The University Of Sheffield.

Circulating Proteomic Biomarkers in Systemic Sclerosis Related Pulmonary Arterial Hypertension

PhD Thesis

Dr. Peter Mark Hickey

Registration No: 150284442

Donald Heath Clinical Research Fellow

Pulmonary Vascular Research Group

Department of Infection, Immunity and Cardiovascular Disease

Faculty of Medicine, Dentistry and Health

University of Sheffield

Supervisors:

Prof. Allan Lawrie

Dr. Robin Condliffe

Acknowledgements

I would like to express my sincerest gratitude for the guidance and support of Prof. Allan Lawrie who conceived this project, supervised my work and drove the multinational corporate collaboration who had an aligned interest in the initial protein dataset and parallel analysis. Arriving as a clinician with little experience of research method, Prof. Lawrie provided the support required for me to develop the research skills needed to complete this work.

I would like to thank Dr. Robin Condliffe who supported my application and introduced me to the Sheffield Pulmonary Vascular Research Group and the pre-clinical research arm at the University of Sheffield. Dr. Condliffe acted as second supervisor to this project, providing much needed review and oversight during the completion of this work. Dr. Condliffe supervised my clinical work which was required within the funding agreement for my research time.

Thanks to the Donald Heath Research Fellowship for providing the financial support necessary for this work.

Grateful thank to the friends I have made in the Sheffield Pulmonary Vascular Research Group. James Iremonger for guidance and training in statistical coding and for entertaining my constant brainstorming. Grateful thanks to Dr. Josephine Pickworth who provided time and training for the skills required in the laboratory element of my project. Thanks also to the other members of the team who provided support and training during my research: Nadine Arnold, Helen Casbolt, and Dr. Laura West.

Heartfelt thanks to my family – Faye, Maria and Isaac – who have provided unerring love and support during my time in research.

Contents:

List of Figures

List of tables

Glossary of Terms

6MWD:	6-minute walk distance
AIC:	Akaike information criterion
ALK1:	Activin receptor like kinase type 1
AUC:	Area under curve
AU-ROC:	Area under receiver operator curve
BMPR2:	Bone morphogenetic protein receptor type 2
cGMP:	Cyclic guanosine mono-phosphate
CHD:	Congenital heart disease
CI:	Cardiac index
CLEC3B:	Tetranectin
CTD:	Connective tissue disease
CTEPH:	Chronic thromboembolic pulmonary hypertension
CTPA:	Computed tomography pulmonary angiography
ECG:	Electrocardiogram
ECHO:	Echocardiogram
FEV1:	Forced expiratory volume in 1 second
FVC:	Forced vital capacity
GDF-15:	Growth differentiation factor-15
HHT:	Hereditary haemorrhagic telangiectasia
hPAEC:	Human pulmonary artery endothelial cell
HPAH:	Heritable pulmonary arterial hypertension
hPASMC:	Human pulmonary artery smooth muscle cell
HV:	Healthy volunteer
ILD:	Interstitial lung disease
IPAH:	Idiopathic pulmonary arterial hypertension
LA:	Left atrium
LASSO:	Least absolute shrinkage and selection operator
LV:	Left ventricle
mPAP:	Mean pulmonary artery pressure
MRI:	Magnetic resonance imaging

NPV:	Negative predictive value
PAEC:	Pulmonary artery endothelial cell
PAH:	Pulmonary arterial hypertension
PAH-CHD:	Pulmonary arterial hypertension owing to congenital heart disease
PAH-CTD:	Pulmonary arterial hypertension owing to connective tissue disease
PAOP:	Pulmonary artery occlusion pressure (often referred to as “pulmonary capillary wedge pressure” or “PCWP”)
PARK7:	Protein DJ-1
PASMC:	Pulmonary artery smooth muscle cell
PBS:	Phosphate buffered saline
PCA:	Principle component analysis
PH:	Pulmonary hypertension
PH-ILD:	Pulmonary hypertension resulting from interstitial lung disease
PH-LHD:	Pulmonary hypertension resulting from left heart disease
PH-lung:	Pulmonary hypertension resulting from lung disease
PH-misc:	Pulmonary hypertension from miscellaneous causes
PPV:	Positive predictive value
PVOD:	Pulmonary veno-occlusive disease
PVR:	Pulmonary vascular resistance
RA:	Right atrium
RAP:	Right atrial pressure
RHC:	Right heart catheter
ROC:	Receiver operator curve
RV:	Right ventricle
RVSP:	Right ventricular systolic pressure
sPAP:	Systolic pulmonary artery pressure
SSc:	Systemic sclerosis
SSc-no PH:	Systemic sclerosis without pulmonary hypertension
SSc-PAH:	Systemic sclerosis with pulmonary arterial hypertension
SSc-PAH-ILD:	Systemic sclerosis with pulmonary arterial hypertension and an element of interstitial lung disease

STH-Obs:	Sheffield Teaching Hospitals Observational Study of Patients with Pulmonary Hypertension, Cardiovascular and Lung Disease
TGFβ:	Transforming growth factor beta
TL_{CO}:	Transfer factor for carbon monoxide
UoS:	University of Sheffield
VEGF:	Vascular endothelial growth factor
VIF:	Variance inflation factor
Wu:	Woods units

Abstract

Achieving the diagnosis of pulmonary hypertension (PH) is difficult as, given the non-specific nature of symptoms, this condition can masquerade as multiple other conditions, most of which are more common. PH is known to be a complication of certain other disease processes such as systemic sclerosis, and for this reason screening for PH in this disease population is now standard practice. The optimal screening process remains unclear, with multi-modality testing as part of the ERS and DETECT protocols being current practice, but for which improvements are needed.

Survival for patients with systemic sclerosis (SSc) related pulmonary arterial hypertension (PAH) is significantly poorer than in patients with idiopathic pulmonary arterial hypertension. The reasons for this are incompletely understood and may include differences in the nature of underlying pulmonary vasculopathy as well as differences in the ability of the myocardium to compensate for increased right ventricular afterload. A greater understanding of underlying pathophysiological mechanisms, and biomarkers with the ability to aid early diagnosis and guide therapy is therefore needed.

Using pre-treatment serum samples from a tightly phenotyped cohort of systemic sclerosis patients both with and without pulmonary arterial hypertension, through collaboration with industry partners, we have data from a large unbiased proteomic screen of 296 serum protein concentrations. We hypothesise that these data either for individual proteins, or a combined panel of proteins could be used to accurately classify for pulmonary arterial hypertension in these patients.

Several bio-informatic techniques were tested, with a final two-step process separating variable selection from classification modelling. Final modelling was done using logistic regression with backward step-AIC optimisation. A final panel of consisting of 3 serum proteins was derived including Tetranectin, Protein DJ-1 and Growth differentiation factor-15.

When combined in a predictive model, these three proteins classify PAH in SSc with an accuracy of 85%, which compares favourably when measured against the current ERS/ESC guideline screening at 86%, and the DETECT protocol at 74%. When tested in an external validation cohort this model performed well with an AU-ROC 0.79.

Neither Tetranectin nor Protein DJ-1 have previously been described in PAH. We have demonstrated the presence of these proteins in lung tissues of patients with PAH, and have presented cell culture results which go some way to support theoretical mechanisms of action for these proteins in the pathophysiology of this condition.

1 Introduction

1.1 Defining the problem

1.1.1 Pulmonary Circulation

The primary role of the pulmonary circulation is to achieve efficient gas exchange. The pulmonary circulation also has multiple secondary functions such as participation in endocrine control and blood volume reserve. In normal physiology the pulmonary circulation is a high flow with low pressure circuit allowing the right ventricle to operate without significant energy expenditure. Pulmonary blood flow is matched to ventilation for maximum efficiency in gas exchange, through poorly understood oxygen sensing mechanisms coupled to vascular muscle tone. It is poorly adapted for increases in loading pressure such as an acute increase caused by pulmonary thromboembolism, or a chronic increase due to pulmonary hypertension.

1.1.2 Pulmonary Hypertension

Pulmonary hypertension (PH) is a highly heterogeneous group of conditions characterized by a mean pulmonary arterial pressure (mPAP) of at least 25 mmHg. It ranges from very rare conditions such as pulmonary arterial hypertension and chronic thromboembolic pulmonary hypertension (CTEPH) for which there are specific treatment pathways, to a much more common secondary complication of cardiac disease or advanced stages of respiratory disease. Many clinical features and physiological changes are common despite differing aetiology.

1.1.3 Pulmonary Arterial Hypertension

Pulmonary arterial hypertension (PAH) is a rare disease, and is in itself a highly heterogeneous group of conditions, grouped due to similarities in underlying pathophysiology and responses to treatment. It is characterized by the same basic criterion of pulmonary hypertension, with a mPAP $\geq$25 mmHg, but with the added criteria of a pulmonary artery occlusion pressure (PAOP) <15 mmHg, an elevated pulmonary vascular resistance (PVR) >3 Wu (Woods units) in the absence of significant left heart-, lung-, or thromboembolic disease (A revised mPAP criterion of >20mmHg was recently proposed during the 6th World Symposium on Pulmonary Hypertension, Nice, 2018 (Galie et al., 2019)).

It is a life limiting condition, with an untreated life expectancy of 2.8 years.(Dalonzo et al., 1991) Patients experience progressive breathlessness, right heart failure and premature death.

A record of “pulmonary vascular sclerosis” by Ernst Von Romberg in 1891 is believed to be the first description of PAH.(Fishman, 2004) Although understanding of the underlying pathophysiology has progressed substantially since then, the condition remains incompletely understood. It is a disease of the pre-capillary pulmonary vasculature characterized by pathological remodelling of the vascular bed and a subsequent rise in pulmonary pressures.

To date, treatments have been aimed at ameliorating pulmonary vasoconstriction via the prostacyclin, endothelin and cyclic GMP pathways, therefore acting in an approach designed to retard disease progression.(Humbert et al., 2016) No current treatments exist which arrest or reverse the underlying pathological changes.

1.1.4 Clinical Presentation

Diagnosis of pulmonary hypertension requires a high index of suspicion and for the physician to have an awareness of this rare condition. It is based on both the systematic assessment and investigation of the breathless patient and screening of high-risk groups.

By the time of diagnosis, patients have often had multiple contacts with both primary and secondary care, with investigations of multiple other differential diagnoses before this diagnosis is even considered. The features of early PH are non-specific, leading to an average delay in time from symptom onset to diagnosis measured at 2.8 yrs.(Badesch et al., 2010) The diagnosis is often therefore made when once the disease is at an advanced stage.(Edelman, 2007)

The most common presenting symptom of PAH is progressive exertional breathlessness, and as the disease progresses, patients may experience other features of right heart failure such as oedema and ascites due to venous back pressure; exertional pre-syncope or syncope due to reduced cardiac output; or chest pains due to poor coronary perfusion in the setting of cardiac hypertrophy.(Kiely et al., 2013)

1.1.5 Classification

Pulmonary hypertension is a description of the common haemodynamic result of multiple different pathological processes. It is a highly heterogeneous condition with underlying disease processes which can be grouped according to similarities in underlying co-morbidity, similarities in pathophysiology, prognosis and response to treatment.

The first classification of PH was given by the World Health Organization in 1973 (Hatano et al., 1975), but significant advancement in understanding and treatment has led to reform of the classification system, most recently updated during the 6th World Symposium in Nice, France (2018)(Table 1.1).(Simonneau et al., 2019)

Accurate clinical phenotyping depends on rigorous clinical evaluation and appropriate subsequent investigations. Accurate phenotyping is key to defining appropriate treatment pathways, and allows estimation of likely prognosis from analyses done by clinical subgroup shown in the ASPIRE registry (Figure 1.1).(Hurdman et al., 2012)

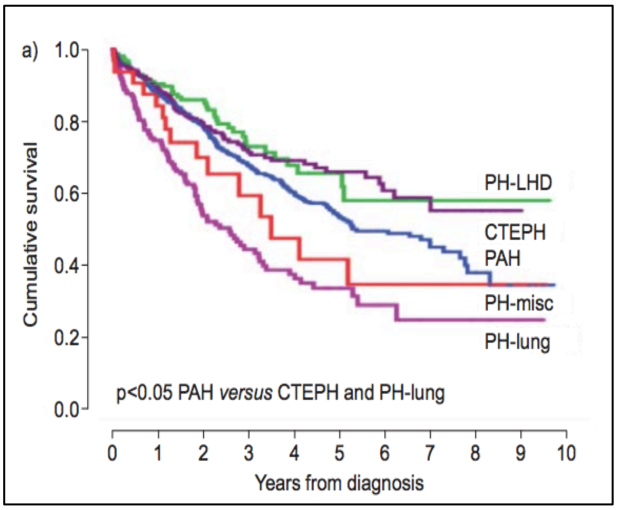


Figure 1.1: Survival by clinical classification. Kaplan Meier curve showing survival for all major groups of pulmonary arterial hypertension from ASPIRE registry follow up data.[1] *Abbreviations: PH-LHD: Pulmonary hypertension due to left heart disease; CTEPH: Chronic thromboembolic pulmonary hypertension; PAH: Pulmonary arterial hypertension; PH-misc: Multifactorial pulmonary hypertension; PH-lung: Pulmonary hypertension due to lung disease.*

[1] Reproduced with permission of the European Respiratory Society ©. European Respiratory Journal Apr 2012, 39 (4) 945-955; DOI: 10.1183/09031936.00078411

Table 1.1: Updated classification of pulmonary hypertension
1. Pulmonary arterial hypertension
1.1. Idiopathic
1.2. Heritable
1.2.1.BMPR-2 mutation
1.2.2.Other mutation
1.3. Drugs and toxin induced
1.4. Associate with:
1.4.1.Connective tissue disease
1.4.2.Human immunodeficiency virus (HIV) infection
1.4.3.Portal hypertension
1.4.4.Congenital heart disease
1.4.5.Schistosomiasis
1.5. PAH long-term responders to calcium channel blockers
1.6. PAH with overt features of venous/capillary (PVOD/PCH) involvement
1.7. Persistent PH of the newborn syndrome
2. Pulmonary hypertension due to left heart disease
2.1. PH due to heart failure with preserved LVEF
2.2. PH due to heart failure with reduced LVEF
2.3. Valvular disease
2.4. Congenital/acquired cardiovascular conditions leading to post-capillary PH
3. Pulmonary hypertension due to lung disease and/or hypoxia
3.1. Obstructive lung disease
3.2. Restrictive lung disease
3.3. Other pulmonary diseases with mixed obstructive/restrictive pattern
3.4. Hypoxia without lung disease
3.5. Developmental lung diseases
4. PH due to pulmonary artery obstructions
4.1. Chronic thromboembolic pulmonary hypertension
4.2. Other pulmonary artery obstructions
4.2.1.Angiosarcoma
4.2.2.Other intravascular tumours
4.2.3.Arteritis
4.2.4.Congenital pulmonary arteries stenosis
4.2.5.Parasites (hydatidosis)
5. Pulmonary hypertension with unclear and/or multifactorial mechanisms
5.1. Haematological disorders: chronic haemolytic anaemia, myeloproliferative disorders, splenectomy
5.2. Systemic and metabolic disorders: sarcoidosis, pulmonary histiocytosis, lymphangioleiomyomatosis, glycogen storage disease, Gaucher disease, thyroid disorders
5.3. Others: pulmonary tumoural thrombotic microangiopathy, fibrosing mediastinitis, chronic renal failure (with or without dialysis), segmental pulmonary hypertension
5.4. Complex congenital heart disease

1.1.6 Epidemiology

Pulmonary arterial hypertension is a rare condition, with an incidence of 1-3.3 per million population for IPAH, and 1.75-3.7 per million for CTEPH.(Kiely et al., 2013) Overall, PAH has a prevalence of 15-52 per million population.(Kiely et al., 2013, Ling et al., 2012) Although rare in the general population, awareness and understanding of PAH is important for physicians managing several associated conditions within which the prevalence is much higher, systemic sclerosis (9%), portal hypertension (2-6%), congenital heart disease (5-10%) and HIV (0.5%).(Avouac et al., 2010, Colle et al., 2003, Gatzoulis et al., 2009, Hadengue et al., 1991, Sitbon et al., 2008)

Within the PAH group, prevalence of sub-classes are shown in Figure 1.2, demonstrating that although the majority of research into PAH has been conducted in idiopathic disease, there is a very significant burden of connective tissue disease associated PH. (PH, 2014)

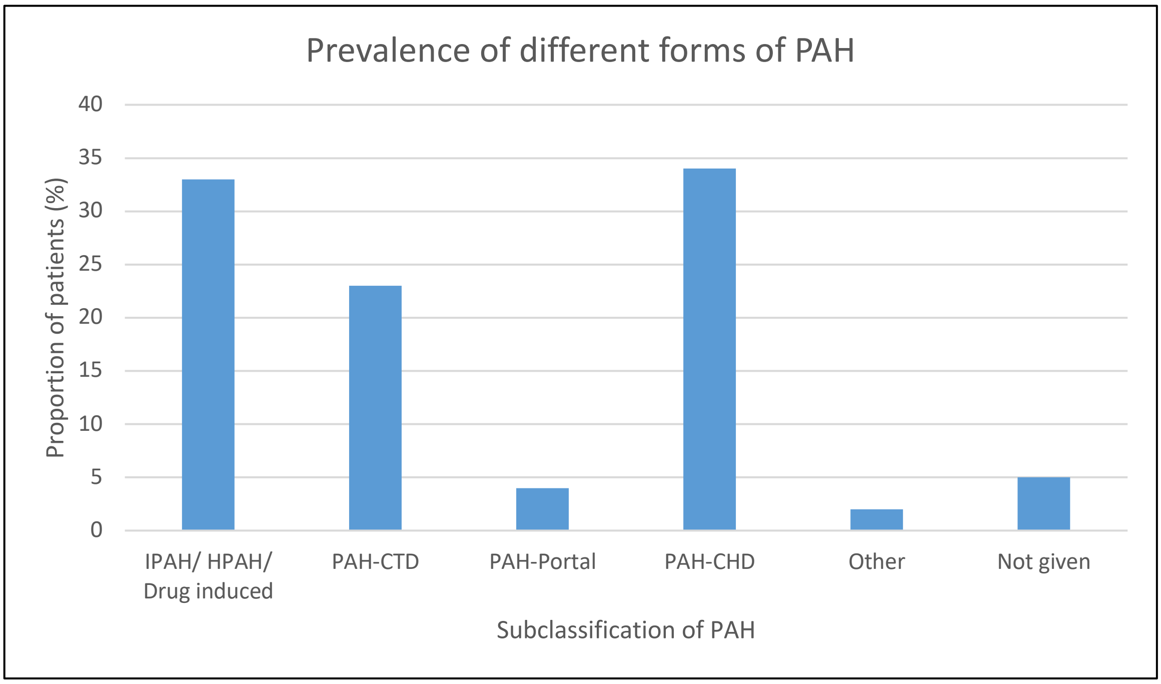


Figure 1.2: Prevalence of subclasses of PAH within PH specialist centres. Data from the National audit of pulmonary hypertension. 2014. Connective tissue disease accounts for 24% of patients in specialist centres. Abbreviations: IPAH: idiopathic pulmonary arterial hypertension; HPAH: heritable pulmonary arterial hypertension; PAH-CTD: Pulmonary arterial hypertension associated with connective tissue disease; PAH-Portal: Pulmonary arterial hypertension associated with portal hypertension; PAH-CHD: Pulmonary arterial hypertension associated with congenital heart disease (+Eisenmengers).

