers. *Am J Epidemiol* **152**, 307-15 (2000).
13. Lawson, W.E. *et al.* Endoplasmic reticulum stress in alveolar epithelial cells is prominent in IPF: association with altered

surfactant protein processing and herpesvirus infection. *Am J Physiol Lung Cell Mol Physiol* **294**, L1119-26 (2008).
14. Stewart, J.P. *et al.* The detection of Epstein-Barr virus DNA in lung tissue from patients with idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* **159**, 1336-41 (1999).
15. Tang, Y.W. *et al.* Herpesvirus DNA is consistently detected in lungs of patients with idiopathic pulmonary fibrosis. *J Clin Microbiol* **41**, 2633-40 (2003).
16. Erwteman, T.M., Braat, M.C. & van Aken, W.G. Interstitial pulmonary fibrosis: a new side effect of practolol. *Br Med J* **2**, 297-8 (1977).
17. Hubbard, R. *et al.* Exposure to commonly prescribed drugs and the etiology of cryptogenic fibrosing alveolitis: a case-control study. *Am J Respir Crit Care Med* **157**, 743-7 (1998).
18. Musk, A.W. & Pollard, J.A. Pindolol and pulmonary fibrosis. *Br Med J* **2**, 581-2 (1979).
19. Baumgartner, K.B., Samet, J.M., Stidley, C.A., Colby, T.V. & Waldron, J.A. Cigarette smoking: a risk factor for idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* **155**, 242-8 (1997).
20. Spira, A. *et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci U S A* **101**, 10143-8 (2004).
21. Steele, M.P. *et al.* Clinical and pathologic features of familial interstitial pneumonia. *Am J Respir Crit Care Med* **172**, 1146-52 (2005).
22. Tsakiri, K.D. *et al.* Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc Natl Acad Sci U S A* **104**, 7552-7 (2007).
23. Armanios, M.Y. *et al.* Telomerase mutations in families with idiopathic pulmonary fibrosis. *N Engl J Med* **356**, 1317-26 (2007).
24. Cogan, J.D. *et al.* Rare variants in RTEL1 are associated with familial interstitial pneumonia. *Am J Respir Crit Care Med* **191**, 646-55 (2015).
25. Stuart, B.D. *et al.* Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat Genet* **47**, 512-7 (2015).
26. Thomas, A.Q. *et al.* Heterozygosity for a surfactant protein C gene mutation associated with usual interstitial pneumonitis and cellular nonspecific interstitial pneumonitis in one kindred. *Am J Respir Crit Care Med* **165**, 1322-8 (2002).

27. van Moorsel, C.H. *et al.* Surfactant protein C mutations are the basis of a significant portion of adult familial pulmonary fibrosis in a dutch cohort. *Am J Respir Crit Care Med* **182**, 1419-25 (2010).
28. Wang, Y. *et al.* Genetic defects in surfactant protein A2 are associated with pulmonary fibrosis and lung cancer. *Am J Hum Genet* **84**, 52-9 (2009).
29. Kropski, J.A., Lawson, W.E., Young, L.R. & Blackwell, T.S. Genetic studies provide clues on the pathogenesis of idiopathic pulmonary fibrosis. *Dis Model Mech* **6**, 9-17 (2013).
30. Noth, I. *et al.* Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *Lancet Respir Med* **1**, 309-317 (2013).
31. Fingerlin, T.E. *et al.* Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. *Nat Genet* **45**, 613-20 (2013).
32. Seibold, M.A. *et al.* A common MUC5B promoter polymorphism and pulmonary fibrosis. *N Engl J Med* **364**, 1503-12 (2011).
33. Evans, C.M. *et al.* Idiopathic Pulmonary Fibrosis: A Genetic Disease That Involves Mucociliary Dysfunction of the Peripheral Airways. *Physiol Rev* **96**, 1567-91 (2016).
34. Praetorius, C. *et al.* A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. *Cell* **155**, 1022-33 (2013).
35. Gupta, R.M. *et al.* A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell* **170**, 522-533 e15 (2017).
36. Tak, Y.G. & Farnham, P.J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).
37. Borie, R. *et al.* The MUC5B variant is associated with idiopathic pulmonary fibrosis but not with systemic sclerosis interstitial lung disease in the European Caucasian population. *PLoS One* **8**, e70621 (2013).
38. Horimasu, Y. *et al.* MUC5B promoter polymorphism in Japanese patients with idiopathic pulmonary fibrosis. *Respirology* **20**, 439-44 (2015).
39. Stock, C.J. *et al.* Mucin 5B promoter polymorphism is associated with idiopathic pulmonary fibrosis but not with development of lung fibrosis in systemic sclerosis or sarcoidosis. *Thorax* **68**, 436-41 (2013).

40. Wei, R. *et al.* Association between MUC5B and TERT polymorphisms and different interstitial lung disease phenotypes. *Transl Res* **163**, 494-502 (2014).
41. Zhang, Y., Noth, I., Garcia, J.G. & Kaminski, N. A variant in the promoter of MUC5B and idiopathic pulmonary fibrosis. *N Engl J Med* **364**, 1576-7 (2011).
42. Peljto, A.L. *et al.* The MUC5B promoter polymorphism is associated with idiopathic pulmonary fibrosis in a Mexican cohort but is rare among Asian ancestries. *Chest* **147**, 460-464 (2015).
43. Wang, C. *et al.* Mucin 5B promoter polymorphism is associated with susceptibility to interstitial lung diseases in Chinese males. *PLoS One* **9**, e104919 (2014).
44. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
45. Peljto, A.L. *et al.* Association between the MUC5B promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *JAMA* **309**, 2232-9 (2013).
46. Roy, M.G. *et al.* Mucin production during prenatal and postnatal murine lung development. *Am J Respir Cell Mol Biol* **44**, 755-60 (2011).
47. Young, H.W. *et al.* Central role of Muc5ac expression in mucous metaplasia and its regulation by conserved 5' elements. *Am J Respir Cell Mol Biol* **37**, 273-90 (2007).
48. Vincent, A. *et al.* Epigenetic regulation (DNA methylation, histone modifications) of the 11p15 mucin genes (MUC2, MUC5AC, MUC5B, MUC6) in epithelial cancer cells. *Oncogene* **26**, 6566-76 (2007).
49. Heazlewood, C.K. *et al.* Aberrant mucin assembly in mice causes endoplasmic reticulum stress and spontaneous inflammation resembling ulcerative colitis. *PLoS Med* **5**, e54 (2008).
50. Romero, F. & Summer, R. Protein Folding and the Challenges of Maintaining Endoplasmic Reticulum Proteostasis in Idiopathic Pulmonary Fibrosis. *Ann Am Thorac Soc* **14**, S410-S413 (2017).
51. Seibold, M.A. *et al.* The idiopathic pulmonary fibrosis honeycomb cyst contains a mucocilary pseudostratified epithelium. *PLoS One* **8**, e58658 (2013).
52. Arason, A.J. *et al.* deltaNp63 has a role in maintaining epithelial integrity in airway epithelium. *PLoS One* **9**, e88683 (2014).
53. Vanoni, S. *et al.* Myeloid-derived NF-kappaB negative regulation of PU.1 and c/EBP-beta-driven pro-inflammatory cytokine

production restrains LPS-induced shock. *Innate Immun* **23**, 175-187 (2017).
54. Chinery, R., Brockman, J.A., Dransfield, D.T. & Coffey, R.J. Antioxidant-induced nuclear translocation of CCAAT/enhancer-binding protein beta. A critical role for protein kinase A-mediated phosphorylation of Ser299. *J Biol Chem* **272**, 30356-61 (1997).
55. Coetzee, S.G., Coetzee, G.A. & Hazelett, D.J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847-9 (2015).
56. Yeh, W.C., Cao, Z., Classon, M. & McKnight, S.L. Cascade regulation of terminal adipocyte differentiation by three members of the C/EBP family of leucine zipper proteins. *Genes Dev* **9**, 168-81 (1995).
57. Krebs, W. Ph.D Thesis, Rheinischen Friedrich-Wilhelms-Universität Bonn (2016).
58. Kono, Y. *et al.* Sphingosine kinase 1 regulates mucin production via ERK phosphorylation. *Pulm Pharmacol Ther* **23**, 36-42 (2010).
59. Atherton, H.C., Jones, G. & Danahay, H. IL-13-induced changes in the goblet cell density of human bronchial epithelial cell cultures: MAP kinase and phosphatidylinositol 3-kinase regulation. *Am J Physiol Lung Cell Mol Physiol* **285**, L730-9 (2003).
60. Hewson, C.A., Edbrooke, M.R. & Johnston, S.L. PMA induces the MUC5AC respiratory mucin in human bronchial epithelial cells, via PKC, EGF/TGF-alpha, Ras/Raf, MEK, ERK and Sp1-dependent mechanisms. *J Mol Biol* **344**, 683-95 (2004).
61. Zhu, Y., Saunders, M.A., Yeh, H., Deng, W.G. & Wu, K.K. Dynamic regulation of cyclooxygenase-2 promoter activity by isoforms of CCAAT/enhancer-binding proteins. *J Biol Chem* **277**, 6923-8 (2002).
62. Pless, O. *et al.* G9a-mediated lysine methylation alters the function of CCAAT/enhancer-binding protein-beta. *J Biol Chem* **283**, 26357-63 (2008).
63. Roy, S.K. *et al.* MEKK1 plays a critical role in activating the transcription factor C/EBP-beta-dependent gene expression in response to IFN-gamma. *Proc Natl Acad Sci U S A* **99**, 7945-50 (2002).
64. Kinoshita, S., Akira, S. & Kishimoto, T. A member of the C/EBP family, NF-IL6 beta, forms a heterodimer and transcriptionally synergizes with NF-IL6. *Proc Natl Acad Sci U S A* **89**, 1473-6 (1992).
65. Roy, M.G. *et al.* Muc5b is required for airway defence. *Nature* **505**, 412-6 (2014).