1.1.7 Pathophysiology of PAH

Pathological changes consistent with pulmonary arterial hypertension were first described in post-mortem specimens by Ernst von Romberg, although reports suggest that 'Primary pulmonary hypertension' was not fully described until a syndrome of breathlessness, cyanosis and polycythaemia was recognised to be in association with these changes by Dr. Abel Ayerza in 1901.(Fishman, 2004)

In 1958, Heath and Edwards (Heath et al., 1958) described grading of histological changes observed in patient with PAH (mainly CHD):

1. Medial hypertrophy
2. Medial hypertrophy + intimal proliferation
3. Medial hypertrophy, intimal proliferation and intimal fibrosis
4. Progressive vascular dilatation and fibrotic luminal occlusion
5. Plexiform lesions, cavernous lesions and dilation lesions
6. Necrotizing arteritis.

These changes were confirmed and developed by Wagenvoort and Wagenvoort, who demonstrated the onion skin type changes from ongoing intimal proliferation and fibrosis which is seen in adult PAH, but not paediatric PAH or CTEPH.(Wagenvoo.Ca et al., 1970) They also noted that plexiform lesions are characteristic of PAH, and not seen in CTEPH.

Although initially thought to be a vasoconstrictive disorder, it is now well accepted that PAH is a disorder characterised by dysregulated inflammation and cellular proliferation, within all three tissue layers (intima, media and adventitia) of the pulmonary artery. The cause of dysregulation is unknown, however it is thought that a combination of genetic predisposition, abnormality of endothelial function, inflammation, autoimmune factors, and possibly even viral insults may play a role.(Budhiraja et al., 2004, Cool et al., 2011, Deng et al., 2000, Nicolls et al., 2005, Price et al., 2012)

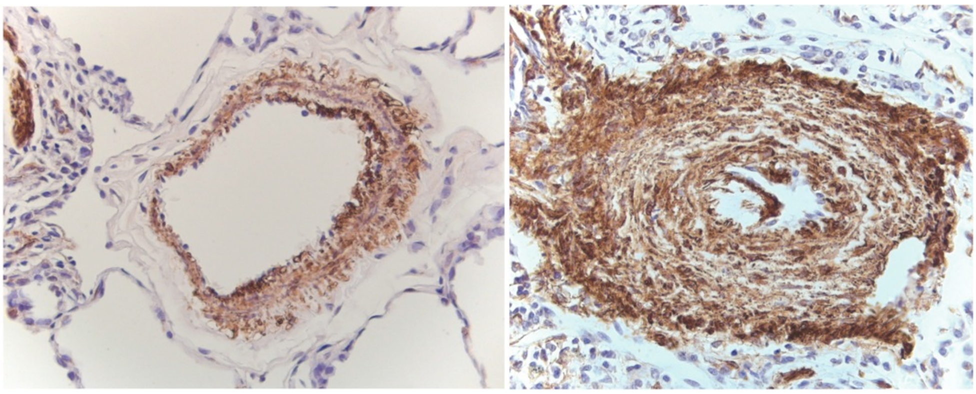


Figure 1.3: Histological changes with smooth muscle hypertrophy in PAH.[2] *Sections from normal pulmonary artery (left) and hypertrophied smooth muscle from a subject with pulmonary arterial hypertension (right) stained with smooth muscle actin.*

1.1.8 Defining Systemic Sclerosis

Systemic sclerosis is a disease characterised by dysregulated collagen deposition in the tissues leading to the disease phenotype. The underlying mechanisms remain incompletely understood, however autoimmunity and pathological inflammation are key components. Typical autoantibody profiles can be detected in 95% of patients with SSc at first presentation.(Steen, 2005)

Cellular mechanisms include abnormalities in the function of endothelial, fibroblast and immunological cell lines.

Endothelial cell dysfunction has been considered as one of the early changes in SSc, and potentially involved in the initiation of disease.(Gabrielli et al., 2009) Endothelial dysfunction promotes vasoconstriction and fibrogenesis through manipulation of counterbalanced cytokines. Endothelin-1 (ET-1) is known to be upregulated, leading to vasoconstriction, and is also known to stimulate fibroblasts to increase collagen deposition.(Abraham et al., 2007, Horstmeyer et al., 2005)

[2] Reprinted from The American Journal of Pathology, Vol. 172 (1), Lawrie A et al, Evidence of a Role for Osteoprotegerin in the Pathogenesis of Pulmonary Arterial Hypertension, Pg: 256-64, 2008, with permission from Elsevier.

Fibroblasts play a key role in the maintenance of the extracellular matrix, balancing the deposition of fibrillar procollagens and fibronectin with control of matrix degradation by proteases.(Abraham et al., 2009) In SSc, fibroblasts are inappropriately activated by pro-fibrotic and pro-inflammatory cytokines and are responsible for the aberrant deposition of connective tissues.

1.1.9 SSc-PAH and distinction from idiopathic disease

Pulmonary arterial hypertension represents a significant complication of systemic sclerosis, affecting between 7-12% of patients with this condition. (Hachulla et al., 2005, Mukerjee et al., 2003) It is one of the leading causes of death in this patient cohort, accounting for 26% of deaths attributable to the effects of the SSc disease process.(Tyndall et al., 2010) It accounts for between 15-20% of all PAH in Europe(Humbert et al., 2006), and given geographical differences, is likely to be a little higher in the US.

Although grouped with, and long considered to have similar pathological characteristics as IPAH, it has become clear that systemic sclerosis related pulmonary arterial hypertension (SSc-PAH) is a very distinct disease from IPAH. SSc-PAH has a higher mortality (Median survival IPAH 7.8 yrs vs SSc-PAH 3 yrs (Ramjug et al., 2017)), and remains less responsive to treatment.(Rhee et al., 2015) Histologically, Overbeek et al found significant differences between the vasculopathy of SSc-PAH and that of IPAH. A plexogenic vasculopathy is characteristic of IPAH, however plexiform lesions were much less apparent in the vasculopathy of SSc-PAH.(Cool et al., 1997, Overbeek et al., 2009) Intimal fibrosis of the vascular tree was found to affect all levels of the arterial tree in SSc-PAH, and importantly was also noted to affect the pulmonary venule in keeping with more pulmonary veno-occlusive type disease (PVOD) in this patient cohort and was not seen in IPAH.(Overbeek et al., 2009)

1.2 Biomarkers, biomarkers panels and screening for pulmonary hypertension

1.2.1 Biomarkers

The biomarker definitions working group has defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to therapeutic interventions".(Biomarkers Definitions Working, 2001)

Further to this, a good biomarker in the assessment of pulmonary hypertension should be easily measurable, and stable such that it is not influenced by other factors such as dietary changes, or renal clearance. For diagnostic utility a biomarker should be measurable for the early detection of PAH, rather than just as a marker of disease progression.

1.2.2 Proteomics in PAH

Previous work by Rhodes et al has investigated the utility of proteomic biomarkers in patients in the wider group of PAH.(Rhodes et al., 2017) It was noted in this international study, similar to our proposed work, that this would include a heterogenous disease group, however to control for this it focussed primarily on patients with IPAH and HPAH which have a similar prevalence to those with CTD-PAH. Protein measurements were determined using the SOMAscan aptamer based platform.(Gold et al., 2010) This study was designed to investigate the prognostic potential of proteins in a prevalent group of patients with PAH.

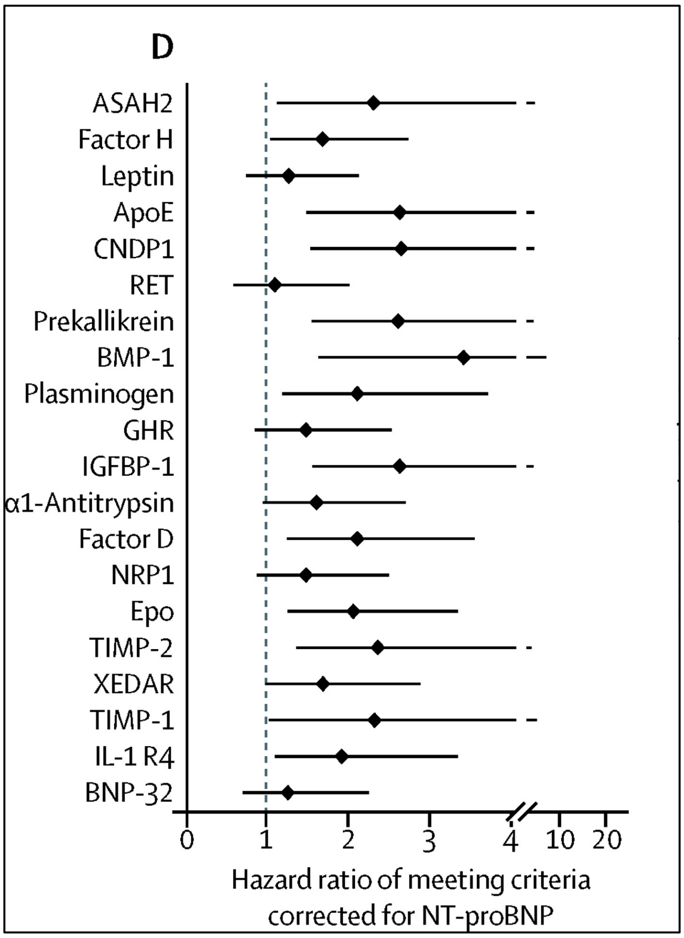


Figure 1.4: Hazard ratios and 95% CI from Cox regression analysis comparing 20 prognostic proteins with established prognostic marker, NT-proBNP.

Figure 1.4 is included as an excerpt from the study results to show the top 20 prognostic proteins identified as an outcome from this work. Among these proteins are those from the complement family suggesting a possible link to immune regulation; growth factors which as a family have previously been identified as playing a key role in the development of PAH; members of the metalloproteinase family involved in regulation of the extracellular matrix; and plasminogen. These protein groups are of particular interest as either these, or protein groups with which they commonly interact, feature strongly in the outcomes of my work in protein selection to be presented in section 4.5.3.5 – Final protein selection.

1.2.3 Proteomics in SSc-PAH

Abnormal concentrations of many potential protein biomarkers have been reported in both the tissue and the circulating compartments in patients with PAH. Fewer reports have been made for protein concentrations specifically in SSc-PAH. The following summary was prepared as a literature review for this thesis, and published by our group as a mini review of potential protein biomarkers in SSc-PAH.(Hickey et al., 2018)

1.2.3.1 Literature review method

To identify suitable primary research articles on this topic, a literature search was conducted using Ovid Medline and PubMed. Keywords used were “Systemic sclerosis”, “Scleroderma”, “Pulmonary hypertension”, “Pulmonary arterial hypertension”, “Protein”, and “Biomarker”. Date of publication was limited from 1990 to present day. 148 publications were returned from this search.

We included studies identifying a cohort of patients diagnosed with systemic sclerosis with pulmonary arterial hypertension with comparator groups including healthy volunteers (HV), systemic sclerosis without pulmonary hypertension (SSc-no PAH) and/or idiopathic pulmonary arterial hypertension.

Studies were included if they reported data on differential protein expression between subgroups which were related to objective measurements of pulmonary hypertension.

Table 1.2: Summary of potential protein biomarkers in SSc-PAH

Protein	Comparison Groups	Number of Patients	Outcome	Correlations in SSc-PAH	Reference
NT-proBNP	SSc-PAH vs SSc	109	Significantly higher in SSc-PAH vs SSc. Sens 55.9%, Spec 95.1%. Correlated with invasive haemodynamics	mPAP (r=0.62; p<0.0001) PVR (r=0.81; p<0.0001)	(Williams et al. 2006)
	SSc-PAH vs SSc	329	NT-proBNP superior to BNP for detection of PAH in SSc		(Chung et al. 2017)
	SSc-PAH vs IPAH	98	Significantly higher in SSc-PAH, correlated with haemodynamics and predicted survival in SSc-PAH group.	CI (r=-0.58; p<0.01) PVR (r=0.54; p<0.01)	(Mathai et al. 2010)
Endoglin	SSc-PAH vs SSc vs HV	60	Serum levels significantly higher in SSc-PAH than control		(Coral-Alvarado et al. 2010)
sFLT-1	SSc-PAH vs SSc	77	Plasma levels significantly higher in SSc-PAH and correlate with RVSP and inversely with DL_{CO}. Possible predictor of PH progression.	RVSP (r=0.32; p=0.01) DL_{CO} (-0.29; p=0.01)	(McMahan et al. 2015)
PlGF	SSc-PAH vs SSc	77	Plasma levels significantly higher in SSc-PAH. Correlates with severity of Raynaud's phenomenon and inversely with DL_{CO}.	DL_{CO} (r=-.031; p=0.01)	(McMahan et al. 2015)
VEGF-A	SSC-PAH vs SSc vs HV	53	Serum levels significantly higher in SSc-PAH than either SSc or HV. Levels correlate with echocardiographic sPAP, dyspnoea score and DL_{CO}.	sPAP (r=0.58; p<0.01) DL_{CO} (r=-0.47; p<0.01)	(Papaioannou et al. 2009)
GDF-15	SSc-PAH vs SSc	54	Plasma levels significantly higher in SSc-PAH, correlate with echocardiographic RVSP and circulating NT-proBNP. Discriminates between PH and non-PH.	RVSP (r=0.56; p<0.001)	(Meadows et al. 2011)
RELM-ß	SSc-PAH vs IPAH vs HV	26	Tissue concentrations significantly higher in SSc-PAH than in IPAH or HV.		(Angelini et al. 2009)
sThrombomodulin	SSc-PAH vs SSc vs HV	92	Significantly higher plasma levels in SSc-PAH compared to either SSc or HV.		(Stratton et al. 2000)

Abbreviations: NT-proBNP - N-terminal pro-brain type natriuretic protein; sFLT-1 - soluble vascular endothelial growth factor receptor 1; PlGF - placenta growth factor; VEGF-A - vascular endothelial growth factor A; GDF-15 - growth differentiation factor-15; RELM-ß - resistin like molecule-ß; sThrombomodulin - soluble thrombomodulin; SSc-PAH - systemic sclerosis related pulmonary arterial hypertension; SSc - systemic sclerosis; IPAH - idiopathic pulmonary arterial hypertension; HV - healthy volunteer; Sens - sensitivity; PH - pulmonary hypertension; Spec - specificity; mPAP – mean pulmonary artery pressure; PVR – pulmonary vascular resistance; CI – cardiac index; RVSP - right ventricular systolic pressure; DL_{CO} - diffusing capacity for carbon monoxide; sPAP - systolic pulmonary artery pressure; EC - endothelial cells.

1.2.3.2 N-terminal pro-brain natriuretic peptide (NT-proBNP)

NT-proBNP is a marker of myocardial stress and therefore a non-specific marker for pulmonary hypertension (PH). Brain type natriuretic peptide (BNP) and NT-proBNP remain the only blood-based biomarkers suggested by guidelines for routine clinical use.(Galie et al., 2016) NT-proBNP is an inactive cleavage product released during the activation of BNP from its prohormone. BNP is released in response to ventricular stretch and stimulates natriuresis and diuresis via the kidney in order to reduce ventricular preload. NT-proBNP is elevated in PH of any cause (Warwick et al., 2008) and correlates with echocardiographic, haemodynamic and functional measurements.(Avouac et al., 2015, Fijalkowska et al., 2006, Leuchte et al., 2004)

NT-proBNP may be elevated in systemic sclerosis in the absence of pulmonary hypertension as a result of left ventricular disease and primary myocardial involvement.(Avouac et al., 2015)

In a prospective observational study of 109 patients with systemic sclerosis, including 68 with PH and 41 without PH at right heart catheter, Williams *et al* set out to evaluate the utility of NT-proBNP concentrations as a screening tool for PAH. NT-proBNP concentration was significantly higher in patients with PAH than without (1474 pg/ml vs 139 pg/ml respectively, p=0.0002). The authors also reported a significant correlation between NT-proBNP concentration and mean pulmonary arterial pressure (mPAP) (r=0.62, p<0.0001), pulmonary vascular resistance (PVR) (r=0.81, p<0.0001) and right atrial pressure (RAP) (r=0.53, p<0.0001) at right heart catheterisation (RHC). For the ability to accurately diagnose PAH a threshold of 395 pg/ml was selected, returning a sensitivity 55.9%, specificity 95.1%, PPV 95.1% and NPV 56.5%. Longitudinal analysis of baseline and change in serial NT-proBNP measurements both demonstrated significant prognostic utility.(Williams et al., 2006) More recent work has provided validation, with Chung *et al.* reporting a sensitivity and specificity of 73% and 78% respectively for NT-proBNP at a threshold of 210 pg/ml, slightly superior to that of BNP at 71% and 59% respectively at a threshold concentration of 64 pg/ml.(Chung et al., 2017)

Comparing between PAH phenotypes in a study of 98 prevalent PAH patients (SSc-PAH n=55; IPAH n=38; Anorexigen n=5), Mathai *et al* found that NT-proBNP levels were significantly higher in the SSc-PAH group vs IPAH group (1846 pg/ml vs 808.5 pg/ml respectively, p<0.01), and this was despite a significantly higher mPAP in the patients with IPAH (41 mmHg vs 48 mmHg, SSc vs IPAH respectively, p<0.01). The authors also noted stronger correlations between NT-proBNP concentrations and haemodynamic measures of PAH for patients with SSc-PAH than for those with IPAH; cardiac index (CI) (r=-0.58, p<0.01 vs r=-0.46, p<0.01 respectively); PVR (r=0.54, p<0.01 vs r=0.41, p<0.01 respectively). When serial protein measurements were analysed in each subgroup, the prognostic value of NT-proBNP for predicting death remained only in the group with SSc-PAH (SSc-PAH: hazard ratio (HR) 3.07, p<0.01; IPAH: HR 2.02, p=0.29).(Mathai et al., 2010)

The DETECT study investigated a population of SSc patients who were enriched for the presence of PAH by the inclusion of patients with a DL_{CO} <60% predicted.(Coghlan et al., 2014) NT-proBNP was included in a final 2-step algorithm which also included electrocardiography to select patients to proceed to RHC. Sensitivity for the detection of PAH was high (96%) but specificity was only 48%.