66. Cain, D.W. *et al.* Identification of a tissue-specific, C/EBPbeta-dependent pathway of differentiation for murine peritoneal macrophages. *J Immunol* **191**, 4665-75 (2013).
67. Tamura, A. *et al.* Accelerated apoptosis of peripheral blood monocytes in Cebpb-deficient mice. *Biochem Biophys Res Commun* **464**, 654-8 (2015).
68. Satoh, T. *et al.* Identification of an atypical monocyte and committed progenitor involved in fibrosis. *Nature* **541**, 96-101 (2017).
69. Silva, M.A. & Bercik, P. Macrophages are related to goblet cell hyperplasia and induce MUC5B but not MUC5AC in human bronchus epithelial cells. *Lab Invest* **92**, 937-48 (2012).
70. Helling, B.A. *et al.* Regulation of MUC5B Expression in Idiopathic Pulmonary Fibrosis. *Am J Respir Cell Mol Biol* **57**, 91-99 (2017).
71. Chen, Y., Zhao, Y.H., Di, Y.P. & Wu, R. Characterization of human mucin 5B gene expression in airway epithelium and the genomic clone of the amino-terminal and 5'-flanking region. *Am J Respir Cell Mol Biol* **25**, 542-53 (2001).
72. Viart, V. *et al.* Phosphorylated C/EBPbeta influences a complex network involving YY1 and USF2 in lung epithelial cells. *PLoS One* **8**, e60211 (2013).
73. Vittal, R. *et al.* Type V collagen induced tolerance suppresses collagen deposition, TGF-beta and associated transcripts in pulmonary fibrosis. *PLoS One* **8**, e76451 (2013).
74. Walters, M.S. *et al.* Generation of a human airway epithelium derived basal cell line with multipotent differentiation capacity. *Respir Res* **14**, 135 (2013).
75. Halldorsson, S. *et al.* Differentiation potential of a basal epithelial cell line established from human bronchial explant. *In Vitro Cell Dev Biol Anim* **43**, 283-9 (2007).
76. Gomis, R.R., Alarcon, C., Nadal, C., Van Poznak, C. & Massague, J. C/EBPbeta at the core of the TGFbeta cytostatic response and its evasion in metastatic breast cancer cells. *Cancer Cell* **10**, 203-14 (2006).
77. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-56 (2005).

End Notes:

Acknowledgements

We would like to thank members of the Stem Cell Research Unit, especially Jennifer Kricker and Bryndis Valdimarsdottir for various help with cell culture and staining. We would like to thank Helgi Isaksson, MD for providing IPF samples. Funding was provided by the Icelandic Research Council project (grant number 141090-051), the University of Iceland Research Fund and Landspitalinn University Hospital Scientific Fund.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions:

Conception and design: AUA, MKM;

Analysis and interpretation: AUA, EM, AJA, OAS, GG, TG, MKM;

Drafting the manuscript for important intellectual content: AUA, EM, MKM

Figure legends

Fig 1. Risk allele (T) is associated with a higher expression of MUC5B. a) Luciferase activity on stable cell lines transduced with 4,1Kb of MUC5B cis-regulatory region. Luminescence was measured after 24h in a monolayer culture. b) WT allele was replaced on A549 cells by CRISPR to get five heterozygous cell lines [G/T]. RT-qPCR shows MUC5B relative expression (mRNA) in different cell lines. Culture was performed on monolayer, 48h before RNA extraction. c) IF staining on A549CRISPR clones (1-5) shows higher MUC5B expression levels compared to A549WT. d) The control A549WT and BCi_NS1.1WT (e) were transfected with Cas9 alone as a control and with the gRNA and subsequently selected by blasticidin to generate a CRISPR pool of cells (A549CRISPR and BCi_NS1.1CRISPR respectively). CRISPR pool of cells shows higher MUC5B expression compared to the WT Cas9 control in A549 (d) and BCi_NS1.1 (e) cell lines. MUC5B overexpression was corroborated by immunofluorescence staining (f) for A549 (top) and BCi_NS1.1 (bottom) cell line. Scale bars = 50μm. (*p<0.05 **p<0.01 and ***p <0.001 with error bars representing SD).

Fig 2. Methylation plays a role in MUC5B upregulation driven by T allele. a) IF staining shows an increased expression of MUC5B (left) and MUC5AC (right) after 48h with 5'AZA'2 treatment compare to the untreated sample under ALI conditions. Scale bar = 10μm. b) Transient transfection luciferase assay with (lower panel) or without (upper panel) in vitro methylation of the 4.1kb cis-regulatory region. Without in vitro methylation no difference is seen in VA10 or A549 cells between T-allele and WT while after in vitro methylation a significant increase is seen in the T-allele carrying luciferase plasmid compared to WT in all three cell lines. c) Schematic representation of a bisulfite sequencing experiment shows the differential methylation on the rs35705950 region between cell lines (c) and

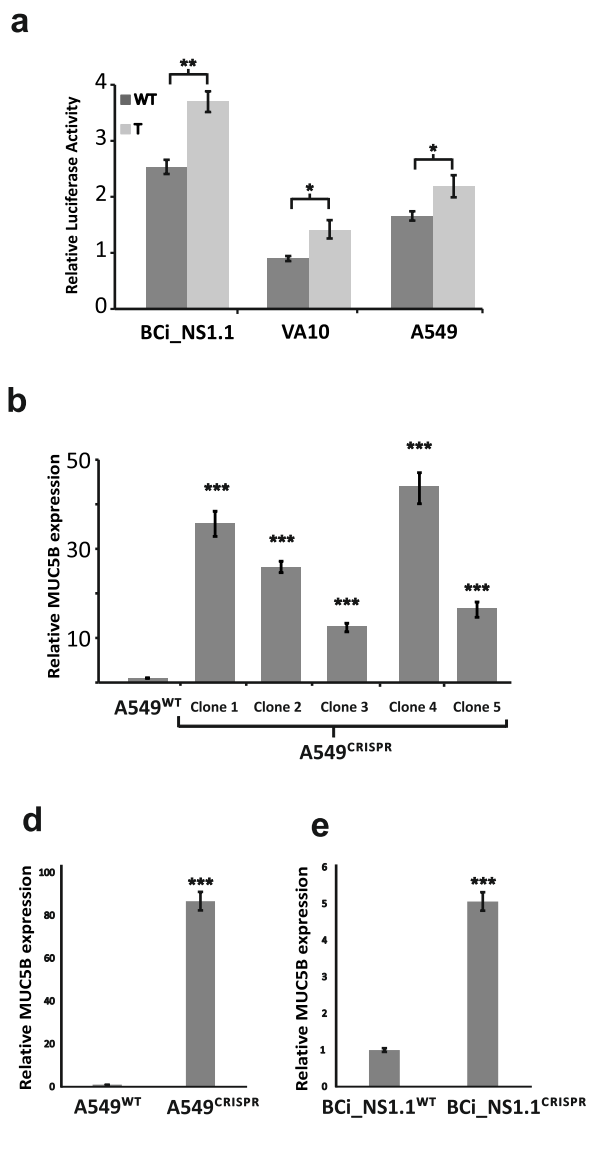
how it changes with the presence of the T allele on CRISPRed cells (d). Legend indicates the range of methylation (*p<0.05 and **p <0.01 with error bars representing SD).

Fig 3. C/EBP is predicted to bind rs35705950 MUC5B cis-regulatory region only in presence of the T allele and C/EBPβ is co-expressed with MUC5B in IPF. a) Results of a weight matrix-based program (Match) for predicting transcription factor binding sites in MUC5B promoter sequence using the DNA sequence flanking rs35705950 using TRANSFAC©. ** indicates a novel binding motif in the T-variant allele. * indicates the loss of a transcription factor binding motif, while (+) and (-) indicates the strand where the bindings occurs. b) IF-P in an IPF sample shows the co-expression of MUC5B (green) and C/EBPβ (red). c) IL-13 (20ng/mL) increases MUC5B (green) and C/EBPβ (red) expression at protein (c) and at RNA level (d), analysed by RT-qPCR, compared to the untreated sample. Scale bars = 50μm.