Both BNP and NT-proBNP levels have been demonstrated to be important prognostic predictors at baseline in PAH.(Andreassen et al., 2006, Nagaya et al., 2000) Subsequently, the change in NT-proBNP level after therapy was shown to be a powerful independent predictor of survival.(Nickel et al., 2012) More recently three large studies have confirmed the importance of changes in NT-proBNP in the risk stratification of patients with PAH during follow-up.(Boucly et al., 2017, Hoeper et al., 2017, Kylhammar et al., 2018)

1.2.3.3 Endoglin

Transforming growth factor beta (TGF-ß) signalling has been strongly implicated in the pathogenesis of PAH, and extensively studied, particularly with regard to bone morphogenetic protein receptor type-2 mutations.(Machado et al., 2015) TGF-ß signalling regulates several processes including cellular proliferation and angiogenesis. Endoglin (Eng) is a transmembrane protein expressed in endothelial cells which acts as a TGF-ß signalling complex component.(Conley et al., 2000)

Both TGF-ß serum concentration and Eng level are raised in IPAH patients, with Eng localised to endothelial cells in tissue samples.(Gore et al., 2014) Germline Eng mutation have shown a protective effect against the development of pulmonary hypertension in heterozygous models exposed to chronic hypoxia.(Gore et al., 2014)

Coral-Alvarado *et al* investigated circulating Eng concentration in 60 patients (20 SSc-PAH; 20 SSc-no PAH; 20 HV). PH was diagnosed by estimation of systolic pulmonary artery pressure >35mmHg, or tricuspid regurgitant jet velocity >3m/s. The authors report higher Eng concentrations in the SSc-PAH group, however possibly due to small study numbers, the difference is only statistically significant between SSc-PAH vs healthy volunteer (HV) groups (SSc-PAH: 6.89 ng/ml, SSc-no PAH: 6.2 ng/ml, HV: 5.42 ng/ml; SSc-PAH vs SSc-no PAH p=0.2447, SSc-PAH vs HV p=0.0006, SSc-no PAH vs HV p=0.057). There was no correlation noted between Eng concentration and echocardiographic measurements of PH.(Coral-Alvarado et al., 2010)

There is some evidence for altered Eng expression in PAH, however in SSc specifically, this evidence is weak in part due to small study sizes and study design. Given the potential role of Eng in TGF-ß signalling, a role in the pathogenesis of PAH remains reasonable, however further work in this area is needed to establish its role.

1.2.3.4 VEGF-A

Vascular Endothelial Growth Factor-A (VEGF-A) is a member of the PDGF superfamily of growth factors. It is one of the most potent regulators of angiogenesis, and acts on vascular endothelial cells through stimulation of KDR (VEGF receptor 2) and FLT-1 (VEGF receptor 1) to promote angiogenesis, increase vascular permeability and stimulate endothelial cell migration.(Shibuya, 2011, Voelkel et al., 2014)

Serum VEGF-A concentrations are known to be elevated in patients with PAH and have been demonstrated within plexiform lesions of remodelled vasculature.(Eddahibi et al., 2000, Papaioannou et al., 2009)

In a study including 53 participants (SSc-PAH n=20, SSc-no PAH n=20, HV n=13) Papaioannou *et al* examined the relationship of serum VEGF-A concentration to echocardiographic markers of pulmonary hypertension. In this study, participants were treatment naive, and any patients with pulmonary fibrosis were excluded. Estimated sPAP >35mmHg was used to define patients with SSc-PAH. The authors found significantly higher VEGF-A concentrations in all patients with SSc as compared to HV (267 pg/ml vs 192 pg/ml respectively, $p<0.01$), and further found that those with SSc-PAH had higher levels than those with SSc-no PAH (352 pg/ml vs 240 pg/ml respectively, $p<0.01$). In patients with SSc, significant correlations were found between serum VEGF-A concentration and systolic pulmonary arterial pressure (sPAP) ($r=0.58$, $p<0.01$); MRC dyspnoea score ($r=0.34$, $p=0.031$); and DLco ($r=-0.47$, $p<0.01$). In multivariable modelling of sPAP as the dependent variable, VEGF-A concentration remained a significant predictor when adjusted for age and gender.(Papaioannou et al., 2009)

VEGF-A expression is known to be upregulated in both patients PAH, and with systemic sclerosis, both conditions characterised by pathologically excessive endothelial activation. In patients with SSc-PAH the VEGF pathway is upregulated, however baseline levels have not been assessed for utility as diagnostic biomarkers.

1.2.3.5 Placenta Growth Factor (PlGF) and Soluble vascular endothelial growth factor (VEGF) receptor 1 (sFLT-1)

Placenta growth factor is a member of the vascular endothelial growth factor family of proteins which binds with high affinity for VEGF receptor 1 (FLT-1/ VEGF-R1), but not for VEGF receptor 2 (KDR/ VEGFR2) - regarded as the main effector protein of VEGF signalling.(Park et al., 1994) PlGF alone does not stimulate tyrosine kinase phosphorylation or proliferation in human endothelial cell lines, however the addition of PlGF potentiates the effect of VEGF-A in stimulating proliferation of cultured endothelial cells.(Park et al., 1994) sFLT-1 is a variant of VEGF receptor 1 (FLT-1) which can bind VEGF-A, VEGF-B, and Placenta growth factor (PlGF). It functions as a decoy receptor, downregulating free ligand and therefore thought to control excessive endothelial activity.(Olsson et al., 2006)

Recognising the need for further study into diagnostic biomarkers for patients with SSc-PAH, McMahan *et al* designed a case-control study of 77 patients with SSc (37 with PH, 40 without PH). The groups were unbalanced for age (64.9 vs 55.9 respectively, p<0.01) and lung volumes (FVC% 67.5 vs 88.1 respectively, p<0.01). Diagnosis of PH was based on mPAP >= 25 mmHg at right heart catheterisation. The authors report that both PlGF (24.8 pg/ml vs 19.1 pg/ml, p=0.02) and sFLT-1 (101.8 pg/ml vs 89.7 pg/ml, p=0.02) are significantly upregulated in patients with PH than in those without. Both proteins were significantly inversely correlated to DLco (PlGF: r=-0.31, p=0.01 and sFLT-1: r=-0.29, p=0.01). sFLT-1 was also correlated to RVSP (r=0.32, p=0.01).(McMahan et al., 2015)

This study was designed to evaluate potential biomarkers of pulmonary hypertension in systemic sclerosis. No comment is made about extent of pulmonary fibrosis, so it remains conceivable that these are imbalanced given the difference in baseline pulmonary function tests, and it is not clear why no comparisons were given for protein concentration and invasive right heart catheter measurements. Although protein concentration changes are noted, no statistics have been given for the performance of these proteins as diagnostic markers.

1.2.3.6 GDF-15

Growth Differentiation Factor-15 (GDF-15) is a member of the TGF-ß superfamily of cytokines playing an important role in cell growth and differentiation. It is a stress responsive cytokine associated with tissue damage and inflammation. Increased levels have been reported in heart failure, atherosclerosis, endothelial dysfunction and diabetes and have been linked to disease progression and prognosis.(Adela et al., 2015)

In treatment naïve IPAH, serum GDF-15 is increased and is a significant predictor of survival.(Nickel et al., 2008) In a mixed cohort of PAH patients, tissue levels of GDF-15 are increased - localising to the pulmonary endothelium, and in remodelled vessels strong signals are identified in plexiform lesions.(Nickel et al., 2011) In-vitro studies using pulmonary endothelial cells and varying concentrations of GDF-15 resulted in reduction in hypoxia induced apoptosis suggesting a potential pathological mechanism in PAH.(Nickel et al., 2011)

Meadows *et al* studied a cohort of 111 patients (SSc-PAH n=30, SSc-no PAH n=24, IPAH n=44, HV n=13) for circulating GDF-15 concentrations. PH was defined at right heart catheterisation. Patients with PAH were already established on PH specific therapy at the time of entry to study. Both plasma and tissue levels of GDF-15 were elevated in SSc-PAH (442 pg/ml), and differentiated it from SSc without PAH (108 pg/ml, p=0.0004), IPAH (173 pg/ml, p=0.0003) and HV (66 pg/ml, p=0.0013). Within the SSc subgroup, GDF-15 levels correlated with echocardiographic RVSP (r=0.556, p<0.001), and with NT-proBNP concentration (r=0.484, p<0.001), but not with other invasive haemodynamics. On diagnostic ROC analysis, GDF-15 has been shown to have good discriminative power with area under curve (AUC) 0.91 for differentiation of SSc-PAH from SSc without PH with an optimal threshold for GDF-15 of 125 pg/ml demonstrating 93% sensitivity and 88% specificity for the presence of SSc-PAH. Furthermore, patients below this threshold were found to have significantly improved survival.(Meadows et al., 2011)

1.2.3.7 <u>Resistin-like molecule-ß (RELM-ß)</u>

RELM-ß is a member of a relatively newly described resistin family. Largely studied through their effects on animal models, these proteins have been shown to induce angiogenesis and vascular remodelling.(Angelini et al., 2009)

Following the identification that hypoxia induced mitogenic factor (HIMF) is upregulated in animal models of PH, Angelini et al sought to evaluate this in human tissues. In a small study involving 26 prevalent patients (SSc-PAH n=9, IPAH n= 11, HV n=6), the authors found that in human lung tissue samples, RELM-ß (a close human homolog to HIMF) is upregulated in patients with SSc-PAH as compared to healthy control (p<0.01, measured by relative intensity on western blot) and localises to remodelled vasculature. In comparison, although some expression of RELM-ß was noted in remodelled vessels of patients with IPAH, this was inconsistent, and relative quantification showed no difference between IPAH and HV concentrations. Additional in-vitro study showed mitogenic activity of RELM-ß on both human lung microvascular endothelial cells and human pulmonary artery smooth muscle cells.(Angelini et al., 2009)

This is a relatively novel candidate protein, which appears to show higher expression in SSc-PAH, however more work is needed to assess its concentration in the circulating compartment if it is to be considered further as a biomarker as lung tissue samples are not practical for this purpose.

1.2.3.8 Soluble thrombomodulin (sThrombomodulin)

Thrombomodulin is a glycoprotein expressed on endothelial cells. Its physiological function is to bind thrombin and alter its activity, to subsequently activate protein C.(Stratton et al., 2000) The pathogenesis of both systemic sclerosis and pulmonary arterial hypertension involves and injury to and activation of the vascular endothelium. Soluble thrombomodulin is increased in conditions associated with endothelial damage.(Mercie et al., 1997)

Stratton *et al* studied 92 patients (SSc-PAH n=34, SSc-no PAH n=38, HV n=20) and found that sThrombomodulin was increased plasma of patients with SSc-PAH (65.4 ng/ml) compared to SSc without PH (43.3 ng/ml, p<0.05), and healthy controls (38.1 ng/ml, p<0.05). There was no difference in circulating concentration between SSc without PH and healthy control.(Stratton et al., 2000) This is in contrast to previous studies which have shown a significant decrease in circulating sThrombomodulin concentration in patients with PAH (IPAH and PAH due to Eisenmenger's' syndrome) compared to healthy controls (26 ng/ml vs 44 ng/ml respectively, p=0.0001).(Cacoub et al., 1996)

1.2.3.9 Summary of published data

A biomarker has been defined by the NIH as “a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.” (Strimbu et al., 2010) NT-proBNP is the most widely studied circulating biomarker in clinical use in patients with suspected or known PAH. Elevations in NT-proBNP result from right ventricular (RV) strain as a result of increased RV afterload. As it does not reflect the underlying pathophysiology of the pulmonary arterial vasculopathy resulting in increased RV afterload in PAH, NT-proBNP levels can be elevated due to other pathophysiological processes including increased RV afterload due to PH arising from left heart disease and from disease processes directly affecting the myocardium. As such, the specificity of NT-proBNP in the diagnosis of SSc-PAH tends to be rather low resulting in a significant number of RHCs being performed in patients who do not in the end have PAH.(Coghlan et al., 2014) Furthermore, given the dismal prognosis in SSc-PAH, identifying patients early in their disease process before the development of RV strain is desirable.(Coghlan et al., 2018, Condliffe et al., 2018) The identification of a biomarker or panel of biomarkers which more reflect the underlying pulmonary vasculopathy in SSc-PAH prior to the development of RV strain is therefore of interest.

The data described summarise the current evidence for various candidate circulating diagnostic biomarkers for SSc-PAH, several of which do relate to pathways known to be important in PAH pathogenesis, especially the TGF-ß and VEGF pathways. Further study within well phenotyped cohorts of patients to compare the performance of these candidate circulating biomarkers against NT-proBNP and the DETECT protocol are clearly warranted.

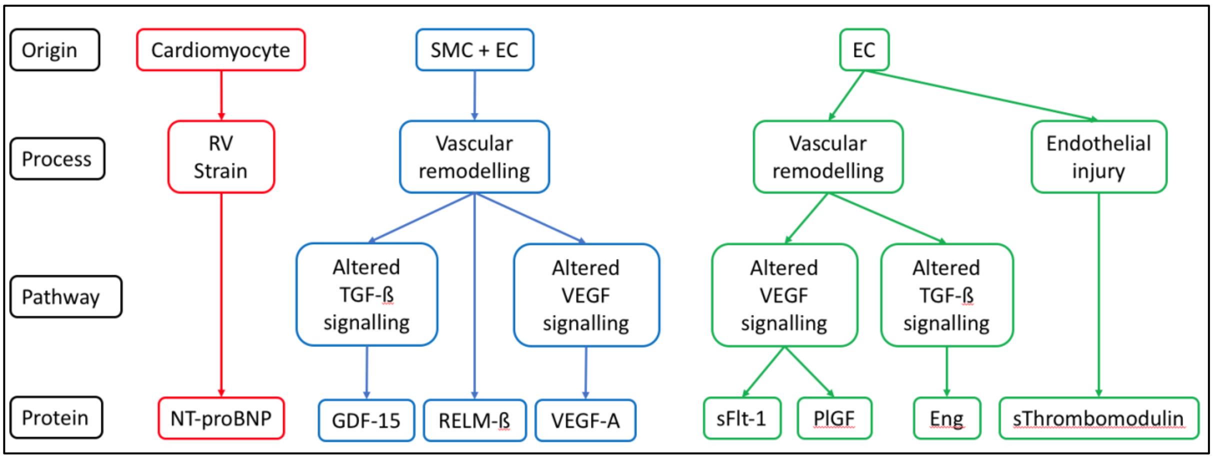


Figure 1.5: Cellular origin and pathways for each protein described in the context of SSc-PAH

Description of the likely origin of each protein, along with the pathophysiological process it has a role in. If a component of one of the pathways known to be relevant to pathogenesis of PAH then this is also given.

Abbreviations: SMC - vascular smooth muscle cell; EC - vascular endothelial cell; RV - right ventricle; TGF-ß – transforming growth factor beta; VEGF – vascular endothelial growth factor; NT-proBNP – N-terminal pro-brain natriuretic peptide; GDF-15 – growth differentiation factor-15; RELM-ß – resistin-like molecule beta; VEGF-A – vascular endothelial growth factor A; sFLT-1 – soluble vascular endothelial growth factor receptor 1; PlGF – placenta growth factor; Eng – endoglin; sThrombomodulin – soluble thrombomodulin.

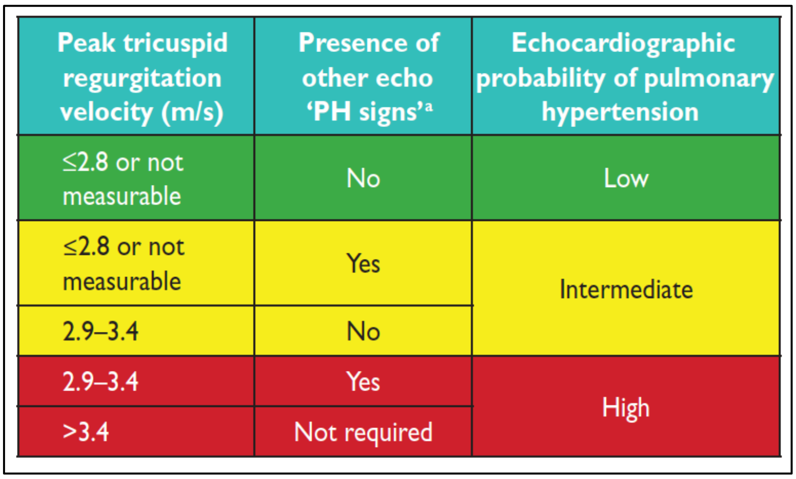
1.2.4 Screening strategies

Although a rare condition overall, the prevalence of PAH in patients with SSc is around 10%.(Avouac et al., 2010) Accordingly it has been standard practice to screen patients with SSc for the presence of PAH.

In 2011, Humbert et al published evidence from a small trial which suggested improved outcomes for patients with SSc-PAH who were diagnosed earlier as a result of screening, rather than waiting for development of clinical symptoms and routine clinical investigation.(Humbert et al., 2011) To adjust for potential lead time bias, the authors subdivided the patients by functional status at diagnosis and the survival difference was maintained.(Humbert et al., 2011)

1.2.4.1 ERS guidelines (2015)

Acknowledging the incidence of PH in certain high risk groups including SSc, the ERS guidelines suggest screening for PH in asymptomatic individuals from these groups.(Galie et al., 2016) For patients with SSc it is suggested that alongside annual ECG and pulmonary function testing, that these patients should undergo annual transthoracic echocardiographic examination as part of screening. It is noted that echocardiography alone is not sufficient to guide treatment decisions, and is therefore used to stratify patients into risk groups, with high risk patients recommended to undergo invasive diagnostic testing with right heart catheterisation (Figure 1.6).



Peak tricuspid regurgitation velocity (m/s)	Presence of other echo 'PH signs'[a]	Echocardiographic probability of pulmonary hypertension
≤2.8 or not measurable	No	Low
≤2.8 or not measurable	Yes	Intermediate
2.9–3.4	No	
2.9–3.4	Yes	High
>3.4	Not required	

Figure 1.6: Echocardiographic probability of pulmonary hypertension in asymptomatic patients with a suspicion of pulmonary hypertension (Galie et al., 2016)[3]

1.2.4.2 DETECT study

In an effort to standardise the approach to screening for PAH more specifically to patients with SSc, the DETECT protocol was developed and includes several candidate parameters including demographic data, physical examination characteristics, clinical history, pulmonary function tests, blood testing, ECG and echocardiographic measurements.(Coghlan et al., 2014) Through multiple levels of variable selection and expert panel consensus, the initial 112 variables were reduced to 8 variables, and stratified into final a two level test; baseline characteristics with blood tests; and echocardiographic parameters (Figure 1.7).(Coghlan et al., 2014) To minimize false negatives, the algorithm was designed to maximize the negative predictive value.

[3] Reproduced from European Heart Journal 37(1): 67-119, with permission from the Oxford University Press.