Fig 4. C/EBPβ mediates the MUC5B upregulation by the rs35705950 T allele. a) The three C/EBPβ isoforms were transfected individually into BCi_NS1.1 (a) and A549 (b) stable cell lines for 24h. The activatory isoform (LAP) increases the MUC5B reporter activity only in presence of the T allele, while the inhibitory isoform (LIP) inhibits the differential activity between WT and T allele. c) siRNA silencing C/EBPβ isoforms, LAP and LIP together or C/EBPβ LAP alone, was performed on A549CRISPR Clone 1. Inhibition of C/EBPβ restores MUC5B expression driven by the T allele to the WT associated MUC5B expression levels. d) siRNA was used to silence C/EBPβ (LAP and LIP or only LAP isoform) in A549 luciferase stable cell lines. Inhibition of C/EBPβ restores MUC5B expression driven by the T allele to the WT associated MUC5B expression levels. e) Chromatin Immunoprecipitation assay shows partial binding of C/EBPβ to the MUC5B cis-regulatory region (rs35705950 region) to

[G/T] genotype A549CRIPSR cell lines but not to [G/G] A549WT genotype. Light grey represents the immunoprecipitation with C/EBPβ antibody, while NS*IgG Isotype where used as an antibody control (dark grey). TNFAIP3 was used as a positive control, and Actin as a negative control. Figure 4e shows a representative ChIP while the other two experiments are included in S6 Figure.

Figure 1


a
Relative Luciferase Activity
WT
T
BCi_NS1.1
VA10
A549
b
Relative MUC5B expression
A549WT
Clone 1
Clone 2
Clone 3
Clone 4
Clone 5
A549CRISPR
d
Relative MUC5B expression
A549WT
A549CRISPR
e
Relative MUC5B expression
BCi_NS1.1WT
BCi_NS1.1CRISPR

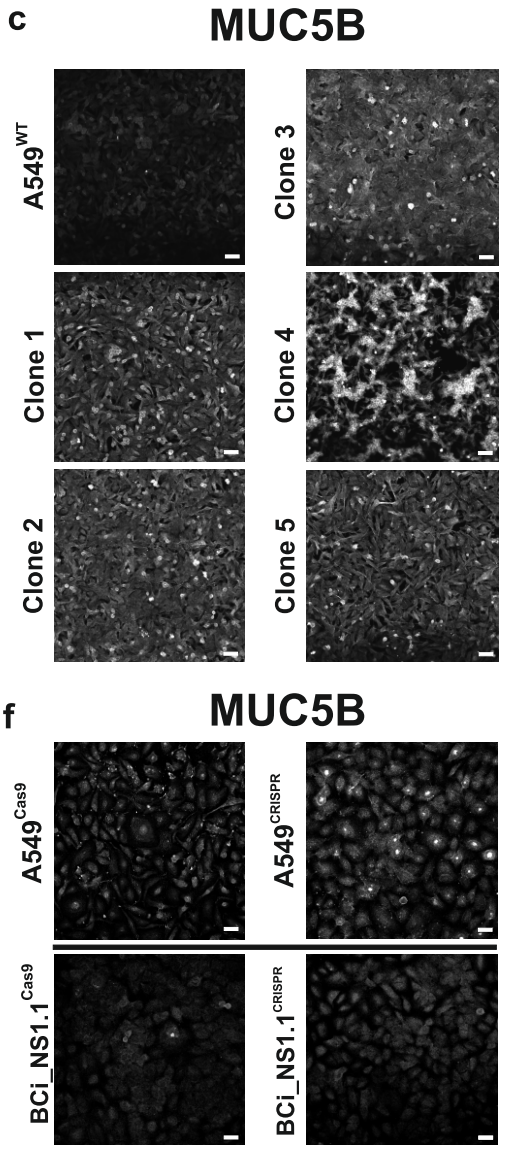

c
MUC5B
A549WT
Clone 3
Clone 1
Clone 4
Clone 2
Clone 5
f
MUC5B
A549Cas9
A549CRISPR
BCi_NS1.1Cas9
BCi_NS1.1CRISPR

Figure 2

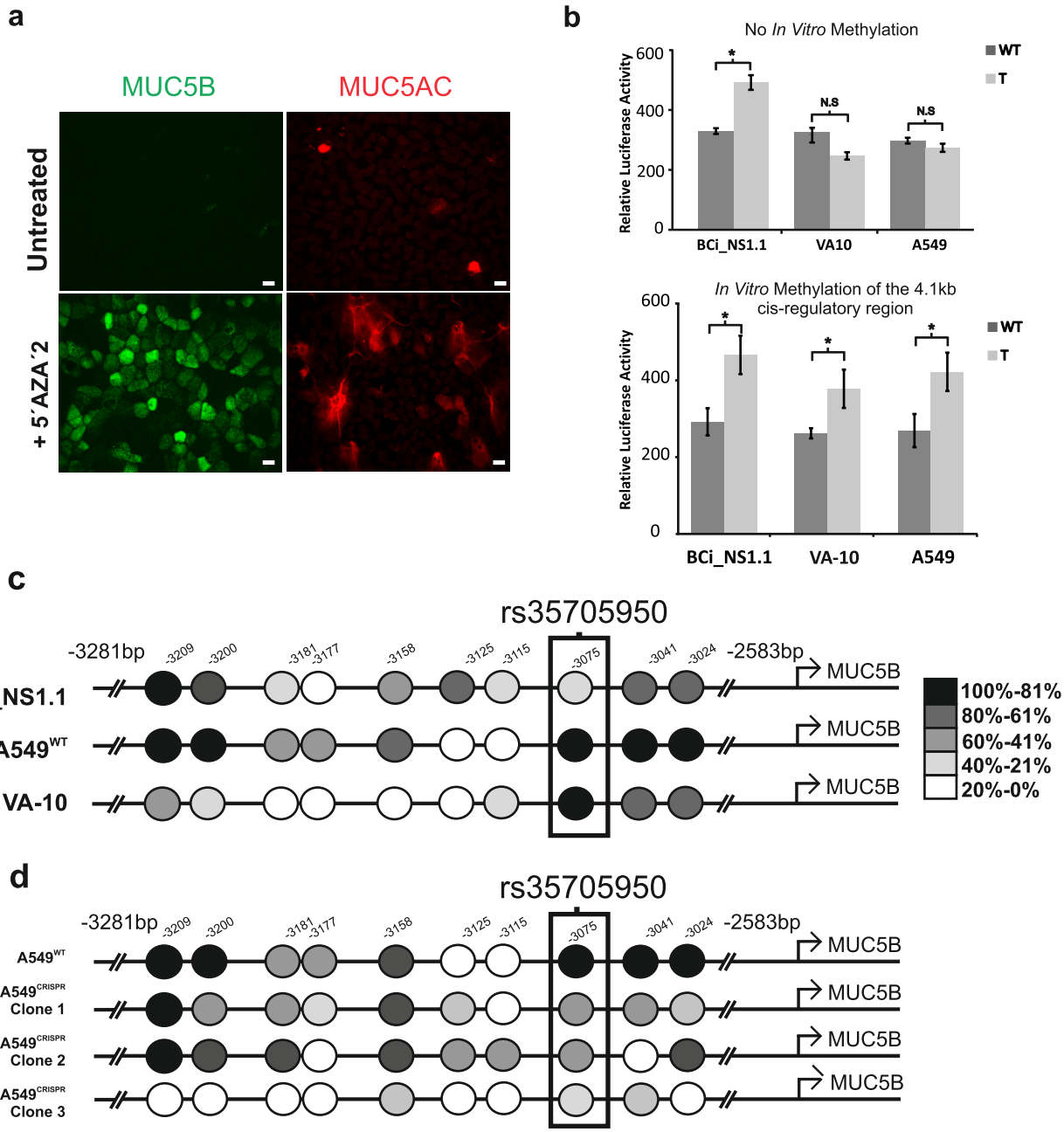

a
MUC5B
MUC5AC
Untreated
+ 5´AZA´2
b
No In Vitro Methylation
Relative Luciferase Activity
600
400
200
0
*
N.S
N.S
WT
T
BCi_NS1.1
VA10
A549
In Vitro Methylation of the 4.1kb cis-regulatory region
Relative Luciferase Activity
600
400
200
0
*
*
*
WT
T
BCi_NS1.1
VA-10
A549
c
rs35705950
-3281bp
-3209
-3200
-3181
-3177
-3158
-3125
-3115
-3075
-3041
-3024
-2583bp
MUC5B
BCi_NS1.1
A549WT
VA-10
100%-81%
80%-61%
60%-41%
40%-21%
20%-0%
d
rs35705950
-3281bp
-3209
-3200
-3181
-3177
-3158
-3125
-3115
-3075
-3041
-3024
-2583bp
MUC5B
A549WT
A549CRISPR Clone 1
A549CRISPR Clone 2
A549CRISPR Clone 3

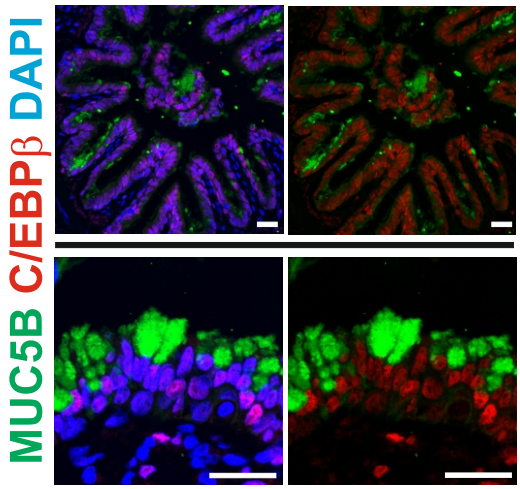
Figure 3

a

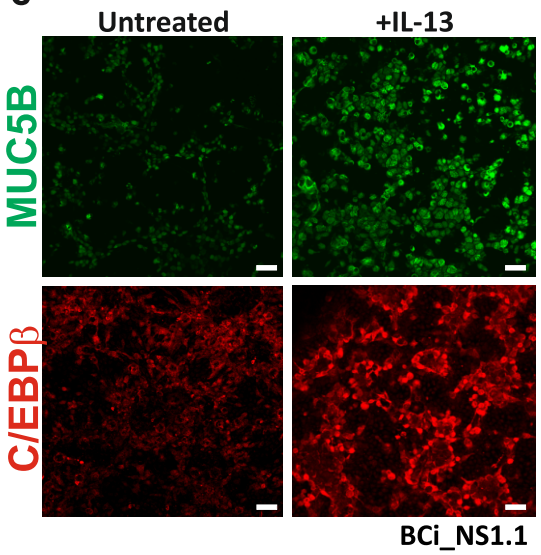
```
                                    gaagAGGTGaac      ZEB-1      (-)
                                  acTGTGAagaggt       C/EBP      (+)
                                caactgtgaAGAGG        KLF4       (+)
                               ttcAACTGtg             c-Myb      (+)
                      tcTGTTT                         SRY        (-)
                   tATCTTctgttttca                    Evi-1      (-)
          ctTTATCttc                                  GATA-3     (-)
CCTTCCTTTATCTTCTGTTTTCAGCGCCTTCAACTGTGAAGAGGTGAACTCTTCAAACA
                                T
                          agctcctTCAACtg              C/EBP      (-) **
                            cgccTTCAActgtgaagag       PAX2       (+)
                            cgcctTCAACtgtgaagaggt      PAX4       (+) *
                             gccttcaactGTGAAgagg       PAX2       (-)

* (disappears with minor allele)
**(appears with minor allele)
```

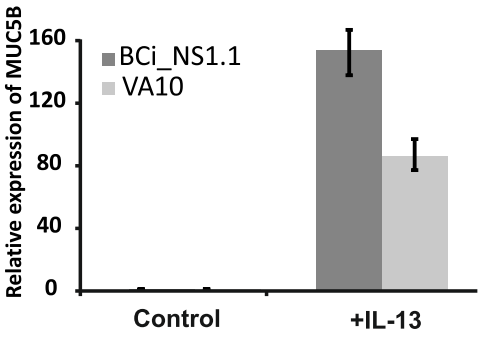
b



c



d



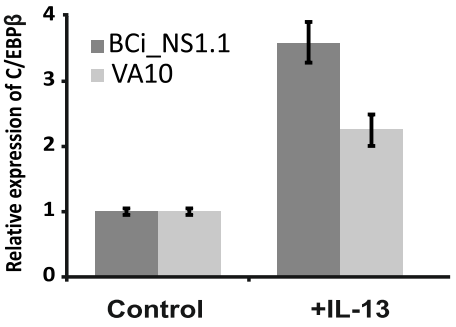
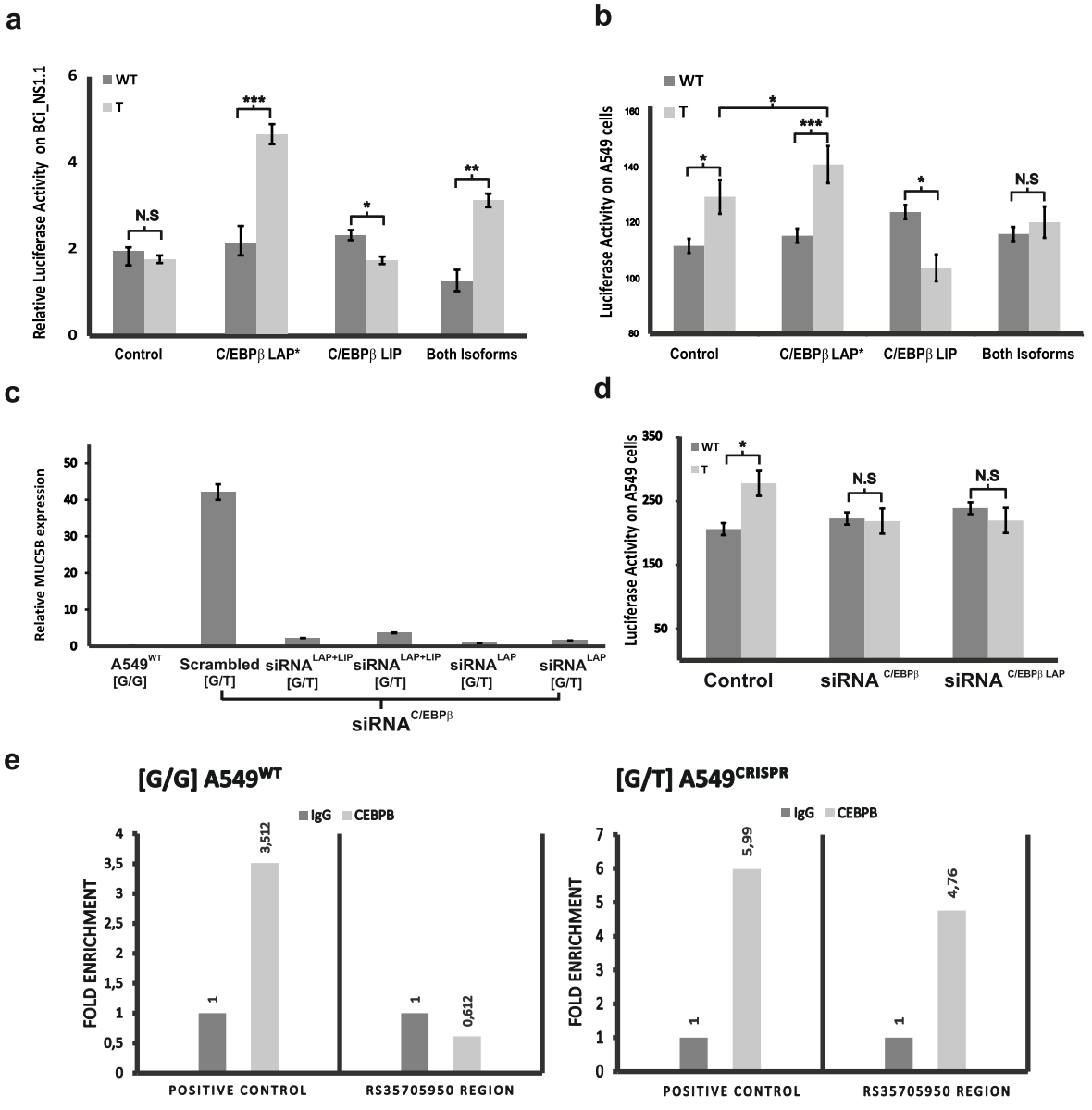


Figure 4

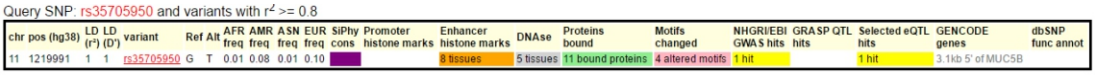

a
Relative Luciferase Activity on BCi_NS1.1
WT
T
N.S
Control
C/EBPβ LAP*
C/EBPβ LIP
Both Isoforms
b
Luciferase Activity on A549 cells
N.S
c
Relative MUC5B expression
A549WT [G/G]
Scrambled [G/T]
siRNALAP+LIP [G/T]
siRNALAP+LIP [G/T]
siRNALAP [G/T]
siRNALAP [G/T]
siRNAC/EBPβ
d
Luciferase Activity on A549 cells
Control
siRNA C/EBPβ
siRNA C/EBPβ LAP
e
[G/G] A549WT
[G/T] A549CRISPR
IgG
CEBPB
FOLD ENRICHMENT
3,512
0,612
5,99
4,76
POSITIVE CONTROL
RS35705950 REGION

A MUC5B polymorphism associated with Idiopathic Pulmonary Fibrosis mediates overexpression through decreased CpG methylation and C/EBPb recruitment driving transcriptional activation