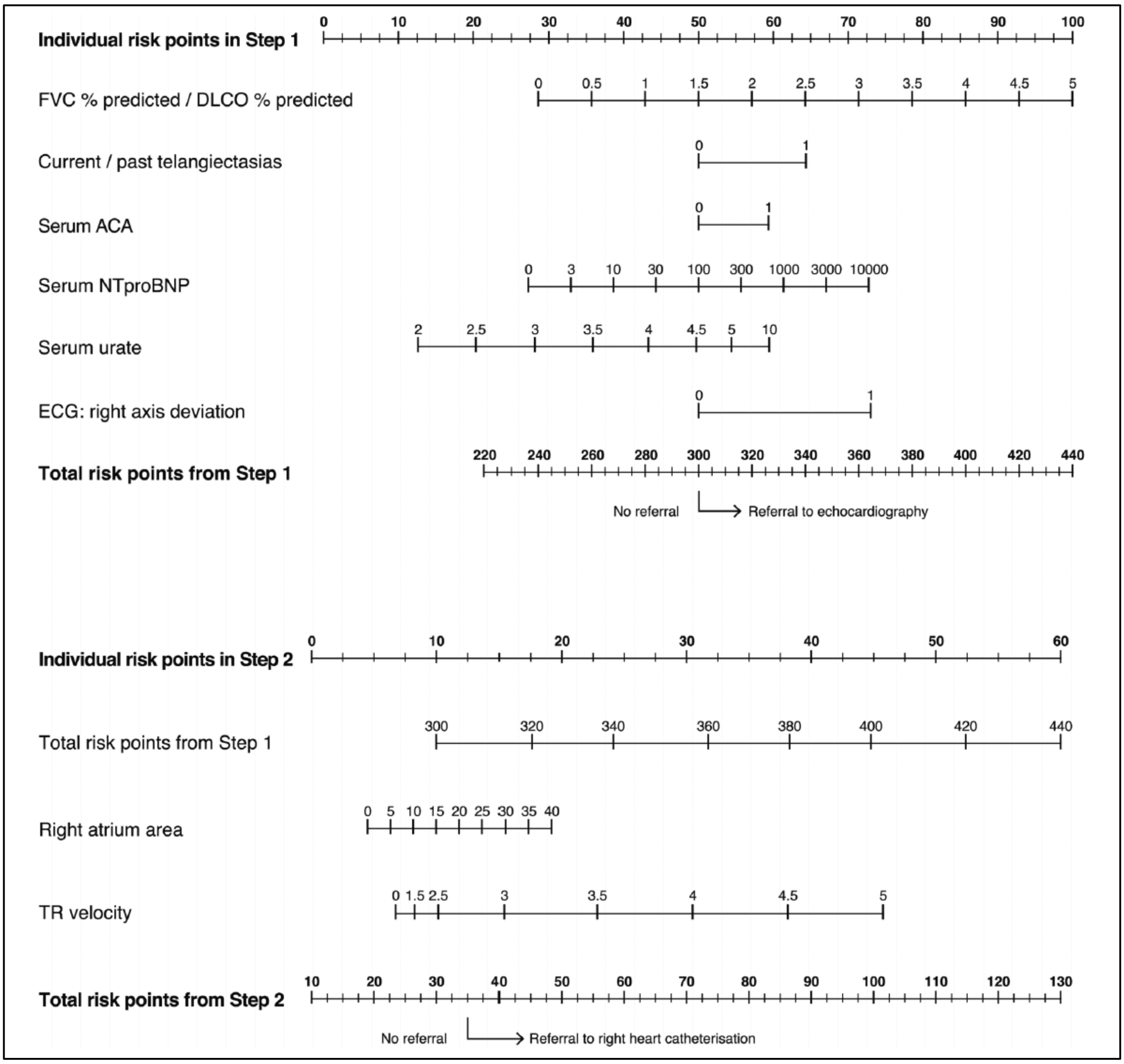


Figure 1.7: DETECT algorithm nomogram for determination of the likelihood of pulmonary hypertension and cut-off for the recommendation of referral for invasive right heart catheterization (Coghlan et al., 2014).[4]

Developed for screening use in rheumatology, the algorithm is suggested for application to patients with at least a 3-year history of diagnosed SSc, with DL_{CO} <60% predicted, and without a previously known diagnosis of PH.(Galie et al., 2016) The final DETECT score has a negative predictive value of 0.98, with 4% false negatives, but a positive predictive value of just 0.35, and a false positive rate of 65%.(Coghlan et al., 2014)

[4] Reproduced from Annals of the Rheumatic Diseases 73(7): 1340-9 with permission from the BMJ Publishing Group Ltd.

Although the DETECT protocol has been proposed as the optimal screening tool for PAH in SSc patients, the test statistics suggest it is not an ideal tool since given a high false positive rate, a large number of patients are referred for potentially inappropriate invasive testing with right heart catheterisation. Further limitations lie both in the ability of the clinician to recognise necessary clinical signs leading to a potential source of error in the calculation; and the availability of echocardiography.

1.2.4.3 Comparison of screening strategies

In 2015, Hao and Thakkar et al published a comparison between results of both the ERS and DETECT screening algorithms on patients with SSc recruited from the Australian Scleroderma Cohort Study. For comparison, only patients with group 1 PH and controls were retained in the final analysis. The DETECT screening algorithm identified PH in this group with sensitivity 100%; specificity 35.3%; PPV 55.1% and NPV 100%, compared to ERS screening with sensitivity 96.3%; specificity 32.3%; PPV 55.3% and NPV 90.9%.(Hao et al., 2015) A strong sensitivity and NPV were the metrics considered most important, and on this basis the DETECT algorithm outperformed the ERS guidelines.

A similar analysis in a separate cohort of Belgian patients, designed to look more at the economics of each protocol highlighted the very poor PPV for the DETECT algorithm at 6%, compared to 11% for the ERS echocardiography guidelines. For the 3 patients found to have PAH, both algorithms correctly included them. Further analysis identified a greatly increased cost associated with DETECT screening compared to ERS methods.(Vandecasteele et al., 2017)

These, amongst other comparison studies, identify the strong sensitivity and negative predictive value for the DETECT protocol in a highly selected group of patients with SSc. It is also clear that the DETECT protocol has a poor positive predictive value and therefore excessive patients are screen positive and suggested for right heart catheterisation via this method. The DETECT protocol is complex, relying on multi-modality testing, and therefore can be expensive and result in a delay in referral for definitive testing.

1.3 Summary - what is needed, and why?

It is unethical to subject patients unnecessarily to tests which carry significant risks. Although at present right heart catheterisation is carried out in most patients referred with a significant DETECT score, this is done on the basis of the risk of complication in those with false positives being outweighed by the improved outcome for those diagnosed with and treated for pulmonary arterial hypertension at an earlier stage.

An ideal test would maintain the negative predictive value of the DETECT study, whilst improving on specificity, and reducing potential sources of error. A diagnostic panel formed of circulating protein markers, easily accessible through a single blood test, without the need for clinician detected signs, or the availability of operator dependent imaging techniques could theoretically provide a test with improved diagnostic potential.

There is a need for improved markers with potential for diagnosis, severity stratification, prognostication and response to treatment so that we can better counsel patients regarding the likely disease process tailored to them as an individual, and so that we can better streamline patients to the most effective treatment strategy.

1.4 Hypothesis

There are there circulating proteins, or a panel of combined circulating proteins for which the expression significantly differs between patients with SSc with or without PAH, and these can be used to reliably detect PAH in a cohort of SSc patients.

1.4.1 Related research questions

1. Can such a model, which detects a signal related to PAH pathobiology, be extended reliably to function for other PAH subtypes?

2. Can such a model be extended from classification of disease to demonstrate prognostic utility?

3. Are there novel proteins identified which play a role in the pathogenesis of pulmonary hypertension? If so:
    a. Are they expressed in lung tissue?
    b. What effect do they exert on cell types relevant to pulmonary arterial hypertension?

1.4.2 Specific thesis objectives

In the work described in this thesis we aimed to examine the above aims as follows:

1. To address question 1, I used my final predictive model (paragraph 4.5.4), derived to predict disease classification between patients with SSc-PAH and SSc-no PAH, with a separate cohort of patients from the initial Myriad DiscoveryMAP cohort (Table 3.1) to determine whether the panel can accurately classify patients with IPAH from healthy volunteers. This can be found in paragraph 5.3.

2. To address question 2, I tested my model score against known patient outcome to determine whether the panel can go beyond disease classification, and predict outcome based on the magnitude of the model score. This work can be found in paragraph 4.5.5.2.

3. To address question 3a, I used immunohistochemistry to identify the presence of proteins in my final model, not previously described in relation to pulmonary hypertension, in lung tissue sections from both human transplant explants, and animal models of pulmonary hypertension. This work is presented in paragraph 6.3.1.

4. To address question 3b, I used human pulmonary arterial endothelial- and smooth muscle- cells in cell culture models to investigate any direct effect of these proteins on these cell lines. This work can be found in paragraph 6.3.2.

2 Methods

2.1 Ethics Statement

Clinical data and blood samples were obtained through license of materials from the Sheffield Teaching Hospitals Observational Study of Patients with Pulmonary Hypertension, Cardiovascular and Lung Disease (STH-ObS) Biobank collection, project registration STH15222, under full NHS Research Ethics Committee approval (Ethics approval No. 08/H1308/193+5).

2.2 Patient Data

Patients samples and data were obtained retrospectively from patients enrolled in STH-Obs. After agreement with Myriad RBM and Actelion regarding study groups of interest. Patients were recruited into 7 study groups:

1. Systemic sclerosis with pulmonary hypertension (SSc-PAH) (Group 1)
2. Systemic sclerosis with pulmonary hypertension and interstitial lung disease (SSc-PAH-ILD) (Group 1)
3. Pulmonary arterial hypertension with another connective tissue disease (PAH-Other CTD) (Group 1)
4. Pulmonary hypertension related to interstitial lung disease (PH-ILD) (Group 3)
5. Systemic sclerosis without pulmonary hypertension (SSc-no PH)
6. Idiopathic pulmonary arterial hypertension (IPAH) (Group 1)
7. Healthy volunteers (HV)

Patients recruited to the STH-ObS are carefully phenotyped at initial clinical contact by consultants in pulmonary vascular diseases, with multidisciplinary review of clinical presentation, radiology and diagnostic investigation prior to final disease classification. All patient notes, radiology and diagnostic investigations were further reviewed to ensure accurate disease classification, and extensive searches undertaken to retrieve missing data from local systems, research databases and patient notes to generate a comprehensive phenotype database. Baseline phenotype data collected included:

1. Demographics – gender, ethnicity, dates of birth and death
2. Visit/Sampling date
3. WHO functional class
4. Medical co-morbidities
5. Specific PH diagnostic classification

6. Right heart catheter data – Date, Cardiac output, RA mean, Cardiac index, mPAP, PCWP, PVR
7. Pulmonary function tests – FEV1 (%), FVC (%), TL_{CO} (%)
8. Incremental shuttle walking test – Distance
9. Current PH specific medication
10. Haematological indices – FBC, UE, LFT, CRP

Additional clinical information was sought to complete the DETECT criteria for comparison between screening methods. This required further review of patient notes for all patients with SSc. The following information was gathered for all patients where possible at the timepoint corresponding to the baseline sample. DETECT criteria data collected included:

1. Presence or absence of telangiectasia
2. Anti-centromere antibody status
3. Serum NTproBNP concentration
4. Serum urate concentration
5. Presence or absence of right axis deviation on ECG
6. Echo: Right ventricular systolic pressure and right atrial pressure.

A measure of right atrial area from echocardiography is a requirement for completion of the DETECT score. This metric was not routinely recorded in our hospitals during the period our patients were recruited to the biorepository. Maximal right atrial area was retrospectively measured in most cases from ECG gated cardiac MRI scans, and in a minority from non-ECG gated CTPA using PACs software, from scans available from the timepoint matched to each sample.

Follow up samples were identified for a subset of patients with SSc for use in longitudinal analyses. These were identified as paired samples to our baseline cohort where available and the matching clinical information gathered through the same search process for the time point corresponding to that sample.

2.3 Commercial Derivation Laboratory Studies

2.3.1.1 Consent, Venesection and Sample storage in the Sheffield Pulmonary Vascular Biorepository

Patients were identified and approached for inclusion in the Sheffield biorepository at first visit, while treatment naïve, by consultants in the Sheffield Pulmonary Vascular Disease unit.

Patients were then consented for inclusion in the biorepository (STH15222) by trained research nurses.

Serum samples were obtained by our research nurses following standard operating protocol (STH16436 SOP05). Blood samples were taken by peripheral venepuncture or at diagnostic RHC into gold top serum vacutainer tubes. Tubes were then mixed thoroughly before being stood upright for 30 minutes at room temperature. Samples were then centrifuged at 1500G for 15 minutes at room temperature. Serum samples were decanted by disposable pipette into cryovials and carefully labelled for storage. Samples were then frozen and stored in the liquid nitrogen biorepository until required.

2.3.1.2 Myriad RBM Luminex Assays

Myriad RBM have extensive experience and a good track record in clinical assay development. The DiscoveryMAP platform from Myriad RBM reported protein concentration data for 296 proteins from our serum samples. These analytes are not biased to any particular disease process and allow for study of protein concentrations across a wide range of physiological and pathological processes. In contrast to other more extensive platforms such as SOMAscan, Myriad RBM develops its own in-house assays and does not rely on commercial kits, they therefore have better control over variability. Myriad RBM uses sandwich capture immunoassays that use two specific antibodies to bind the target, the SOMAscan utilises an ionic binding aptamer molecule which is less stable than an antibody method.(Christiansson et al., 2014) Myriad RBM uses a calibration/control method, and is validated for analyte specificity and cross-reactivity whereas SOMAscan is not. The Myriad RBM platform has been validated to Clinical Laboratory Standards Institute guidelines. The platform provided by Myriad RBM has a track record over the past 18 years and data generated has been used in over 1000 peer-reviewed publications. It is for these reasons that we felt that the Myriad RBM platform would provide the most robust and reproducible data for our analysis.

The Myriad platform is a broad and non-biased protein discovery assay platform, with protein targets covering a wide spectrum of human physiology and pathophysiological pathways. Current understanding of the pathophysiology of early SSc includes endothelial dysfunction as a key early pathway to the development of the disease phenotype(Altorok et al., 2014) , a feature shared with the understanding of the early pathophysiology of PAH.(Budhiraja et al., 2004) Although limited in fine detail due to the low number of proteins measured relative to other assay platforms, the Myriad Discovery assay platform has good coverage of proteins relevant to endothelial dysfunction documented in both systemic sclerosis and pulmonary arterial hypertension.(Odler et al., 2018) Markers of endothelial dysfunction measured on the Myriad DiscoveryMAP platform include von Willebrand Factor, Endostatin, Endoglin and Platelet endothelial cellular adhesion molecule-1. Notable absentees with well documented association with endothelial dysfunction in SSc-PAH include Assymetric dimethylarginine (ADMA) and Endothelin-1.

1 ml serum samples were analysed externally by Myriad RBM under contractual agreement between UoS, Myriad RBM and Actelion pharmaceutical, and results shared between all teams. Assays were performed on a validated fully-automated luminex based platform.

Appropriate multiplexes are determined based on serum concentration of each analyte, and therefore the requirement for similar dilution prior to assay can be controlled.

Myriad RBM has optimized a method for controlling lot-to-lot variation using reference samples. These calibrators are analysed with each new lot of reagents and results adjusted accordingly to control for batch variation.

Serum samples were shipped to Myriad RBM by DHL courier on dry ice and confirmed by Myriad RBM to have been received appropriately packaged, undamaged and in frozen condition upon receipt. Upon receipt, each sample is allocated a unique identifier that tracks it throughout the automated analysis process. Samples were randomized prior to loading on the analysis plate. For plating, samples are removed from the freezer, thawed, vortexed and centrifuged before being pipetted into a 96 well microtiter plate. Two technicians work together to plate the samples manually according to the plate map, with the second present to verify the correct sample placement.

Automated liquid handling instruments are used for all dilutions, reagent additions, and manipulation. A small volume from each of the sample wells is added to a reaction well containing capture beads. These microspheres are conjugated to antibodies and encoded with a unique fluorescent signature specific to the analyte of interest. The beads are allowed time to incubate with the sample before other reagents are added. Biotinylated detection reagents are then added, followed by a fluorescent reporter molecule. The multiplex is then washed to remove any unbound detection reagents prior to reading on the luminex machine.

The luminex technology uses the principle of hydrodynamic focussing to pass the microspheres, one at a time, along a path that is interrogated by two lasers. The excitation beams measure the unique fluorescent signature of each bead, with the amount of fluorescence generated proportional to the analyte concentration in the sample.

All proteins in the dataset are referred to by their corresponding gene names to allow for communication between complementary studies in our research group. (Appendix 1 – Protein decode)

2.4 Statistics and Data Analysis

2.4.1 Computing specification and Software

Data were analysed using MacBook Pro, 2.9 GHz Intel Core i5 processor. For particularly high demand processing tasks a Mac Pro, 3.5 GHz Intel Xeon E5 processor with 12 cores was used.

Data analysis were conducted using R for Mac version 3.5.1, with appropriate statistical packages as separately documented.

2.4.2 Data Preparation

2.4.2.1 Limits of Detection and Missing Data

Several methods of dealing with variable measurements falling outside the limits of detection were considered. Models were developed both replacing the values outside a limit of detection with the absolute value of the corresponding limit of detection, or removal of the value from the dataset. Both methods resulted in identical variable selection, as those with a high proportion of datapoints beyond the limits of detection were not selected by the statistical models, likely due to a lower information content across affected proteins. It was considered that a datapoint falling outside a limit of detection still yields important information regarding that variable. As such for the purposes of retaining the maximum information in the dataset, and dataset accuracy, each datapoint falling outside the limit of detection was replaced with the absolute value of the corresponding limit of detection.

Replacing data values outside the limit of detection with the absolute value of the limit of detection can lead to an absence of informative data given by that variable if all, or a majority of the datapoints fall at that level. To overcome this, I individually examined all protein variables which had >90% of all datapoints recorded at a single value to assess for any informative data. These variables were individually assessed for any relationship to the classification outcome variable. No variable which met this criterion was found to hold any informative data and they were therefore considered to have been analysed and filtered from entry into further statistical analysis on this basis.

In the protein concentration dataset, 45 of 49235 datapoints were missing due to insufficient sample at analysis.(Figure 2.1) These datapoints were therefore deemed to be missing at random from a statistical perspective and were imputed using ‘MissForest’ imputation package for R. MissForest is method for imputation of missing data based on a random forest algorithm. It is therefore appropriate for use with non-parametric data. It functions by creating a random forest model for each variable, and uses the model to predict the missing values with the help of the observed values.

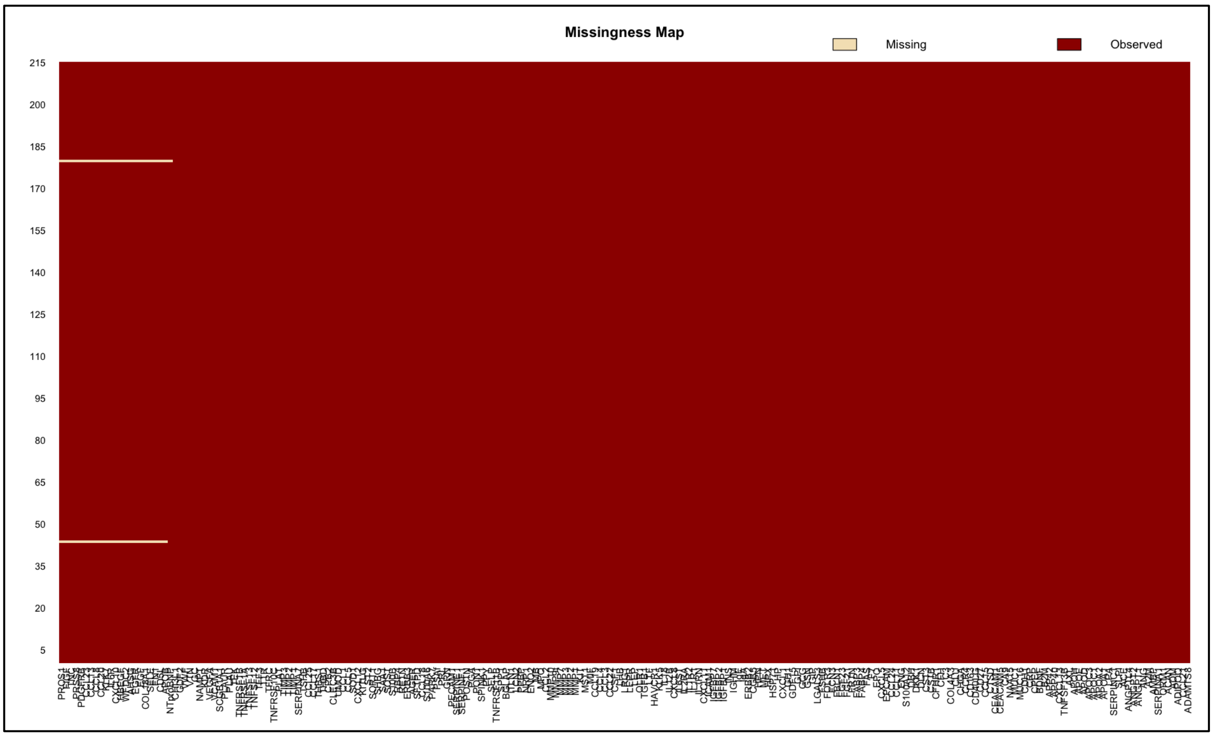


Figure 2.1: Missingness map

Heat map showing the distribution of missing data among the protein variable dataset.

2.4.2.2 Dataset reduction by univariate AU-ROC

To reduce the likelihood of overfitting within classification, univariate proteins were excluded from further analysis in the subset of interest if their area under the ROC curve applicable to the analysis of the particular subset was below a defined cut-off, as this would suggest poor univariate classification potential.

2.4.2.3 Multicollinearity

Multicollinearity was assessed visually using correlation matrix and statistically using Spearman's rho coefficient across all proteins. The effect of collinear interactions was further investigated using variance inflation factors with variables above threshold identified and correlating proteins given individual consideration for removal from further analysis based on known biological relevance.

2.4.2.4 Scaling

Where necessary data were centred and scaled by subtracting the mean to centre at 0, and scaled by division of standard deviation.

2.4.2.5 Normalisation of data

Not all data could be completely normalised, however applying a $\log_2$ transformation yielded the most normalised protein dataset, and this was applied prior to statistical testing where necessary. Resulting variables were subsequently analysed for normality using Shapiro-Wilks test and the appropriate statistical approaches adopted.

2.4.3 Exploratory data analysis: Principle component analysis

Principle component analysis (PCA) is a complex process in statistical mathematics used to display similarities and differences between samples through the analysis and reconfiguration of the dependent variables available to describe them. PCA reduces the dimensionality of a dataset by excluding redundant data present through covariance to create a new set of dimensions each of which represents a combination of the original variables which can then be displayed to demonstrate the similarity and differences between each sample.(Abdi et al., 2010)

2.4.4 Variable selection and classification modelling

Three methods were evaluated for variable reduction and classification modelling; combination panelling based on univariate diagnostic statistics, random forest modelling, and LASSO modelling.

2.4.4.1 Combination panel modelling

Diagnostic panels were created using recursive stepwise combination modelling. For each protein analysed, the corresponding Youden index from ROC modelling was used as a binary cutpoint. To develop the panel, protein values falling above the Youden index were assigned a value 1, and those below a value 0 for that particular protein concentration for each individual patient sample. Every possible combination of between 2 and 5 proteins were modelled, with the additive total for each protein in that combination recorded for each patient. The optimal cutpoint for classification was then derived based on the known patient class, and patients were then assigned a predicted class as either PH or not PH based on panel score. Diagnostic statistics for each combination were calculated from the known and predicted classes and each combination ranked according to diagnostic accuracy.

2.4.4.1.1 Combination Panels: Advantages

This is the most open book approach, built manually and with clear understanding of the relatively simple statistics used to generate each panel, and the outputs generated. Any particular combination of proteins of interest can be reviewed.

2.4.4.1.2 Combination Panels: Disadvantages

This has proven a very time consuming approach, with any iteration, or change in the class of interest requiring a full analysis to be executed at a run time of up to several days to completion in some instances. Panels generated are based on the predictive ability of the individual variables included, combining but not taking into account the influence of the other proteins in the panel, therefore are less likely to generate the optimal results possible from a multivariable statistical model.

The process of combination modelling is based on the univariate statistics and distribution of each included protein in the panel. Given that our derivation cohort is small, the likelihood of error in the distribution is relatively high for each protein, and when proteins are combined, this error is likely to be substantially increased. The chance of overfitting through this process is therefore particularly high.

2.4.4.2 Random Forest

Tree based algorithms are increasingly popular for analysis of large datasets, particularly for categorical dependent variables such as for classification. They can work on both categorical and continuous independent variables, are not particularly influenced by outliers and work well with non-parametric data.

Random forest modelling is an ensemble method designed to increase accuracy in prediction through averaging of large numbers of decision trees (bagging).(Figure 2.2) To derive each decision tree, a random sample of patients are chosen from whom the protein variable data are used to train the tree. At each node a random selection of the protein variables are examined with the predictor yielding the largest information gain used to split that particular node to develop the tree. This continues until the chosen number of splits has been reached. Each leaf is designated the independent variable value corresponding to the mode average of those falling within it. The tree is then tested for error in prediction using the values of the remaining data (the non-training data). Decision trees are generally felt to be easily understood without significant background statistical knowledge, however run a high risk of overfitting a model to a derivation dataset without appropriate constraints.

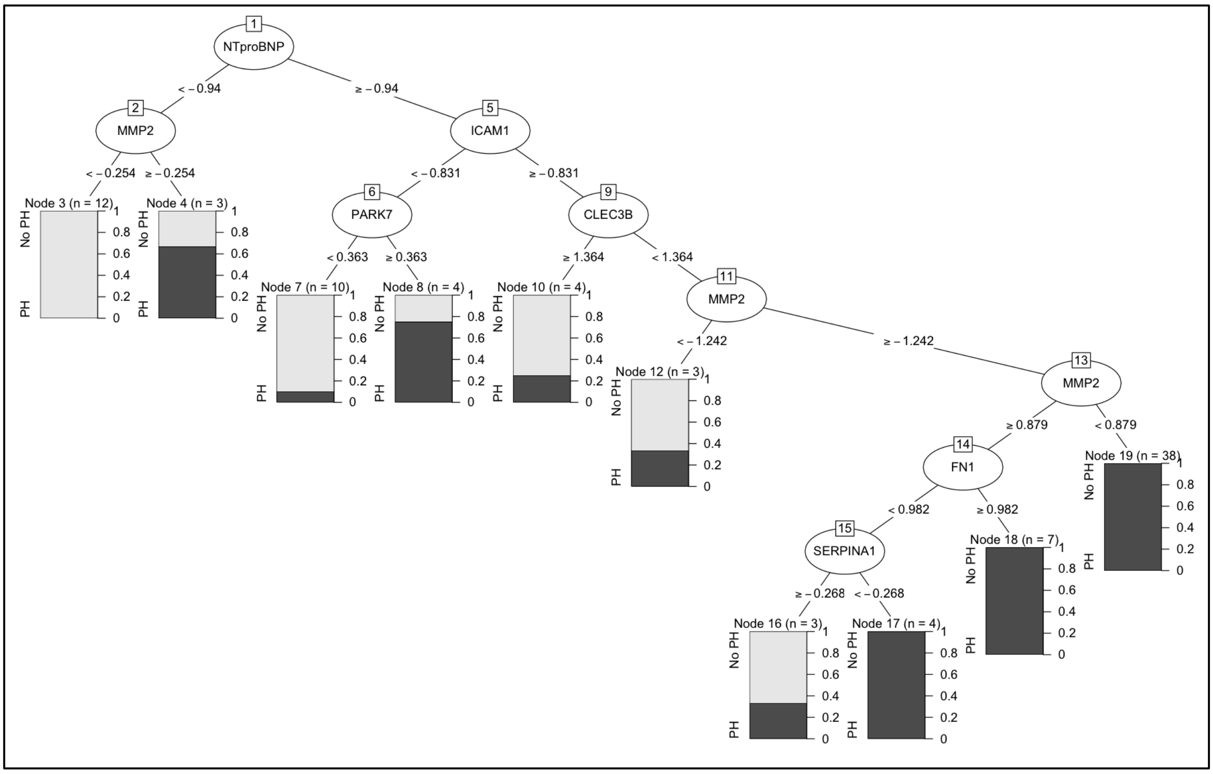


Figure 2.2: Decision Tree

Example of decision tree built in process of random forest generation. At each leaf a stacked bar chart is shown demonstrating the proportion of the total number of patients with or without PH at that leaf.

A random forest is a model generated by growing a large number of decision trees based on the same derivation dataset. The principles for each tree remain as described above, and each tree is grown independently of each other, with its own random selection of patients and proteins variables. When random forest is used for classification, the classification outcome of each individual tree for each new sample is considered a vote, with the forest outcome considered the majority vote. Through this averaged outcome from the decision trees generated, a random forest is built to reduce the error, and the resulting output is a classifying model with an additional optional ranking of important variables.

The model can return an index of variable importance for each protein, derived from a combination of how frequently that protein is selected and the sum of the reduction of heterogeneity in the outcome data of the split using that protein variable (i.e. how pure the groups are downstream of the split).

2.4.4.2.1 Assumptions of Random Forest

Random forest modelling holds few assumptions of the predictors. In particular, it works well with data that is not normally distributed, and is not affected by outliers. It does suggest predictors be independent of each other to prevent influencing the averaging of trees. If predictors are similar ie collinear, they will vote the same way and therefore influence the process of averaging outcome across all trees. Furthermore, if a collinear variable is selected at a particular split, this will adversely influence the calculated variable importance for its covariate at a distal split.

2.4.4.2.2 Application of Random Forest

Random forest models were applied to the dataset using randomForestSRC package for R. Random forest modelling was performed using random forest optimized for classification. Forest size was determined by examination of forest error rate during forest growth, and a forest size of 5000 trees was selected as this was well above size at which model error stabilized. Protein variables were then ranked according to variable importance.

Within this work, random forest modelling was used to support variable selection through variable importance ranking, rather than for classification modelling. As such, acknowledging the issue with overfitting the model developed, decision splits were not constrained and the trees were allowed to develop completely to individual sample terminal leaves.

2.4.4.2.3 Random Forest: Advantages

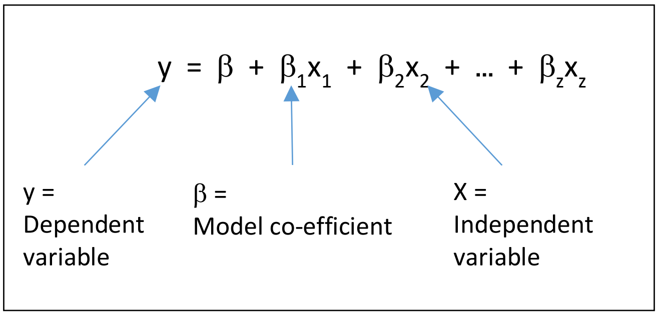
At its most basic form – the decision tree – this methodology is understandable and can simply be applied to the clinical setting. Random forests do not assume normality in the data and can be reliably handle data with outliers. Due to randomness in the selection of proteins to model, collinearity is well handled in the classifying model with minimal effect on the outcome.

2.4.4.2.4 Random Forest: Disadvantages

Due to the depth of ensemble trees required to develop a random forest model, the final output is one that cannot be visualised, and therefore the predictive model exists as a "black box" for examining test data after training. Although the variable importance can be examined, this can only be interpreted as an indication of the way proteins are required within the model. Random forests are very good classifiers, but very susceptible to overfitting without reasonable pre-modelling constraint on the number of predictor variables entered, and due consideration on the level of constraint over the extent of growth of the decision tree. Without this, the tree will inevitably over grow, to produce a long final model which can perfectly identify disease in our derivation cohort and describes the derivation cohort perfectly, but cannot generalise to accurately predict disease in a new validation sample.

2.4.4.3 Least Absolute Selection and Shrinkage Operator (LASSO)

LASSO is an advanced method for linear regression, developed for higher dimensional modelling with large numbers of independent variables which often exceed the number of patient samples.



$$y = \beta + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_z x_z$$

Figure 2.3: Regression model with descriptors

The model applies a shrinkage penalty which reduces the co-efficient of any independent variable to zero if the variable does not add to the predictive value of the model, either through lack of predictive utility, or through collinearity, thereby performing selection of the important variables. In the case of collinearity, one variable is chosen at random to remain in the model with the others removed, therefore prior examination of collinear variables is essential.

LASSO model building uses internal cross validation to find the penalty term which returns the lowest error in the final model.

In diagnostic classification, the dependent variable is the diagnostic grouping, and the independent variables are the clinical parameters and protein concentrations.

2.4.4.3.1 Assumptions of LASSO

Variables should be linear, scaled and independent. Normality need not be assumed and the model can handle multicollinearity.

2.4.4.3.2 Application of LASSO

LASSO models were built using the GLMnet package for R. All data were centred and scaled prior to modelling, and $\log_2$ transformed as previously described. Collinearity was assessed and eliminated by review of collinear proteins and selection of the appropriate protein to retain from each group based on biological relevance after review of published literature and evidence related to protein function.

As described above with regards to the penalty applied to the model, the length of the LASSO model generated for this work was constrained using the lambda min parameter to control final model length.

Protein coefficient data were extracted from the final model and ranked according to absolute value.

2.4.4.3.3 LASSO: Advantages

LASSO returns a single model which can be written out by hand if necessary in the form of a standard linear regression formula and then applied to test data. It is derived through complex machine learning and statistics, but can be applied relatively simply to new data. It is therefore a clear model compared to the random forest and avoids the concern regarding "black box" models. LASSO is reported to be able to handle non-parametric data.

2.4.4.3.4 LASSO: Disadvantages

Although it is stated that the model can handle collinearity, where variables are correlated, LASSO will indiscriminately select one of the correlated variables to take forward and drop the others from the result. As the statistics to generate the final model are complex, it is important to be aware of this and assess for collinearity prior to modelling as there is no report on this in the output.

The final model is reported as independent variable identifiers and their co-efficients, but the LASSO process does not return any measure of statistical significance such as p-values which are returned and expected as standard in other linear regression models. It is therefore more difficult to understand the significance of each individual variable in the model to the dependent variable, which is important when trying to assess whether a model is likely to be overfit.

2.4.4.4 Logistic Regression with backward step Akaike information criterion (AIC)

Logistic regression is a method for modelling classification problems, with a binary outcome. It is used to determine the relationship between features and the probability of a particular outcome.

2.4.4.4.1 Assumptions of Logistic Regression

For this type of classification using binary logistic regression, the dependent variable must be binary categorical. The predictor variable measurements should be independent of each other (ie no repeated measurements). There should not be collinearity between the predictor variables. The dependent variables need not be normally distributed.

2.4.4.4.2 Akaike information criterion (AIC)

AIC is a metric used to assess model fit, and can be used to compare different statistical models. Overfitting is a risk when deriving a statistical model, increased by an increase in the number of predictors in a model compared to the number of samples. AIC uses the log-likelihood metric (a measure of how well the model fits the data) from a logistic regression model and penalizes it for the greater number of variables in the model, returning a metric which can be used to compare between similar models and allow the one which has the optimal balance between fitting the derivation dataset and avoidance of overfitting.

I have used a backward step-AIC model from the MASS package in R to calculate an AIC metric for my logistic regression model. This algorithm starts by calculating the AIC metric for the full model, and then removes variables and recalculates. If the AIC score is lower with variables removed, then this becomes the new accepted model and the algorithm repeats until the lowest AIC metric is found, this then represents the model that best fits the data, with a low enough number of predictors to reduce the likelihood of model overfitting.

2.5 Validation Assays

External serum samples were received from the Vera Moulton Wall Centre for Pulmonary Vascular Diseases, Stanford University School of Medicine and from the Division of Allergy, Pulmonary and Critical Care Medicine, Vanderbilt University, US. These were couriered on dry ice and received intact. They were temporarily stored at -80°C until assayed.

Samples were received without associated clinical data, and therefore assays were performed blind to patient classification.

All samples were randomized as to plate and location, with duplicates plated together and standards plated together as shown. Five samples from the derivation cohort were randomized and plated alongside the validation samples to act as 'plate anchors'. As constant biological samples between plates, and between our assays and the external Myriad RBM assays, these samples can be used in addition to the plate standards and controls to assess for intra- and inter-assay variability, and additionally be used to assess for any consistent statistical correction necessary between assay batches.

2.5.1 CLEC3B ELISA assay

CLEC3B serum concentrations were determined from the external samples using an ELISA assay kit (Abcam ab213832, Detection range 312 pg/ml to 20000 pg/ml). Expected serum concentrations were taken from the derivation dataset results which fell in the range 6.4 µg/ml to 23 µg/ml, therefore each sample was diluted by of 1:1000 prior to assay. Assay standards were made up by diluting a known concentration of lyophilized recombinant human Tetranectin protein with sample diluent buffer and a 1:2 serial dilution, resulting in standard concentrations for standard 7 to standard 1 of 20,000 pg/ml, 10000 pg/ml, 5000 pg/ml, 2500 pg/ml, 1250 pg/ml, 625 pg/ml, and 312.5 pg/ml respectively. Background standard was sample diluent buffer alone. 100 µl of each test sample and standard were plated as per Figure 2.4 in the provided pre-coated anti-human Tetranectin 96-well plate, with each sample position verified by 2 independent operators.

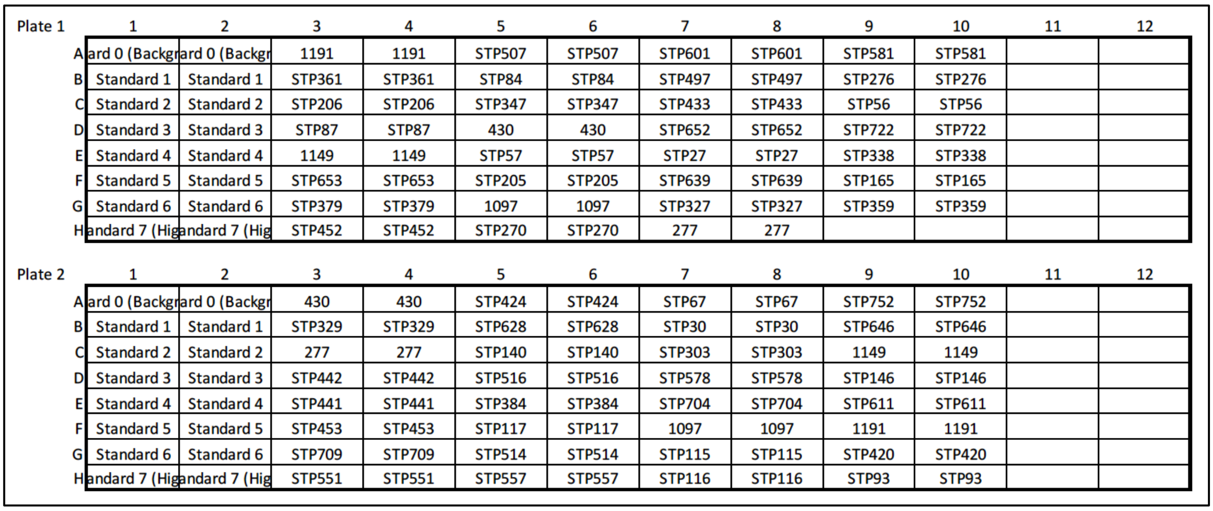


Plate 1	1	2	3	4	5	6	7	8	9	10	11	12
A	ard 0 (Backgr	ard 0 (Backgr	1191	1191	STP507	STP507	STP601	STP601	STP581	STP581		
B	Standard 1	Standard 1	STP361	STP361	STP84	STP84	STP497	STP497	STP276	STP276		
C	Standard 2	Standard 2	STP206	STP206	STP347	STP347	STP433	STP433	STP56	STP56		
D	Standard 3	Standard 3	STP87	STP87	430	430	STP652	STP652	STP722	STP722		
E	Standard 4	Standard 4	1149	1149	STP57	STP57	STP27	STP27	STP338	STP338		
F	Standard 5	Standard 5	STP653	STP653	STP205	STP205	STP639	STP639	STP165	STP165		
G	Standard 6	Standard 6	STP379	STP379	1097	1097	STP327	STP327	STP359	STP359		
H	andard 7 (Hig	andard 7 (Hig	STP452	STP452	STP270	STP270	277	277				

Plate 2	1	2	3	4	5	6	7	8	9	10	11	12
A	ard 0 (Backgr	ard 0 (Backgr	430	430	STP424	STP424	STP67	STP67	STP752	STP752		
B	Standard 1	Standard 1	STP329	STP329	STP628	STP628	STP30	STP30	STP646	STP646		
C	Standard 2	Standard 2	277	277	STP140	STP140	STP303	STP303	1149	1149		
D	Standard 3	Standard 3	STP442	STP442	STP516	STP516	STP578	STP578	STP146	STP146		
E	Standard 4	Standard 4	STP441	STP441	STP384	STP384	STP704	STP704	STP611	STP611		
F	Standard 5	Standard 5	STP453	STP453	STP117	STP117	1097	1097	1191	1191		
G	Standard 6	Standard 6	STP709	STP709	STP514	STP514	STP115	STP115	STP420	STP420		
H	andard 7 (Hig	andard 7 (Hig	STP551	STP551	STP557	STP557	STP116	STP116	STP93	STP93		

Figure 2.4: Plate map for CLEC3B ELISA assay in 96 well plate

Showing plate map for Stanford samples as example. Note differences between platemaps for positions of standards and background.