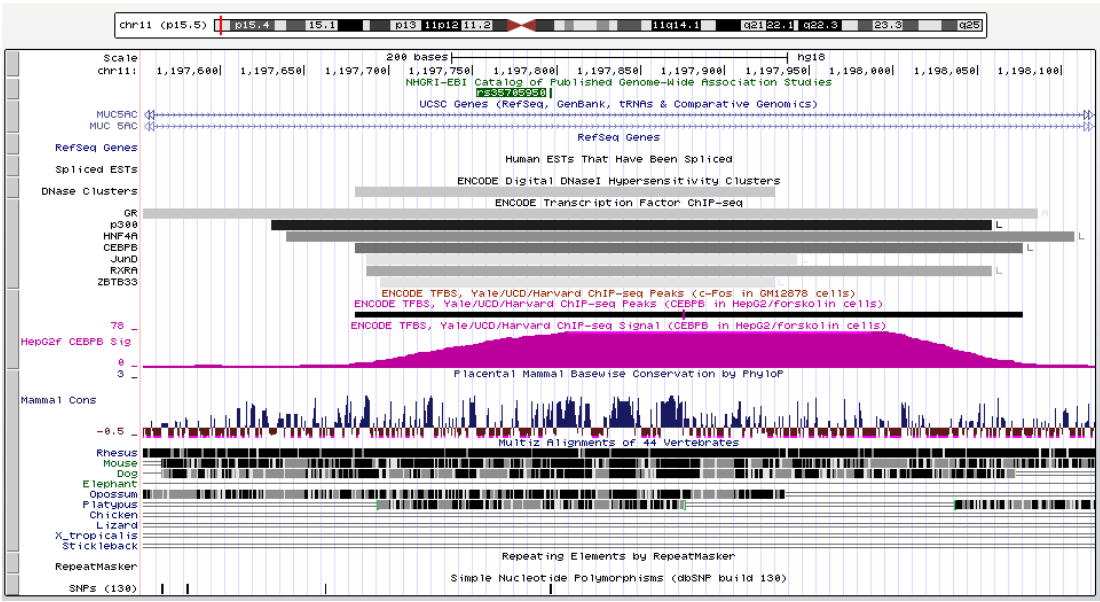
Amaranta U. Armesto, Erna Magnusdottir, Ari J. Arason, Olafur A. Stefansson, Gunnar Gudmundsson, Thorarinn Gudjonsson, Magnus K. Magnusson.

Supplementary material

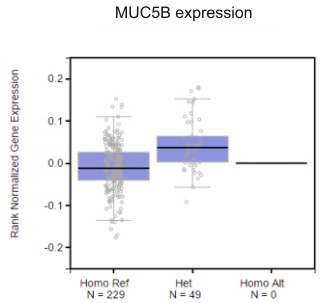
a.



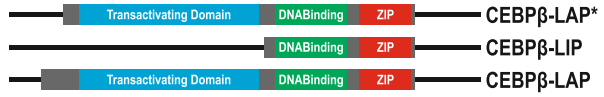
b.



c.

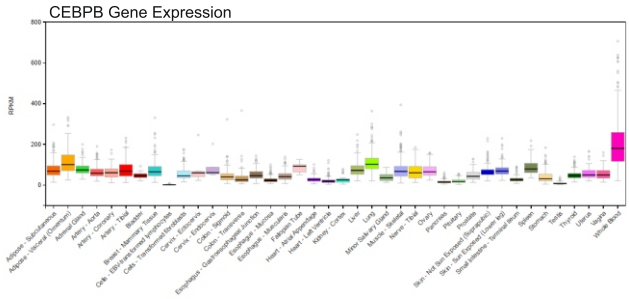


d.

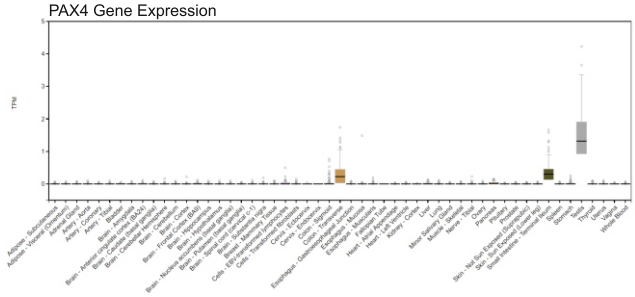


S1 Fig. Schematic representation of rs35705950 on genome browser. Descriptive data for rs35705950 polymorphism obtained from GTEx database (a) and UCSC genome browser (b). c) Differential *MUC5B* expression between homozygous [G/G] and heterozygous [G/T] carriers of rs35705950. Data obtained from GTEx database based on human genotyped samples. d) A schematic representation of the different C/EBPβ isoforms.

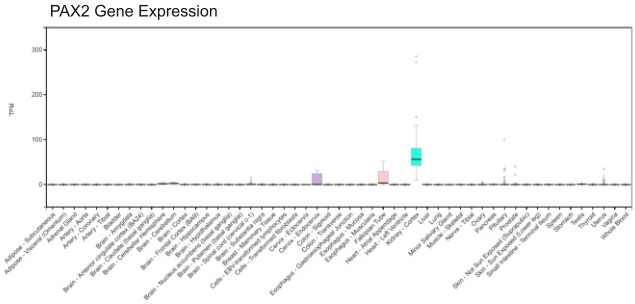
a.



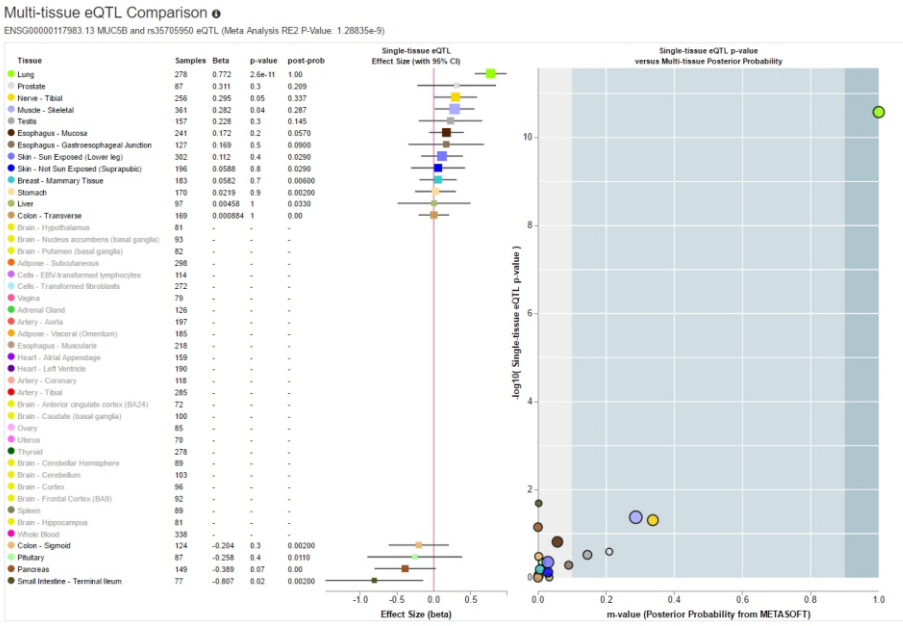
b.



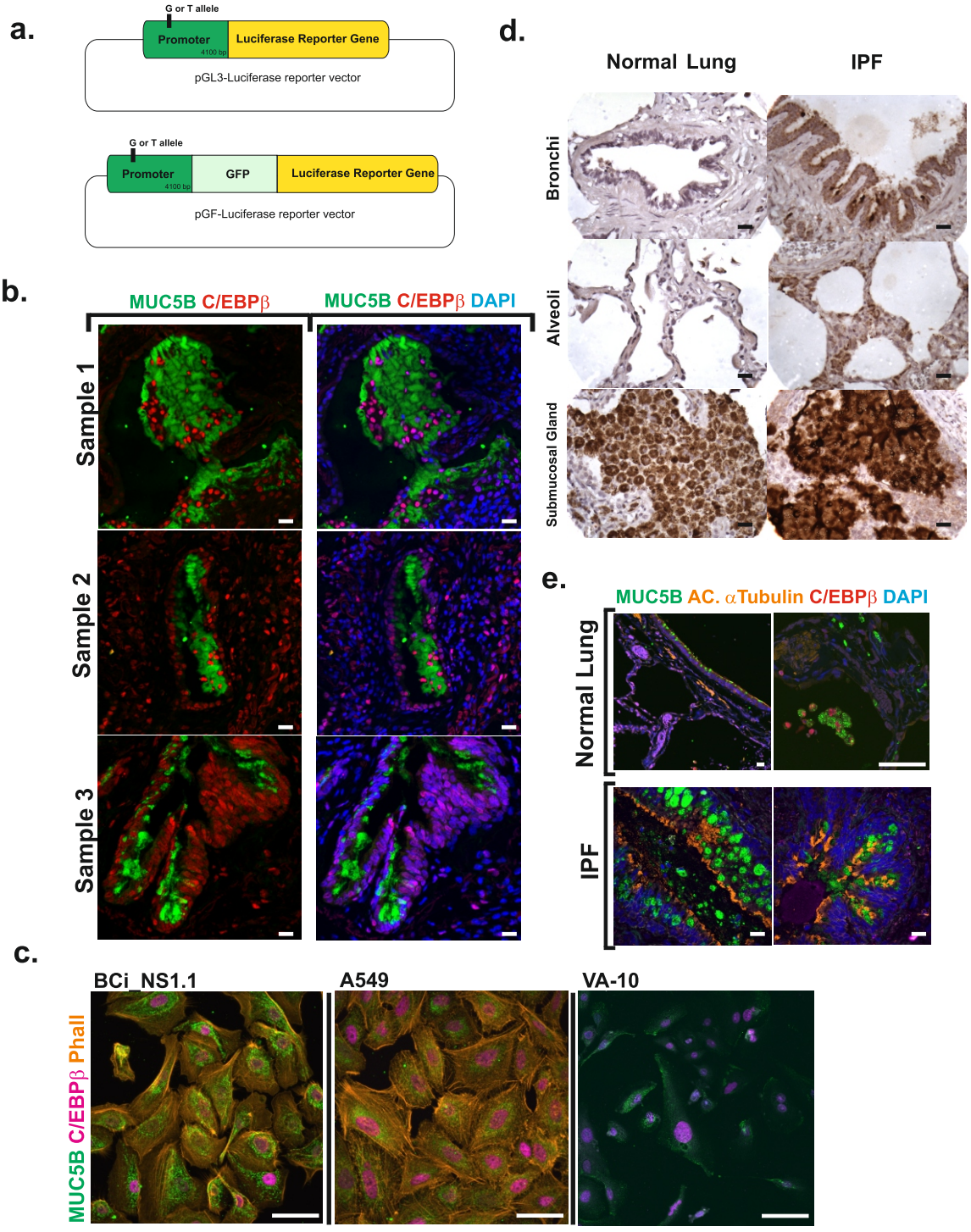
c.



d.

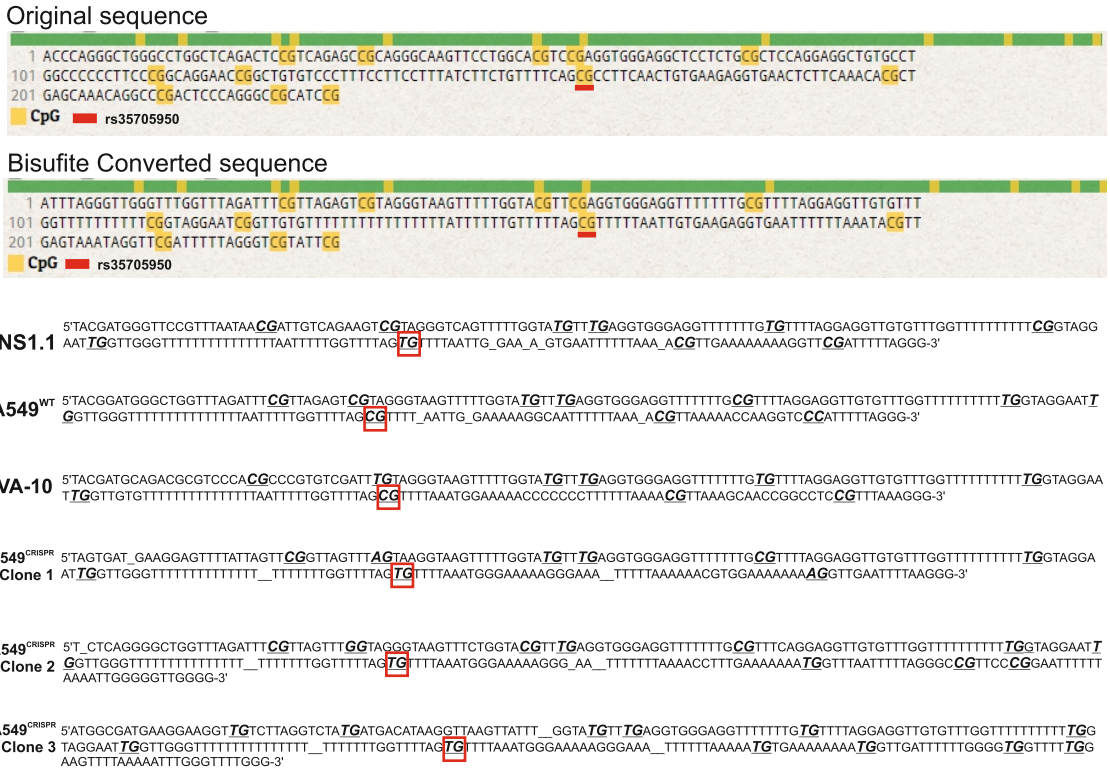


S2 Fig. Analysis of the three different candidates to bind to the MUC5B *cis*-regulatory domain. C/EBPβ (a), PAX4 (b) and PAX2 (c) organ related expression from GTEx. d) rs35705950 polymorphism multi-tissue eQTL plot from GTEx.



S3 Fig. MUC5B and C/EBPβ are co-expressed in IPF and healthy lung samples. a) Schematic view of Luciferase reporter vectors used in Figure 2b (top) and Figure 1a (bottom). b) C/EBPβ and MUC5B are co-expressed in IPF epithelial cells. The pictures are a representation of 3 different samples. c) C/EBPβ (red) and MUC5B (green) are co-expressed in Bci_NS1.1, A549 and VA-10 cell lines. Nucleus (blue) was stained with DAPI. d) MUC5B is overexpress in IPF compared to control samples. e) Comparison of MUC5B and C/EBPβ expression in IPF and control samples. Scale bars = 50μm.

a.



BCi_NS1.1
5'TACGATGGGTTCCGTTTAATAA**CG**ATTGTCAGAAGT**CG**TAGGGTCAGTTTTTGGTA**TG**TT**TG**AGGTGGGAGGTTTTTTTG**TG**TTTTAGGAGGTTGTGTTTGGTTTTTTTTTT**CG**GTAGG
AAT**TG**GTTGGGTTTTTTTTTTTTTTTTAATTTTTGGTTTTAG**TG**TTTAATTG_GAA_A_GTGAATTTTTTAAA_A**CG**TTGAAAAAAAAGGTT**CG**ATTTTTAGGG-3'

A549^{WT}
5'TACGGATGGGCTGGTTTAGATTT**CG**TTAGAGT**CG**TAGGGTAAGTTTTTGGTA**TG**TT**TG**AGGTGGGAGGTTTTTTTG**CG**TTTTAGGAGGTTGTGTTTGGTTTTTTTTTT**TG**GTAGGAAT**T**
GGTTGGGTTTTTTTTTTTTTTTTAATTTTTGGTTTTAG**CG**TTTT_AATTG_GAAAAAGGCAATTTTTTAAA_A**CG**TTAAAAACCAAGGTC**CC**ATTTTTAGGG-3'

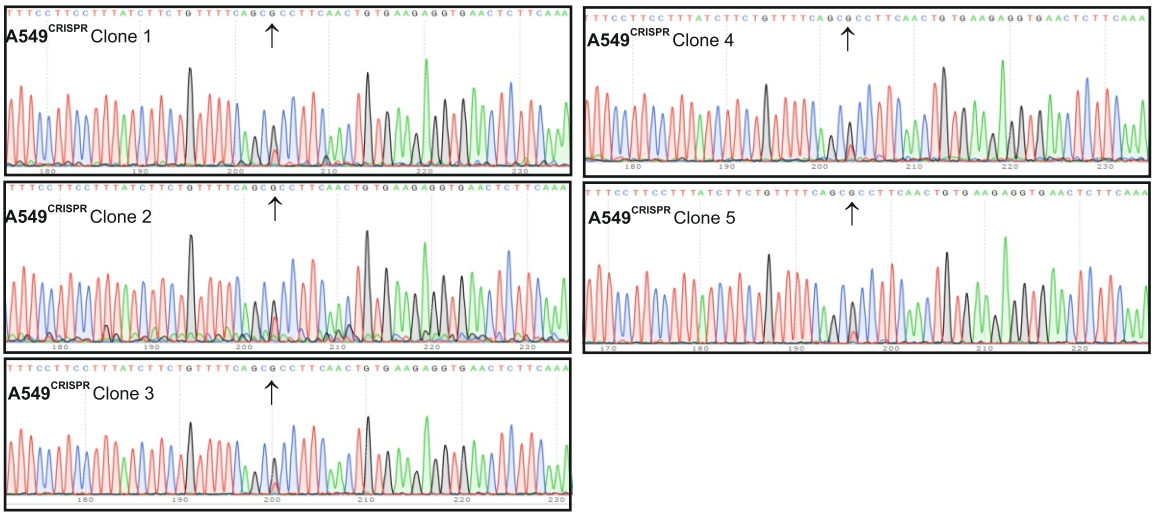
VA-10
5'TACGATGCAGACGCGTCCCA**CG**CCCGTGTCGATT**TG**TAGGGTAAGTTTTTGGTA**TG**TT**TG**AGGTGGGAGGTTTTTTTG**TG**TTTTAGGAGGTTGTGTTTGGTTTTTTTTTT**TG**GTAGGAA
T**TG**GTTGTGTTTTTTTTTTTTTTTTAATTTTTGGTTTTAG**CG**TTTTAAATGGAAAAACCCCCCCTTTTTTAAAA**CG**TTAAAGCAACCGGCCTC**CG**TTTAAAGGG-3'