Following protocol, each plate was sealed and incubated at 37°C for 90 minutes before discarding the remaining contents of the plate and adding 100 µl of biotinylated anti-human Tetranectin working solution to each well. The plate was again sealed and incubated at 37°C for 60 minutes. Each well was washed 3 times with PBS, followed by the addition of 100 µl of prepared ABC working solution to each well. The plate was sealed and incubated at 37°C for 30 minutes. Each well was carefully washed 5 times with PBS, followed by the addition of 90 µl of prepared colour developing agent added to each well. A final incubation period of 15 minutes followed at 37°C in the dark. Finally, 100 µl of prepared TMB stop solution was added to each well. Assay results were determined by reading each well on a plate reader on photometric setting for absorbance of 450 nm. Standard curves were generated using Prism 7 for Mac, and sample protein concentrations interpolated from standard curve. Measured protein concentrations were multiplied x1000 to account for the initial sample dilution step.

2.5.2 NTproBNP Luminex assay

The initial NTproBNP assays were conducted using a Luminex kit assay (Milliplex MAP kit, Human CVD, magnetic bead panel 1, HCVD1MAG-67K, Detection range 34.3 pg/ml to 25000 pg/ml). Following the assay protocol, no serum sample dilution was required for this assay.

To prime the plates, 100 µl of assay buffer was added to each well of 96 well plate. The plate was then sealed and mixed on a plate shaker for 10 minutes at room temperature. The well contents were then discarded. NTproBNP protein standards made up from a known protein concentration by reconstituting panel 1 standard with 250 µl deionized water, mixed and vortexed to produce NTproBNP standard 1 at 25000 pg/ml. Further standards were made by 1:3 serial dilution with assay buffer to make standard 2 to 7 at concentrations 8333.3 pg/ml, 2777.8 pg/ml, 925.9 pg/ml, 308.6 pg/ml, 102.9 pg/ml and 34.3 pg/ml respectively. Background standard was assay buffer alone. The NTproBNP luminex assay kit also provided 2 quality control samples which were also reconstituted with 250 µl deionized water, mixed and vortexed. 25 µl of serum matrix was added to standard, control and background wells and 25 µl of assay buffer was added to each test sample well. 25 µl of standards, samples and controls were added to the appropriate wells according to our platemap (Figure 2.5). All sample placements were verified by 2 independent operators.

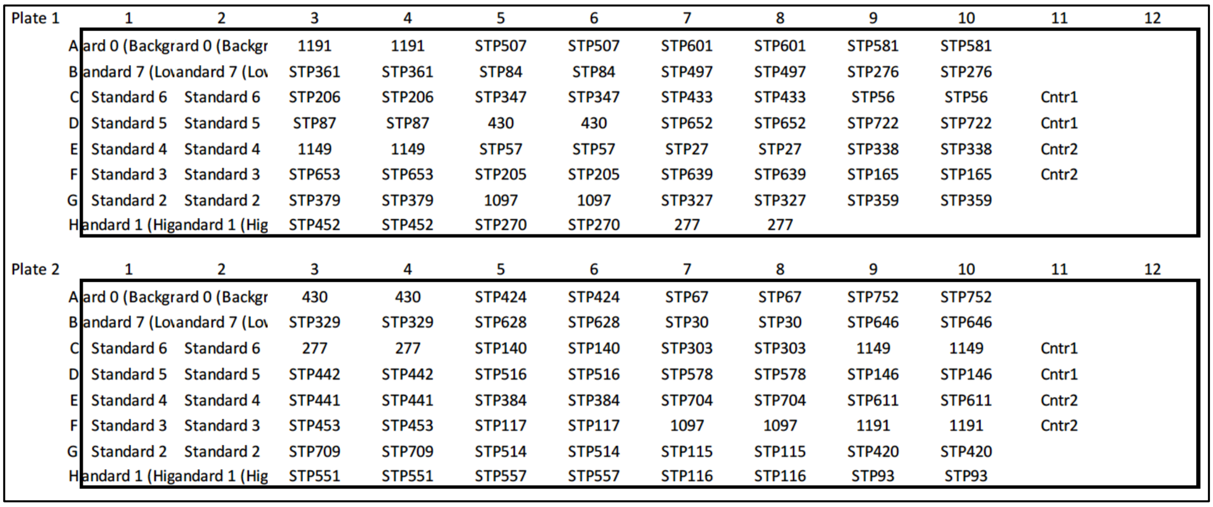


Plate 1	1	2	3	4	5	6	7	8	9	10	11	12
A	ard 0 (Backgr	ard 0 (Backgr	1191	1191	STP507	STP507	STP601	STP601	STP581	STP581		
B	andard 7 (Lo\	andard 7 (Lo\	STP361	STP361	STP84	STP84	STP497	STP497	STP276	STP276		
C	Standard 6	Standard 6	STP206	STP206	STP347	STP347	STP433	STP433	STP56	STP56	Cntr1	
D	Standard 5	Standard 5	STP87	STP87	430	430	STP652	STP652	STP722	STP722	Cntr1	
E	Standard 4	Standard 4	1149	1149	STP57	STP57	STP27	STP27	STP338	STP338	Cntr2	
F	Standard 3	Standard 3	STP653	STP653	STP205	STP205	STP639	STP639	STP165	STP165	Cntr2	
G	Standard 2	Standard 2	STP379	STP379	1097	1097	STP327	STP327	STP359	STP359		
H	andard 1 (Hig	andard 1 (Hig	STP452	STP452	STP270	STP270	277	277				

Plate 2	1	2	3	4	5	6	7	8	9	10	11	12
A	ard 0 (Backgr	ard 0 (Backgr	430	430	STP424	STP424	STP67	STP67	STP752	STP752		
B	andard 7 (Lo\	andard 7 (Lo\	STP329	STP329	STP628	STP628	STP30	STP30	STP646	STP646		
C	Standard 6	Standard 6	277	277	STP140	STP140	STP303	STP303	1149	1149	Cntr1	
D	Standard 5	Standard 5	STP442	STP442	STP516	STP516	STP578	STP578	STP146	STP146	Cntr1	
E	Standard 4	Standard 4	STP441	STP441	STP384	STP384	STP704	STP704	STP611	STP611	Cntr2	
F	Standard 3	Standard 3	STP453	STP453	STP117	STP117	1097	1097	1191	1191	Cntr2	
G	Standard 2	Standard 2	STP709	STP709	STP514	STP514	STP115	STP115	STP420	STP420		
H	andard 1 (Hig	andard 1 (Hig	STP551	STP551	STP557	STP557	STP116	STP116	STP93	STP93		

Figure 2.5: Plate map for NTproBNP Luminex assay in 96 well plate

Showing plate map for Stanford samples as example. Note differences between platemaps for positions of standards and background.

The Luminex bead vials were sonicated for 30 seconds, then vortexed for 1 minute. 150 µl of this was then added to the mixing bottle and made up to 3000 µl with bead diluent and again vortexed. 25 µl of the bead mixture was added to every well. The plate was sealed and incubated at 4°C for 18 hours in the dark. To wash the luminex plates, each plate was coupled with a magnet, and after 60 seconds the plate contents discarded. Wells were washed 3 times by removal of the magnet, addition of wash buffer, before reapplication of the magnet and discard of the well contents after 60 seconds. After the first wash, 50 µl of detection antibodies were added to each well. The plate was sealed with foil and shaken for 1 hour at room temperature. 50 µl Streptavidin-phycoerythrin was added to each well. The plates were again sealed with foil and shaken for 30 minutes at room temperature. Each plate was then washed 3 times as previously described. 100 µl sheath fluid was added to all wells and the plate shaken for 5 minutes. Plates were read on Luminex xMAP machine in house. Standard curves were created in Prism 7 for MAC, and test sample protein concentrations interpolated from the standard curve.

2.5.3 COL18A1, GDF15, IL6ST, IGFBP7, PARK7 Luminex assay

Measurement of COL18A1, GDF15, IL6ST, IGFBP7 and PARK7 serum concentrations was done using a Luminex kit assay (R&D systems, Catalogue: LXSAHM-05, Lot: 424495. Detection ranges: COL18A1 43.1 pg/ml to 31450 pg/ml; GDF15 6.8pg/ml to 4950 pg/ml; IL6ST 131.1 pg/ml to 95540 pg/ml; IGFBP7 101.7 pg/ml to 74150 pg/ml; PARK7 86.6 pg/ml to 63120 pg/ml). Following the assay protocol, test serum samples were diluted 1:2 at the outset. Standards were made up with calibrator diluent RD6-52 as instructed to known protein concentrations, and serial diluted 1:3 to produce standards 1 – 7 at concentrations shown in Figure 2.6. Background standard was calibrator diluent RD6-52 alone.

Standard	IL6ST	PARK7	GDF15	COL18A1	IGFBP7
Standard 0	0	0	0	0	0
Standard 1	95540	63120	4950	31450	74150
Standard 2	31846.7	21040	1650	10483.3	24716.7
Standard 3	10615.6	7013.3	550	3494.4	8238.9
Standard 4	3538.5	2337.8	183.3	1164.8	2746.3
Standard 5	1179.5	779.3	61.1	388.3	915.4
Standard 6	393.2	259.8	20.4	129.4	305.1
Standard 7	131.1	86.6	6.8	43.1	101.7

Figure 2.6: Serial standard protein concentrations (pg/ml)

50 µl of each standard, background or test sample was added to the appropriate wells according to our platemap (Figure 2.7). The plating position was verified by two operators.

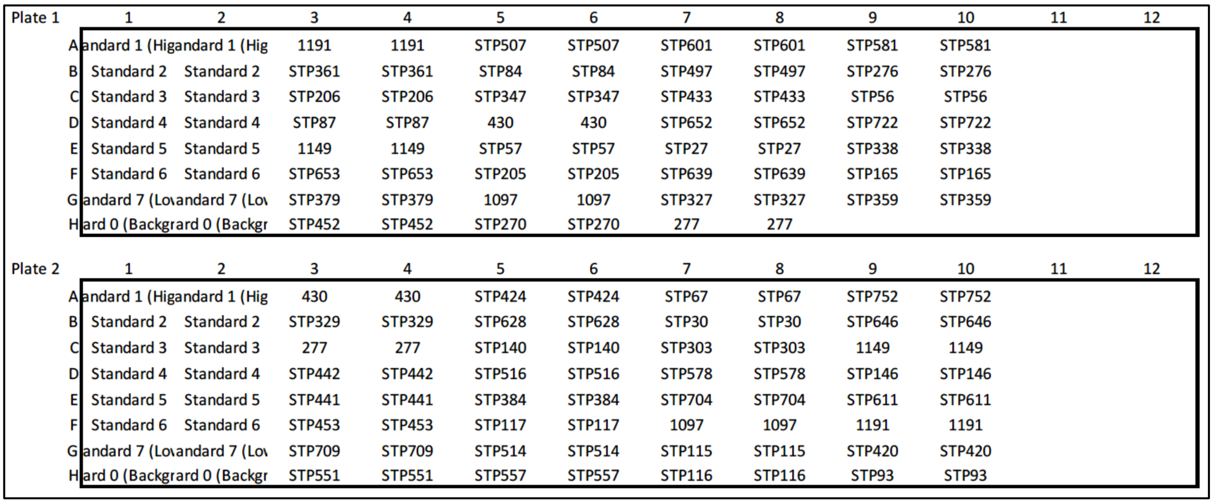


Plate 1	1	2	3	4	5	6	7	8	9	10	11	12
A	andard 1 (Hig	andard 1 (Hig	1191	1191	STP507	STP507	STP601	STP601	STP581	STP581		
B	Standard 2	Standard 2	STP361	STP361	STP84	STP84	STP497	STP497	STP276	STP276		
C	Standard 3	Standard 3	STP206	STP206	STP347	STP347	STP433	STP433	STP56	STP56		
D	Standard 4	Standard 4	STP87	STP87	430	430	STP652	STP652	STP722	STP722		
E	Standard 5	Standard 5	1149	1149	STP57	STP57	STP27	STP27	STP338	STP338		
F	Standard 6	Standard 6	STP653	STP653	STP205	STP205	STP639	STP639	STP165	STP165		
G	andard 7 (Lo	andard 7 (Lo	STP379	STP379	1097	1097	STP327	STP327	STP359	STP359		
H	ard 0 (Backgr	ard 0 (Backgr	STP452	STP452	STP270	STP270	277	277				

Plate 2	1	2	3	4	5	6	7	8	9	10	11	12
A	andard 1 (Hig	andard 1 (Hig	430	430	STP424	STP424	STP67	STP67	STP752	STP752		
B	Standard 2	Standard 2	STP329	STP329	STP628	STP628	STP30	STP30	STP646	STP646		
C	Standard 3	Standard 3	277	277	STP140	STP140	STP303	STP303	1149	1149		
D	Standard 4	Standard 4	STP442	STP442	STP516	STP516	STP578	STP578	STP146	STP146		
E	Standard 5	Standard 5	STP441	STP441	STP384	STP384	STP704	STP704	STP611	STP611		
F	Standard 6	Standard 6	STP453	STP453	STP117	STP117	1097	1097	1191	1191		
G	andard 7 (Lo	andard 7 (Lo	STP709	STP709	STP514	STP514	STP115	STP115	STP420	STP420		
H	ard 0 (Backgr	ard 0 (Backgr	STP551	STP551	STP557	STP557	STP116	STP116	STP93	STP93		

Figure 2.7: Plate map for COL18A1, GDF15, IL6ST, PARK7, IGFBP7 Luminex assay in 96 well plate

Showing plate map for Stanford samples as example. Note differences between platemaps for positions of standards and background.

50 µl of the microparticle cocktail was added to each well, and the plate then covered with foil and shaken for 2 hours at room temperature. Following the same magnet procedure as described for the NTproBNP luminex assay in section 2.5.2, each plate was washed 3 times. 50 µl of Biotin antibody cocktail was added to each well, the plate then covered with foil and shaken for 1 hour at room temperature. Each plate was washed 3 times before 50 µl Streptavidin-PE was added to each well. The plate was covered with foil and shaken for 30 minutes at room temperature. Plates were washed for a final time followed by the addition of 100 µl wash buffer to each well. Plates then shaken for 2 minutes at room temperature. All plates were read on our in house Luminex xMAP machine. Standard curves were created in Prism 7 for MAC, and sample protein concentrations interpolated from the standard curves. Protein concentrations were multiplied x2 to account for initial sample dilution step.

2.5.4 NTproBNP ELISA assay

A second NTproBNP assay was done using an ELISA DuoSet assay (R&D Systems DY3604-05, Lot P148146, Detection range 312 pg/ml to 150000 pg/ml). No serum sample dilution was required for this assay. The capture antibody was diluted to the working concentration of 4 µg/ml with PBS and 100 µl immediately added to each active well of the three 96-well plates (Figure 2.8). The plates were sealed and incubated overnight at room temperature. The contents of each well was aspirated and washed with wash buffer three times, tipping out the contents and blotting the plate at each cycle. Plates were blocked by adding 300 µl of reagent diluent to each well, and incubated for 1 hour at room temperature before washing each plate again. Standards were made up by diluting supplied standard with 0.5 ml of reagent diluent to give Standard 7 at 150000 pg/ml. The next standard was made up by making up 200 µl of Standard 7 to 3000 µl with reagent diluent, giving Standard 6 at 10000 pg/ml. Standards 5 to 1 were made by serial 1:2 dilution of Standard 6 with reagent diluent resulting in concentrations 5000 pg/ml, 2500 pg/ml, 1250 pg/ml, 625 pg/ml, and 313 pg/ml respectively. Background standard was reagent diluent alone. 100 µl of test samples and standards were plated as per Figure 2.8, with sample positions verified by two operators. The plate was sealed and incubated for 2 hours at room temperature.

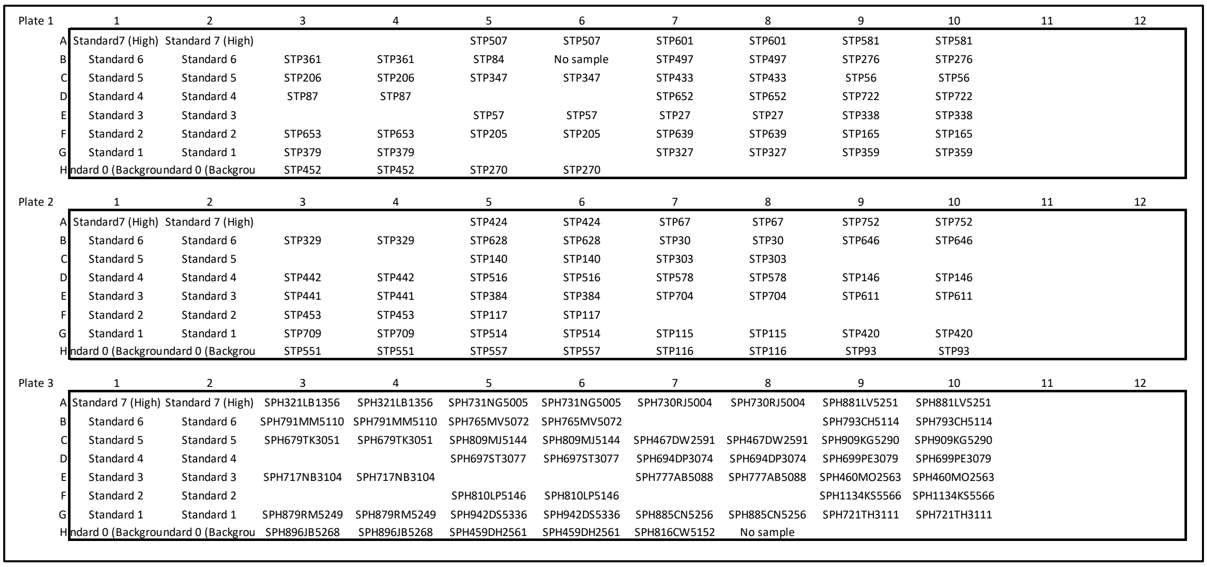


Plate 1	1	2	3	4	5	6	7	8	9	10	11	12
A	Standard7 (High)	Standard 7 (High)			STP507	STP507	STP601	STP601	STP581	STP581		
B	Standard 6	Standard 6	STP361	STP361	STP84	No sample	STP497	STP497	STP276	STP276		
C	Standard 5	Standard 5	STP206	STP206	STP347	STP347	STP433	STP433	STP56	STP56		
D	Standard 4	Standard 4	STP87	STP87			STP652	STP652	STP722	STP722		
E	Standard 3	Standard 3			STP57	STP57	STP27	STP27	STP338	STP338		
F	Standard 2	Standard 2	STP653	STP653	STP205	STP205	STP639	STP639	STP165	STP165		
G	Standard 1	Standard 1	STP379	STP379			STP327	STP327	STP359	STP359		
H	ndard 0 (Backgrou	ndard 0 (Backgrou	STP452	STP452	STP270	STP270						

Plate 2	1	2	3	4	5	6	7	8	9	10	11	12
A	Standard7 (High)	Standard 7 (High)			STP424	STP424	STP67	STP67	STP752	STP752		
B	Standard 6	Standard 6	STP329	STP329	STP628	STP628	STP30	STP30	STP646	STP646		
C	Standard 5	Standard 5			STP140	STP140	STP303	STP303				
D	Standard 4	Standard 4	STP442	STP442	STP516	STP516	STP578	STP578	STP146	STP146		
E	Standard 3	Standard 3	STP441	STP441	STP384	STP384	STP704	STP704	STP611	STP611		
F	Standard 2	Standard 2	STP453	STP453	STP117	STP117						
G	Standard 1	Standard 1	STP709	STP709	STP514	STP514	STP115	STP115	STP420	STP420		
H	ndard 0 (Backgrou	ndard 0 (Backgrou	STP551	STP551	STP557	STP557	STP116	STP116	STP93	STP93		

Plate 3	1	2	3	4	5	6	7	8	9	10	11	12
A	Standard 7 (High)	Standard 7 (High)	SPH321LB1356	SPH321LB1356	SPH731NG5005	SPH731NG5005	SPH730RJ5004	SPH730RJ5004	SPH881LV5251	SPH881LV5251		
B	Standard 6	Standard 6	SPH791MM5110	SPH791MM5110	SPH765MV5072	SPH765MV5072			SPH793CH5114	SPH793CH5114		
C	Standard 5	Standard 5	SPH679TK3051	SPH679TK3051	SPH809MJ5144	SPH809MJ5144	SPH467DW2591	SPH467DW2591	SPH909KG5290	SPH909KG5290		
D	Standard 4	Standard 4			SPH697ST3077	SPH697ST3077	SPH694DP3074	SPH694DP3074	SPH699PE3079	SPH699PE3079		
E	Standard 3	Standard 3	SPH717NB3104	SPH717NB3104			SPH777AB5088	SPH777AB5088	SPH460MO2563	SPH460MO2563		
F	Standard 2	Standard 2			SPH810LP5146	SPH810LP5146			SPH1134KS5566	SPH1134KS5566		
G	Standard 1	Standard 1	SPH879RM5249	SPH879RM5249	SPH942DS5336	SPH942DS5336	SPH885CN5256	SPH885CN5256	SPH721TH3111	SPH721TH3111		
H	ndard 0 (Backgrou	ndard 0 (Backgrou	SPH896JB5268	SPH896JB5268	SPH459DH2561	SPH459DH2561	SPH816CW5152	No sample				

Figure 2.8: Plate layout for NTproBNP ELISA assay

3 plates in total performed to include both Stanford and Vanderbilt samples.

Each plate was washed, then 100 µl of the detection antibody was added to each well, and the plate covered and incubated for 2 hours at room temperature. Each plate was washed, then 100 µl of Streptavidin-HRP added to each well. Plates were covered and incubated for 20 mins at room temperature in the dark. Plates were washed a final time, then 100 µl of substrate solution was added to each well. Plates were covered and incubated for 20 mins at room temperature in the dark. 50 µl of stop solution was added to each well, ensuring thorough mixing. Plates were read on plate reader, with photometric setting and a wavelength of 450 nm. Plates were also read at wavelength 540 nm. Readings at 540 nm subtracted from those at 450 nm to correct for optical imperfections in the plate. Standard curves were created in Prism 7 for Mac, and test sample protein concentrations were interpolated from the standard curve.

2.6 In vitro mechanisms

Statistical modelling identified a mixture of proteins which are either previously well described with regards to their pathophysiology in PAH, or those which are novel candidate proteins for further investigation. We took some of the most frequently reported and highest ranked novel proteins forward for in vitro investigation to identify their role in this disease.

2.6.1 Immunohistochemistry

Immunohistochemistry was performed on healthy and diseased lung tissue sections from human and rat species. Diseased rat lung tissues were from Sugen hypoxic rat models of PAH. Human disease tissues were from human explants at the time of transplantation, with healthy lung tissues from healthy areas of lung from patients undergoing surgery for lung nodules or early stage lung cancers.

Tissues were investigated for the presence or absence of-, and the location of the expression of PARK7, CLEC3B and the receptor of GDF-15 – GDNF family receptor alpha like (GFRAL). Primary antibodies for IHC were obtained from Abcam (polyclonal rabbit anti-PARK7 ab18257, polyclonal rabbit anti-CLEC3B ab202134, polyclonal rabbit anti-GFRAL ab235111).

2.6.1.1 Chromogenic IHC

Slides were de-waxed and rehydrated through graded ethanol to water. (Xylene 10mins, Xylene 10 mins, 100% ethanol 2 mins, 100% ethanol 2 mins, 90% (v/v) ethanol 2 mins, 70% (v/v) ethanol 2 mins, 50% (v/v) ethanol 2 mins, tap water 10 mins). Endogenous peroxidases were blocked by incubation in 3% (v/v) hydrogen peroxide (stock 30% w/v hydrogen peroxide, LP chemicals ltd. Diluted with methanol) for 10 mins followed by a rinse in tap water. Antigen retrieval was performed for PARK7 and GFRAL experiments only by incubating in 10mM citrate buffer at pH 6.0 with 0.05% (v/v) Tween-20 at 95°C for 20 minutes. Non-specific binding of secondary antibody was blocked with 1% (w/v) milk buffer for 30 mins at room temperature. The milk buffer was tipped away (not washed), and excess blotted from each slide. Each slide was incubated with primary antibody (PARK7 1:200 dilution; CLEC3B 1:200 dilution; GFRAL 1:50 dilution) overnight at 4°C. Each slide was washed in 3 changes of PBS with 0.05% (v/v) Tween-20 (vWR chemicals, Cat: 663684B) for 5 minutes. Biotinylated secondary antibodies (PARK7, CLEC3B, GFRAL: 1:200 dilution, Vector laboratories biotinylated goat anti-rabbit antibody BA-1000) were incubated on slides for 30 mins at room temperature. Unbound secondary antibodies were washed with PBS with 0.05% (v/v) Tween-20. Slides were incubated with ABC complex (Vectastain standard, Vector labs. Cat: PK6100) for 30 minutes at room temperature before washing again. DAB substrate (DAB substrate kit, Cell signalling technologies. Cat: 8059s) was added and reaction speed observed by light microscopy. The reaction was stopped after appropriate chromogen development by rinsing in tap water. Slides were counterstained with Carazzi's haematoxylin for 1 minute, then again washed with tap water. Each slide was dehydrated through graded alcohols, to xylene and mounted using DPX mountant.

2.6.1.2 Immunofluorescent IHC

Slides were de-waxed and rehydrated through graded ethanol to water (Xylene 10mins, Xylene 10 mins, 100% ethanol 2 mins, 100% ethanol 2 mins, 90% (v/v) ethanol 2 mins, 70% (v/v) ethanol 2 mins, 50% (v/v) ethanol 2 mins, tap water 10 mins). Endogenous peroxidases were blocked by incubation in 3% (v/v) hydrogen peroxide for 10 mins followed by a rinse in tap water. Antigen retrieval was performed for PARK7 and GFRAL experiments only by incubating in 10 mM citrate buffer at pH 6.0 with 0.05% (v/v) Tween-20 at 95°C for 20 minutes. Non-specific binding of secondary antibody was blocked with 10% (v/v) goat serum (Vector laboratories. Cat: S-1000) for 30 mins at room temperature. Goat serum was tipped away (not washed), and excess blotted. Each slide was incubated with primary antibody (PARK7 1:200; CLEC3B 1:200; GFRAL 1:50) overnight at 4°C then washed in 3 changes of PBS with 0.05% (v/v) Tween-20 for 5 minutes. Each slide was incubated with immunofluorescent secondary antibody (PARK7, CLEC3B, GFRAL: 1:200 dilution Invitrogen Alexa fluor 488 goat anti-rabbit antibody A11008) for 30 mins at room temperature in the dark, then again washed in 3 changes of PBS with 0.05% (v/v) Tween-20 for 5 minutes, this time in the dark. Slides were mounted using Vectashield vibrance with DAPI (Vector labs, Cat: H-1800) mount in the dark. Each slide was individually scanned using a Zeiss Imager Z2 microscope at x20 magnification set for fluorescence at the appropriate excitation wavelength.

2.6.2 Cell culture

To investigate the effect of novel candidate proteins in cell culture, an appropriate cell line was required to approximate to those found in the pulmonary vasculature, as such human pulmonary arterial endothelial cells (Lonza, Cat: CC-2530) and human pulmonary arterial smooth muscle cells (Lonza, Cat: CC-2581) were used.

Cells were bought at passage 3, and stored in liquid nitrogen until used. On waking, cells were passaged when confluent and growth media changed every 24-36 hours. Experiments were conducted between passages 3 and 8.

2.6.2.1 Phenotyping of H-PAEC

Endothelial cells were grown on coverslips coated in 0.1% (v/v) gelatin in a 6-well plate until confluent. Each well was washed in 3 changes of PBS. Cells were fixed by adding 2 ml of ice cold paraformaldehyde to each well for 10 mins at room temperature. Each well was washed in 3 changes of PBS. Cells were incubated in 2 ml Triton x100 for 15 mins at room temperature. Each well was again washed in 3 changes of PBS. Non-specific antibody binding was blocked by incubating cells in 10% (v/v) goat serum for 90 mins at room temperature. The contents of each well was aspirated, but wells were not washed at this step. Primary antibody was diluted to appropriate concentrations in 10% (v/v) goat serum and added to the appropriate wells. Plates were incubated for 90 mins at room temperature (SMA – 1:150 dilution, mouse monoclonal antibody against SMA, Abcam. Cat: ab7817; vWF – 1:200 dilution, rabbit polyclonal antibody against vWF, Dako. Cat: A0082; Vimentin – 1:500 dilution, rabbit monoclonal antibody against vimentin, Abcam. Cat: ab92547). Each well was washed in 3 changes of PBS. 1:200 dilution in PBS of the appropriate fluorescent secondary antibody was added to each well and incubated at room temperature in the dark for 60 mins. (Secondary antibodies: Alexa fluor 488 goat anti-rabbit antibody, Invitrogen. Cat: A11008; Alexa fluor 555 goat anti-mouse antibody, Invitrogen. Cat: A21422). Each well was washed in dark conditions with 3 changes of PBS. Coverslips were mounted onto slides using Vectashield Vibrance with DAPI (Vector labs, Cat: H-1800) mount and left to set overnight in the dark. Slides were imaged using a Zeiss Imager Z2 microscope at x20 magnification set for fluorescence at the appropriate excitation wavelength.

2.6.2.2 Proliferation assays

HPAEC proliferation assays

The assay was performed in a clear bottom, white wall 96-well plate coated with 30 µl 0.1% (v/v) gelatin and incubated for 30mins at 37°C before tipping the plate and blotting away the excess. Human endothelial cells were plated at 5000 cells/well in full growth media (EGM-2, Lonza, Cat: CC-3162) and incubated overnight at 37°C in a CO2 incubator. The growth media was removed, the plate washed with sterile PBS, and replaced with 200 µl of quiescent media (EBM-2 supplemented with 0.5% (v/v) foetal calf serum, Lonza, Cat: CC-3156) followed by a period of incubation at 37°C in a CO2 incubator for 24 hours. Media excess was tipped away and blotted. The plate was washed with sterile PBS, and 200 µl quiescent media with appropriate stimulator was added to each well (Recombinant human PARK7 protein, Abcam, Cat: ab124312; Recombinant human CLEC3B protein, R&D systems, Cat: 5170-CL-050; Recombinant human VEGFA protein, R&D systems, Cat: 293-VE-010; Recombinant human FGF protein, R&D systems, Cat: 233-FB-025). The plate was then incubated at 37°C in a CO2 incubator for 48 hours. A standard curve of cells was added to spare wells on the plate in quiescent media at total volume or 200 µl per well. 100 µl CellTiter-Glo (Promega, Cat: G7571) was added to each well and the plate was then read on plate reader on a luminescence detection setting.

HPASMC Proliferation assays

For this assay a clear bottom, white wall 96-well plate was coated with 30 µl 13.5 ng/ml fibronectin and incubated for 30mins at 37°C before tipping and blotting away the excess. Human pulmonary artery smooth muscle cells were plated at 5000 cells/well in full growth media (SMGM-2, Lonza, Cat: CC-3182) and incubated overnight at 37°C in a CO2 incubator. Media was removed, the plate was washed with sterile PBS, and replaced with 200 µl of quiescent media (1:20 dilution of SMGM-2 in SMBM (Lonza, Cat: CC-3181)). The plate was then Incubated at 37°C in a CO2 incubator for 48 hours. Excess media was tipped away and blotted. The plate was washed with sterile PBS, followed by the addition of 200 µl of quiescent media with the appropriate stimulator to each well (Recombinant human PARK7 protein, Abcam, Cat: ab124312; Recombinant human CLEC3B protein, R&D systems, Cat: 5170-CL-050; Recombinant human PDGF protein, R&D systems, Cat: 220-BB-050). The plate was incubated at 37°C in a CO2 incubator for 72 hours. A standard curve of cells was added to spare wells on the plate in quiescent media at total volume or 200 µl per well. 100 µl CellTiter-Glo (Promega, Cat: G7571) was added to each well. The plate was read on plate reader, with a luminescence detection setting.

2.6.2.3 Migration assays

This experiment was started when passaging cells were near confluence in a T75 flask. The cells were washed 3 times in sterile PBS, and excess removed with a pipette. 30 ml of quiescent media (as 2.6.2.2 Proliferation assays) was added to the T75 flask. The flask was incubated at 37°C in a CO2 incubator for 24 hours for HPAEC and 48 hours for HPASMC. Inserts from 24-well Multiwell plates (BD Falcon 351185, 8 µm pore size) were coated in 0.1% (v/v) gelatin. Cells were washed with sterile PBS and suspended in quiescent media at $12x10^4$ cells/ml. Excess gelatin was removed from the inserts and wells. 750 µl of quiescent media with the required stimulator was added to each well (Recombinant human PARK7 protein, Abcam, Cat: ab124312; Recombinant human CLEC3B protein, R&D systems, Cat: 5170-CL-050; Recombinant human VEGFA protein, R&D systems, Cat: 293-VE-010; Recombinant human FGF protein, R&D systems, Cat: 233-FB-025). 250 µl of the cell suspension was added to each insert. Plates were incubated at 37°C in a CO2 incubator for 5 hours. The bottom of the insert was carefully washed in PBS. Non migrated cells were removed from inner membrane of the insert with a cotton bud. Cells remaining on the outer surface of the membrane were fixed and stained using a Kwik-Diff kit (Thermofisher, Cat: 9990700). The inserts were left to dry for 24 hours, and images were taken from random locations in the well using an inverted light microscope (Figure 2.9). Migrated cells were counted using the cell counter application in ImageJ (Schindelin et al., 2012).

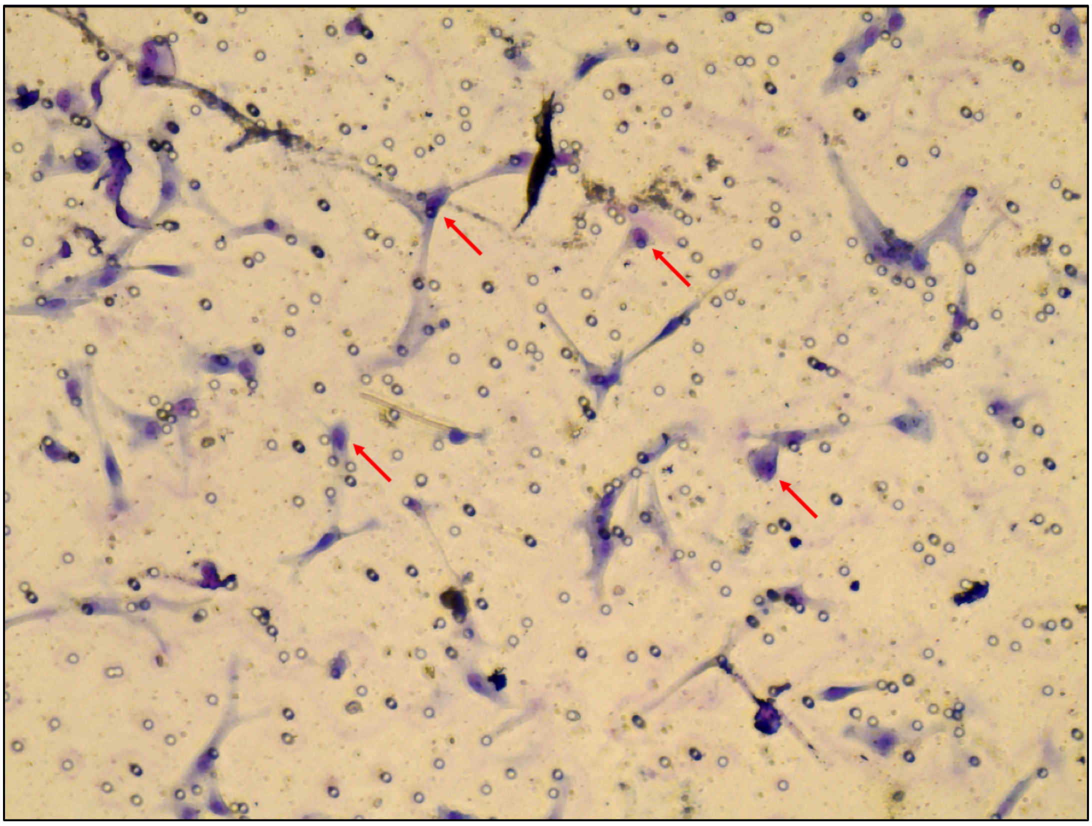


Figure 2.9: Migration assay: an example.

Image from light microscopy used to quantify migration assay. Red arrows indicate a sample of the cell bodies counted as an example of the cells counted on this slide.