A549^{CRISPR} Clone 1
5'TAGTGAT_GAAGGAGTTTTATTAGTT**CG**GTTAGTTT**AG**TAAGGTAAGTTTTTGGTA**TG**TT**TG**AGGTGGGAGGTTTTTTTG**CG**TTTTAGGAGGTTGTGTTTGGTTTTTTTTTT**TG**GTAGGA
AT**TG**GTTGGGTTTTTTTTTTTTTTTT__TTTTTTGGTTTTAG**TG**TTTTAAATGGGAAAAAGGGAAA__TTTTTAAAAAACGTGGAAAAAAA**AG**GTTGAATTTTAAGGG-3'

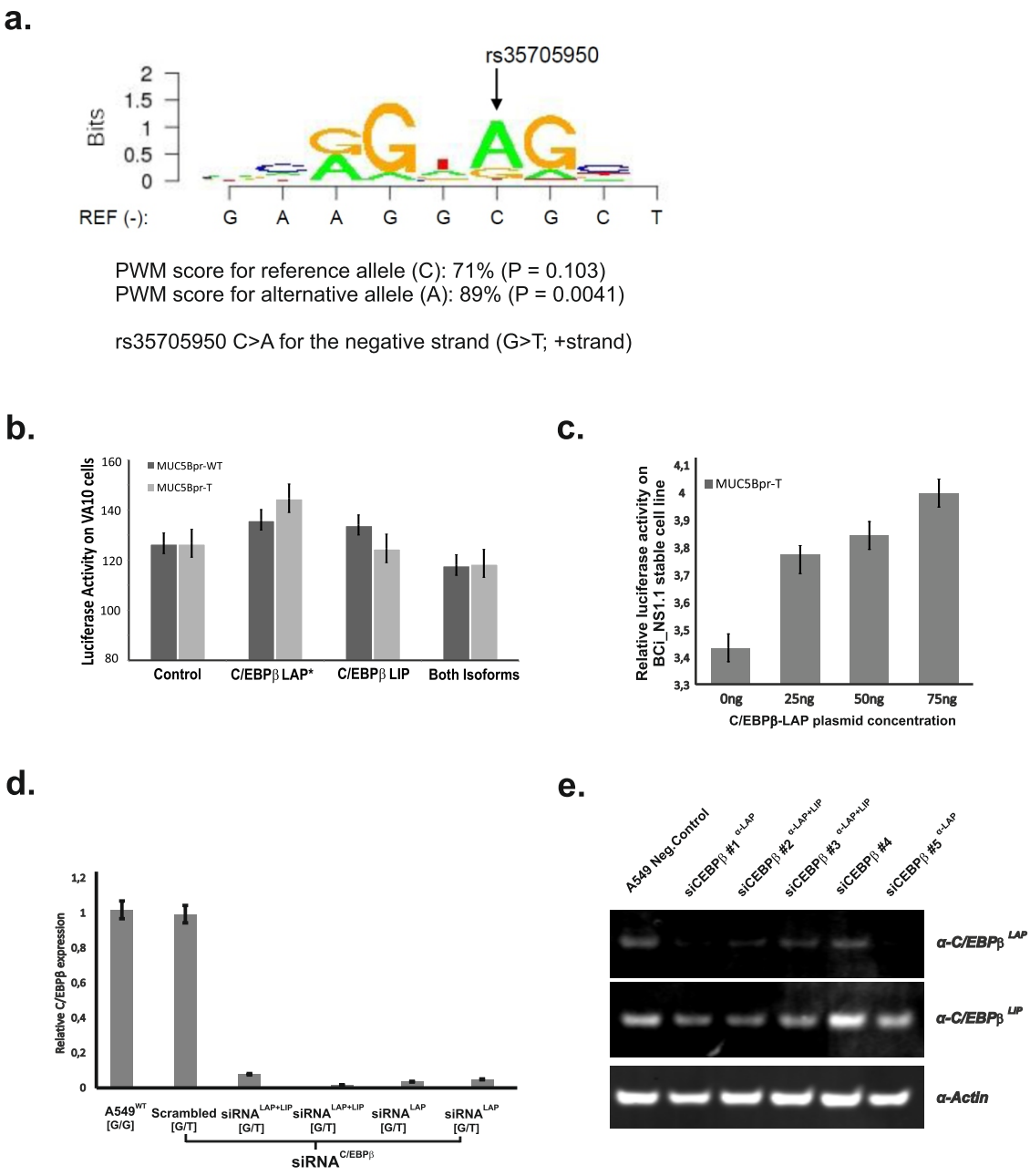
A549^{CRISPR} Clone 2
5'T_CTCAGGGGCTGGTTTAGATTT**CG**TTAGTTT**GG**TAGGGTAAGTTTCTGGTA**CG**TT**TG**AGGTGGGAGGTTTTTTTG**CG**TTTCAGGAGGTTGTGTTTGGTTTTTTTTTT**TG**GTAGGAAT**T**
GGTTGGGTTTTTTTTTTTTTTTT__TTTTTTGGTTTTAG**TG**TTTTAAATGGGAAAAAGGG_AA__TTTTTTAAAACCTTTGAAAAAAA**TG**GTTTAATTTTTAGGGC**CG**TTCC**CG**GAATTTTTT
AAAATTGGGGGTTGGGG-3'

A549^{CRISPR} Clone 3
5'ATGGCGATGAAGGAAGGT**TG**TCTTAGGTCTA**TG**ATGACATAAGGTTAAGTTATTT__GGTA**TG**TT**TG**AGGTGGGAGGTTTTTTTG**TG**TTTTAGGAGGTTGTGTTTGGTTTTTTTTTT**TG**G
TAGGAAT**TG**GTTGGGTTTTTTTTTTTTTTTT__TTTTTTTGGTTTTAG**TG**TTTTAAATGGGAAAAAGGGAAA__TTTTTTAAAAA**TG**TGAAAAAAAA**TG**GTTGATTTTTTGGGG**TG**GTTTT**TG**G
AAGTTTTAAAAATTTGGGTTTTGGG-3'

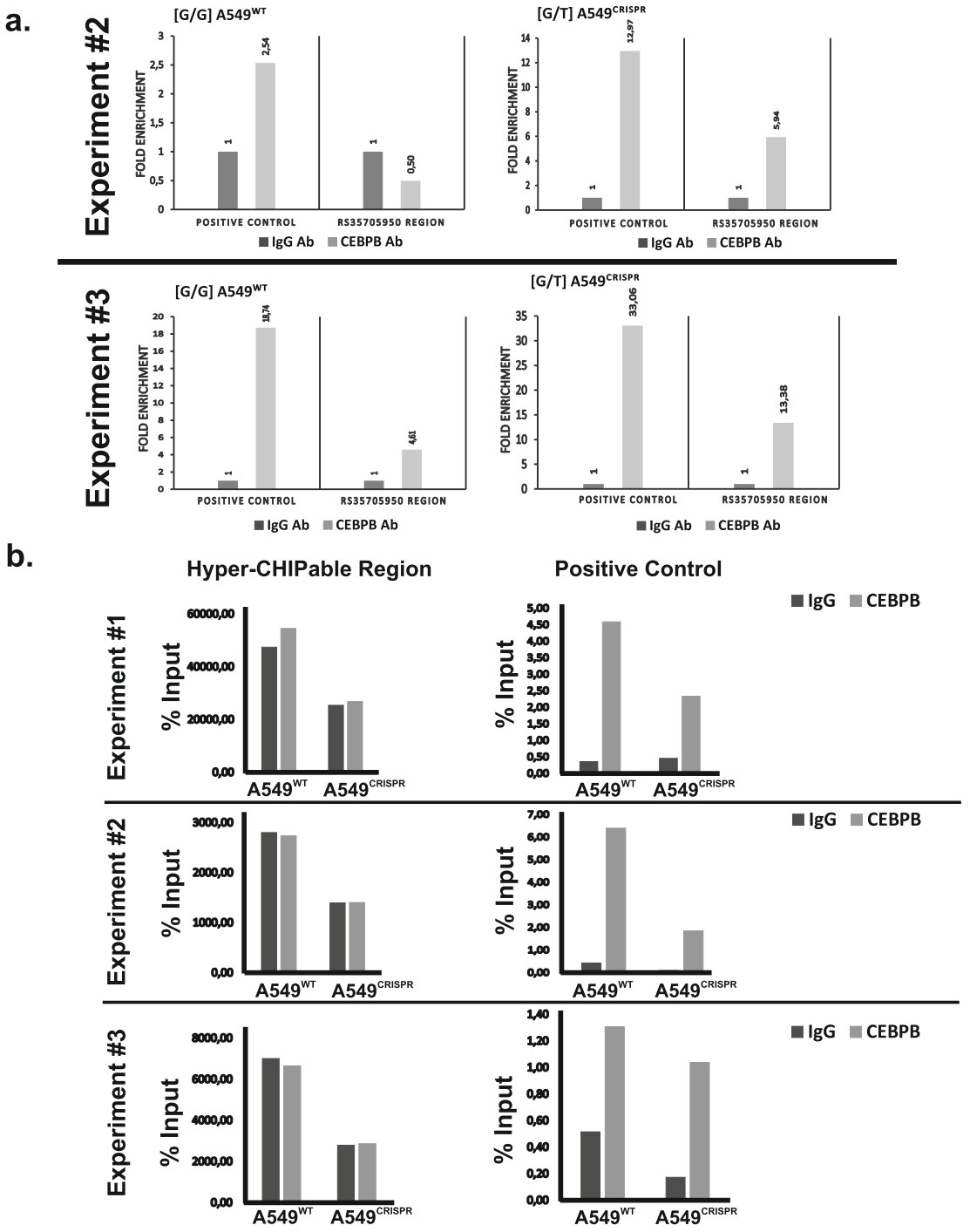
b.



S4 Fig. Sequencing data of bisulfite sequencing experiment and A549^{CRISPR} clones. a) Bisulfite sequencing of BCi_NS1.1, A549, VA10 and A549^{CRISPR} clones (1-5). Yellow box represents CpG site and red line the rs35705950 position, overlapping a CpG site. Bisulfite expected converted sequence was obtained from ZymoResearch Bisulfite Online Tool. b) Genotype of A549^{CRISPR} clones.



S5 Fig. C/EBPβ increases differential MUC5B expression between WT and T alleles. a) The alternative allele for rs35705950 contains a candidate binding site for C/EBPβ on the negative strand. C/EBPβ motif derived from MotifDb (Bioconductor package for R). REF(-)=reference sequence for the negative strand. The position weight matrix (PWM) score for reference allele is 71% while alternative allele has a PWM score of 89%. b) The three C/EBPβ isoforms were transfected individually into VA10 cell line with non-significant difference. c) Three different concentration of C/EBPβ-LAP plasmid were transfected in BCi_NS1.1 luciferase stable cell line, showing a dose dependent MUC5B expression through T allele. d) Corroboration of C/EBPβ relative expression after siRNA knock down in Figure 4c. e) Western blot shows C/EBPβ expression as a corroboration of C/EBPβ siRNA knock down in Figure 4d. siC/EBPβ#4 was discarded and not used on the experiments due to the low knock-down ratio.



S6 Fig. Supplementary data for ChIP experiment presented at Figure 4e. a) ChIP was done in triplicate. A representative ChIP was shown in Figure 4e, while the other two ChiP experiment are represent in S.Figure 6a, comparing C/EBPβ binding to WT and CRISPR A549 cell lines. b) A hyper-ChIPable region was target in ChIP experiment to corroborates the specific detection of C/EBPβ binding sites by IgG and C/EBPβ antibodies. On the left, a hyper-ChIPable region of H3F3a gene was amplified by specific primers in [G/G]A549^{WT} cell line and [G/T] A549 ^{CRISPR} cell line in pools obtained with IgG antibody (dark grey) and C/EBPβ antibody (light grey). Similar results were obtained in the three ChIP experiments, were H3F3a is significantly higher than positive control (represented at the right graph.

S1 Table. Summary of primer sequences.

ChIP	Forward	Reverse
ChIP1-M5Bpr	AGCTATTGAGACATCCCGGA	GCTGTGTCCCTTTCCTTCCT
ChIP2-M5Bpr	GAGTCGGGCCTGTTTGCT	AACCGGCTGTGTCCCTTT
ChIP3-M5Bpr	TGGGAGTCGGGCCTGTTT	CAGGAACCGGCTGTGTCC
ChIP4-M5Bpr	CTGCAGATGACGCTGTCTGT	GGGGCCCCAGCTTATGTAG
Negative Control - GAPDH	TCG AAC AGG AGG AGC AGA GAG CGA	TAC TAG CGG TTT TAC GGG CG
Positive Control - TNFAIP3	GCTGTTGCTCAATTGCTAGTC	CTTCTTGTGCTTACTTTCAGTTCTT
CRISPR	**Forward**	**Reverse**
Sequencing 1	ATGCTACTGGAAGCCTCGAA	CATCAGCTCCCAGGCACT
Sequencing 2	ATGCTACTGGAAGCCTCGAA	CATCAGCTCCCAGGCACT
Sequencing 3	GTTGGACCACAGGCACTGA	CTGCACAGCGACGTGAAC
Sequencing 4	CTGCAGATGACGCTGTCTGT	GGGGCCCCAGCTTATGTAG
Sequencing 5	GAATTCATGCTACTGGAAGCCTCGAA	TCTAGAGGGTCTGTCCCCAGAAGAAC
MUC5B gRNA1	AAAACTGGGAGTCGGGCCTGTTTGC	ACACCGCAAACAGGCCCGACTCCCAG
MUC5B gRNA2	AAAACCAAACACGCTGAGCAAACAGC	ACACCGCTGTTTGCTCAGCGTGTTTG
MUC5B gRNA4	ACACCGGGAGTCGGGCCTGTTTGCT	AAAACAGCAAACAGGCCCGACTCCC
MUC5B gRNA5	AAAACGATGCGGCCCTGGGAGTCGGC	ACACCGCCGACTCCCAGGGCCGCATC
Homologous region 1	TTATCTTCTGTTTTCAGCTCCTTCAACTGTGAAGAGGTGAACTCTTCAAACACGCTGAGCAAACAGGCCCGACTCCCA	
	GGGCCGCATCCGGGATGTCTCAATAGCTGTGGCCTTGACGTCCACCTCGGACCCCTGCCCCGGACCCAGCCCA	
Homologous region 2	AGGTGGACGTCAAGGCCACAGCTATTGAGACATCCCGGATGCGGCCCTGGGAGTCGGGCCTGTTTGCTCAGCGTG	
	TTTGAAGAGTTCACCTCTTCACAGTTGAAGGAGCTGAAAACAGAAGATAAAGGAAGGAAAGGGACACAGCCGGTTC	
Clonning primers	**Forward**	**Reverse**
SEQ1	ATGCTACTGGAAGCCTCGAA	CATCAGCTCCCAGGCACT
SEQ2	GGCTCTGAGCAGACCAAGAG	CTCAGGCAGCTCCTCTGTC
SEQ3	GTTGGACCACAGGCACTGA	CTGCACAGCGACGTGAAC
SEQ4	AGCCATGAGGGGTGACAG	TGACGAGCGTCATCTACAGG
SEQ5	CTGCAGATGACGCTGTCTGT	GGGGCCCCAGCTTATGTAG
MUC5Bpr-pGL3-MluI-NheI	ACGCGTATGCTACTGGAAGCCTCGAA	GCTAGCGGGTCTGTCCCCAGAAGAAC
MUC5Bpr-pGF-XbaI-EcoRI	GAATTCATGCTACTGGAAGCCTCGAA	TCTAGAGGGTCTGTCCCCAGAAGAAC
rs35705950_SEQ_587bp	CGGGTTCTGTGTGGTCTAGG	GCATCAGCGAGATAGCGTTT
Site Directed Mutagenesis [T]	CCTACGAAGGtGACAAAGAGTCCG	TGTGGCAGCGTCCTGGGA
Site Directed Mutagenesis [T]	CCTACGAAGGaGACAAAGAGTCCG	TGTGGCAGCGTCCTGGGA
Bisulfite sequencing	**Fordward**	**Reverse**
Bisulfite Sequencing	TTTGGTTAGAATGAGGGATAGTGAT	CAAAACCACAACTATTAAAACATCC
qRT-PCR	**IDT Identification number**	
MUC5B	Hs.PT.5822513172g	
MUC5AC	Hs.PT.5115096441g	
GAPDH	Hs.PT.39a.22214836	
C/EBPβ	Hs.PT.5827185099g	

S2 Table. Summary of antibodies.

Protein	Identification number
MUC5B	ab87376
MUC5AC	ab3649
C/EBPβ	ab32358
Acethylated α-tubulin	ab11323
Phalloidin	a22283
ChIP	**Identification number**
C/EBPβ	sc-150
Rabbit (DA1E) mAb IgG XP® Isotype Control	#3900

S3 Table. Summary of plasmids.

MUC5B promoter Cloning vectors	Catalogue Number	Company
pGL3-LUC reporter vector	E1751	Promega
Renilla Reporter vector	E2231	Promega
pGF-LUC-GFP Lentivector	TR010PA-N	System Bioscience
pGF-LUC-GFP Negative Control	TR000PA-1	System Bioscience
psPAX2	#12260	Addgene
pMD2.G	#12259	Addgene
CRISPR Vectors	**Catalogue Number**	**Company**
CAS9-WT	#44758	Addgene
gRNA backbone (MLM3636)	#43860	Addgene
C/EBPβ overexpression	**Catalogue Number**	**Company**
C/EBPβ LAP	#15738	Addgene
C/EBPβ LIP	#15737	Addgene