2.6.2.4 Angiogenesis assays

This experiment was started once cells were near confluence in a T75 flask. Cells were washed with sterile PBS, and excess removed with a pipette. 30 ml of quiescent media (as 2.6.2.2 Proliferation assays) was added to the T75 flask. The flask was incubated at 37°C in a CO2 incubator for 24 hours. Growth factor reduced Matrigel (Corning, Cat: 11523550) was thawed gently overnight in ice. A flat bottom 96-well plate, and appropriate pipette tips were also cooled overnight in a -20°C freezer. The pre-cooled plate and pipettes were used and 40 µl GFR-Matrigel was added to each well to cover the entire growth surface while the plate was kept chilled. The plate was centrifuged at 4°C at 300G for 10 mins. The plate was then incubated at 37°C in a CO2 incubator for 30 mins to allow the Matrigel to set. Cells were suspended in quiescent media at 1x10^5 cells/ml. 100 µl of the cell suspension was added gently to each well. 100 µl of quiescent media with double concentration of the desired stimulator was added to each well (Recombinant human PARK7 protein, Abcam, Cat:

ab124312; Recombinant human CLEC3B protein, R&D systems, Cat: 5170-CL-050; Recombinant human VEGFA protein, R&D systems, Cat: 293-VE-010; Recombinant human FGF protein, R&D systems, Cat: 233-FB-025). The plate was incubated at 37°C in a CO2 incubator for 6 hours. Images were obtained from random locations in the well by inverted light microscope and analysed using the Angiogenesis analyser application in ImageJ (Carpentier, Schindelin et al., 2012)(Figure 2.10).

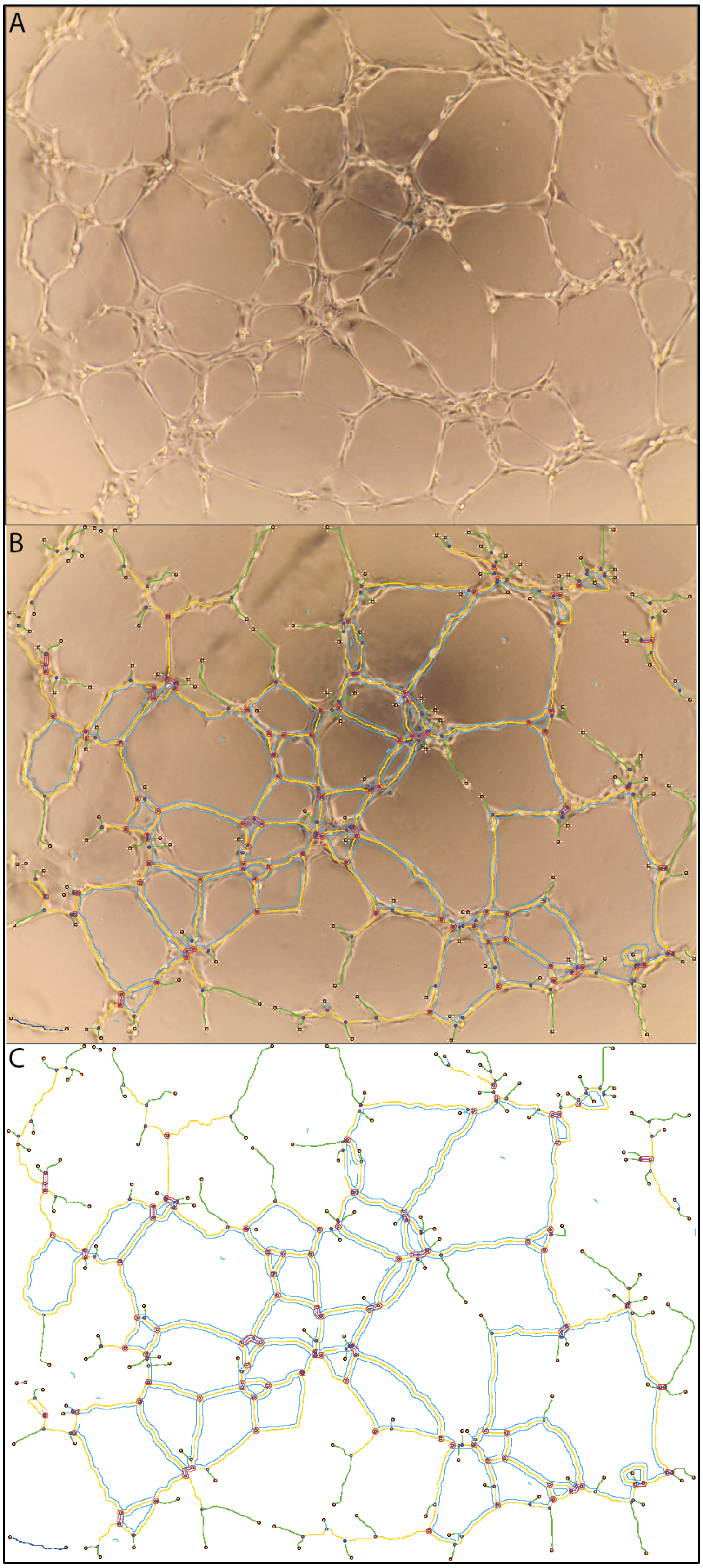


Figure 2.10: Angiogenesis analysis: an example.

An example of the angiogenesis network analysis from ImageJ. A: the original angiogenesis image from light microscopy; B: the original image overlaid with the skeleton network; C: the computer derived skeleton network for tube network measurement.

2.6.2.5 Transfection

This experiment started with cells at confluence in a T75 flask. Cells were washed and 15 ml of full growth media (EGM-2, Lonza, Cat: CC-3162) added. Lipofectamine RNAiMAX reagent (Invitrogen, Cat: 13778-150) was diluted in Opti-MEM medium (Gibco, Cat: 31985-070). 59 μl of RNAiMAX reagent was diluted in 750 μl Opti-MEM. 20 μl of the appropriate 10 μM siRNA solution (200 pmol siRNA) was added to 750 μl Opti-MEM medium (CLEC3B siRNA: Thermofisher Silencer Select, Cat: s14246; PARK7 siRNA: Thermofisher Silencer Select, Cat: s22306; Non-targeting control siRNA, Dharmacon, Cat:D-001810-01-05) The diluted Lipofectamine RNAiMAX reagent and diluted siRNA solution were combined together in 1:1 ratio to a total volume of 1500 μl. This was then incubated at room temperature for 5 mins. The siRNA-Lipofectamine RNAiMAX solution was added to the cells in the T75 flask. Cells were incubated for 6 hours at 37°C in a CO2 incubator, then washed with sterile PBS and transfection media replaced with full growth media overnight (EGM-2, Lonza, Cat: CC-3162). At 24hrs from transfection, cells were washed again, and media replaced with quiescent media (EBM-2 supplemented with 0.5% (v/v) foetal calf serum, Lonza, Cat: CC-3156). Cells were incubated overnight at 37°C in a CO2 incubator. Transfected cells were then ready to be used in a migration assay as per section 2.6.2.3 Migration assays.

2.6.2.6 Protein quantification and Western blot

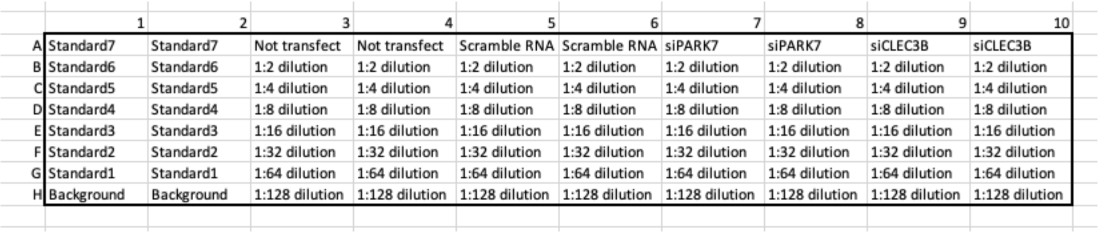
This experiment begins with the cells remaining from the post-transfection migration assay. These cells were centrifuged in falcon tubes at 2500 xG for 10 mins. Make up M-PER mammalian protein extraction reagent (Thermo, Cat: 78501) with 30 μl of Halt protease inhibitor cocktail (x100 concentration) (Thermo, Cat: 1861278) and 30 μl of Halt phosphatase inhibitor cocktail (100x concentration) (Thermo, Cat: 1861277). Remove the supernant from all the tubes and resuspend the pellet in 500 μl of M-PER solution. Shake gently for 10 minutes, then centrifuge at 14000 xG for 15 minutes to remove any cell debris. Keep the supernant for analysis.

Protein quantification

Using a 96 well plate, plate out pre-diluted protein assay standards: (Bovine serum albumin set, Thermo, Cat: 23208).

a. Standard 7: 2000 μg/ml
b. Standard 6: 1500 μg/ml
c. Standard 5: 1000 μg/ml
d. Standard 4: 750 μg/ml
e. Standard 3: 500 μg/ml
f. Standard 2: 250 μg/ml
g. Standard 1: 125 μg/ml

Plate out 10 μl of standards in the appropriate wells of the 96 well plate (Figure 2.11). Plate out 10 μl of neat protein samples in the top row of the plate and produce 2x serial dilution curve (with PBS) in case neat samples are above the top of the standard curve. Add 150 μl of Pierce 660 nm protein assay reagent (Thermo, Cat: 22660) to each well. Shake for 1 minute and stand for 5 minutes before reading on a plate reader, colorimetric setting at 660 nm. Convert results to protein concentrations using the known standard curve.



	1	2	3	4	5	6	7	8	9	10
A	Standard7	Standard7	Not transfect	Not transfect	Scramble RNA	Scramble RNA	siPARK7	siPARK7	siCLEC3B	siCLEC3B
B	Standard6	Standard6	1:2 dilution	1:2 dilution	1:2 dilution	1:2 dilution	1:2 dilution	1:2 dilution	1:2 dilution	1:2 dilution
C	Standard5	Standard5	1:4 dilution	1:4 dilution	1:4 dilution	1:4 dilution	1:4 dilution	1:4 dilution	1:4 dilution	1:4 dilution
D	Standard4	Standard4	1:8 dilution	1:8 dilution	1:8 dilution	1:8 dilution	1:8 dilution	1:8 dilution	1:8 dilution	1:8 dilution
E	Standard3	Standard3	1:16 dilution	1:16 dilution	1:16 dilution	1:16 dilution	1:16 dilution	1:16 dilution	1:16 dilution	1:16 dilution
F	Standard2	Standard2	1:32 dilution	1:32 dilution	1:32 dilution	1:32 dilution	1:32 dilution	1:32 dilution	1:32 dilution	1:32 dilution
G	Standard1	Standard1	1:64 dilution	1:64 dilution	1:64 dilution	1:64 dilution	1:64 dilution	1:64 dilution	1:64 dilution	1:64 dilution
H	Background	Background	1:128 dilution	1:128 dilution	1:128 dilution	1:128 dilution	1:128 dilution	1:128 dilution	1:128 dilution	1:128 dilution

Figure 2.11: Plate layout for protein quantification assay

Western blot

First denature proteins. To do so, make up a solution of 25 μl (4x concentration) of loading buffer (Li-Cor, Cat: 928-40004) with 10 μl of (10x concentration) reducing agent (Novex, Cat: B0004) and 65 μl of the protein sample and heat to 95°C for 2 mins. Take a gel (Invitrogen, Cat: NW04120B0X) and place in 1L Bolt running buffer (20x concentration made up with water) (Novex, Cat: B0002) together with 500 μl of Bolt antioxidant (Invitrogen, Cat: BT0005) in an electrophoresis tank. Remove the comb and add 5 μl Ladder (Li-Cor Chameleon duo, Cat: 928-60000), 35 μl Non-transfected sample, 35 μl Scramble transfected sample, 35 μl PARK7 transfected sample and 35 μl CLEC3B transfected sample to the appropriate gel wells. Run a sample for each primary antibody being tested. Run the electrophoresis for 30 mins at 200 volts. Remove the gel from the electrophoresis tank and remove the wells and foot from the gel. Using an iBlot2 dry blotting system, and Blot2 NC regular stacks (Invitrogen, Cat: IB23001), transfer to a nitrocellulose membrane using 23 Volts for 7 mins. Wash the nitrocellulose membrane in PBS, then block with Odyssey blocking buffer (Li-Cor, Cat: 927-50000) for 60 mins on a rocker at room temperature. Make up primary antibodies in Odyssey blocking buffer to 10ml total volume. Rabbit anti-PARK 7 primary antibody at 1 μg/ml (Abcam, ab18257), Rabbit anti-CLEC3B primary antibody at 1:1000 dilution from neat (Abcam, ab202134). Remove the blocking buffer from the nitrocellulose membranes and add blocking buffer with primary antibodies diluted, one container for each primary antibody and incubate overnight at 4°C. Wash the nitrocellulose membrane in PBS with 0.05% (v/v) Tween-20, repeating 3 times, each time rocking for 5 mins at room temperature. Remove wash fluid and incubate the nitrocellulose membrane in 15 ml of 1:15000 dilution of secondary antibody for 60 minutes on a rocker – IRDye 680RD Goat anti-rabbit secondary antibody (Li-Cor, Cat: 925-68071) made up in Odyssey blocking buffer. Wash the nitrocellulose membrane in PBS with 0.05% (v/v) Tween-20, repeating 3 times, each time rocking for 5 mins at room temperature. Scan the nitrocellulose membrane using Li-Cor Odyssey machine.

GAPDH

Strip the antibodies from the previously analysed nitrocellulose membranes (with anti-PARK7 and anti-CLEC3B primary antibodies) to allow for analysis of GAPDH as a normalizing protein (this has a similar molecular weight to PARK7 and CLEC3B so must be assessed separately). Place the previously analysed nitrocellulose membranes in Reblot Plus antibody stripping solution made up with deionised water (Millipore, Cat: 2502). Incubate for 10 mins at room temperature on a rocker. Wash once in PBS (without Tween-20). Scan each nitrocellulose membrane on a Li-Cor Odyssey machine to ensure there is no remaining antibody detection evident. Place the nitrocellulose membranes in Odyssey blocking buffer for 1 hour at room temperature on a rocker. Replace the buffer with 10 ml 1:1000 dilution of anti-GAPDH primary antibody made up in Odyssey blocking buffer (Cell signalling, Cat: 2118L). Incubate at room temperature for 2 hours on a rocker. Wash the nitrocellulose membrane in PBS with 0.05% (v/v) Tween-20, repeating 3 times, each time rocking for 5 mins at room temperature. Remove the wash fluid and incubate the nitrocellulose membrane in 15 ml of 1:15000 dilution of secondary antibody for 60 minutes on a rocker – IRDye 680RD Goat anti-rabbit secondary antibody (Li-Cor, Cat: 925-68071) made up in Odyssey blocking buffer. Wash the nitrocellulose membrane in PBS with 0.05% (v/v) Tween-20, repeating 3 times, each time rocking for 5 mins at room temperature. Scan the nitrocellulose membrane using a Li-Cor Odyssey machine.

Analysis

All the nitrocellulose membrane images were acquired using Li-Cor Image Studio, and signal intensity from both the primary antibody acquisitions and GAPDH acquisitions were recorded. Protein measurements were normalized to the corresponding GAPDH measurement from the same nitrocellulose membrane. Protein knockdown was calculated from the appropriate protein measurement against non-transfected cells.

3 Classification modelling and understanding sources of error

3.1 Introduction

Following the publication of the DETECT screening protocol, the bioinformatic department of Actelion pharmaceutical began work in an attempt to improve on the diagnostic accuracy by exploring the utility of adding serum protein biomarkers. They gained access to samples from the original DETECT derivation cohort and were able to analyse these in collaboration with MyriadRBM on the DiscoveryMAP platform to generate a derivation dataset with both the clinical parameters known from the DETECT derivation cohort, and the new serum protein measurements. This work generated a hybrid screening tool using both clinical and protein variables to develop a new screening tool. Actelion pharmaceutical initially approached the Sheffield pulmonary vascular disease biorepository to request serum samples across the spectrum of PH diagnostic groups to use as an external validation cohort to their work.

Through this collaboration we received the protein assay concentration dataset for all the Sheffield patients included (Table 3.1). As part of our sample sharing agreement, we were able to use these data for our own research purposes. Given the limitations of the DETECT protocol described previously (Sections: 1.2.4.2 & 1.2.4.3) we elected to used these protein concentration data to determine whether a protein only screening tool could out-perform one with clinical parameters included.

Table 3.1: Initial Sheffield cohort

		SSC-PAH	SSC-no PH	SSC-PAH-ILD	IPAH	PAH-Other CTD	PH-ILD	HV
n		22	22	16	30	9	14	29
Age (IQ range)		69 (63.3-72)	61.5 (56.3-67.8)	66.5 (62.8 - 71.3)	65 (56 - 71.8)	63 (51 - 69)	67 (53.8 - 70.8)	38 (23 - 52)
Gender (M/F)		7/15	2/20	5/11	11/19	2/7	11/3	16/13
Deaths		15	4	11	14	5	9	0
WHO FC	**1**	0	1	0	1	0	0	NA
	2	3	8	2	3	0	3	NA
	3	19	13	12	22	7	5	NA
	4	0	0	2	4	1	6	NA
Co-Morbidity	**COPD (%)**	4 (18.2)	1 (4.6)	1 (6.3)	5 (16.7)	1 (11.1)	0 (0)	0 (0)
	Haemolysis (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	Myeloproliferative (%)	1 (4.6)	0 (0)	0 (0)	1 (3.3)	0 (0)	0 (0)	0 (0)
	AF (%)	2 (9.1)	1 (4.6)	0 (0)	3 (10)	1 (11.1)	1 (7.1)	0 (0)
	A Flutter (%)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
	ILD (%)	0 (0)	9 (40.9)	16 (100)	1 (3.3)	1 (11.1)	14 (100)	0 (0)
	Asthma (%)	1 (4.6)	1 (4.6)	0 (0)	3 (10)	2 (22.2)	0 (0)	2 (6.9)
	Sarcoidosis (%)	0 (0)	1 (4.6)	0 (0)	0 (0)	0 (0)	1 (7.1)	0 (0)
	OSA (%)	2 (9.1)	0 (0)	0 (0)	1 (3.3)	0 (0)	0 (0)	0 (0)
	VTE (%)	2 (9.1)	2 (9.1)	0 (0)	2 (6.7)	2 (22.2)	0 (0)	1 (3.5)
PFTs (median + IQ))	**FEV1 Percent**	85.9 (82.3-97.6)	77.3 (68.6-101.6)	73.3 (68.6-83.6)	90.6 (67.8-104.2)	74.9 (61.2-83.9)	66 (61.6-81.5)	NA
	FVC Percent	100.8 (90.9-111.5)	94.7 (73.4-107.9)	66.6 (60-89.8)	106.8 (87.4-117.3)	89.2 (60.8-89.5)	71.1 (62-83.6)	NA
	TLCO Percent	43.6 (39.6-47)	55.1 (46.7-67.3)	32.2 (26.2-37.5)	41 (28.6-65.8)	36.3 (33.4-44.1)	27.1 (17.6-33.4)	NA
ISWT (median + IQ)	**Distance**	115 (72.5-247.5)	210 (75-338)	120 (92.5-220)	250 (120-410)	95 (45-183)	180 (55-245)	NA
RHC (median + IQ)	**mPAP**	38 (30-52)	21 (19-22)	33.5 (28.5-41.3)	48 (44.3-64.8)	35 (29-45)	42 (33.3-56.8)	NA
	RA pressure	9 (6-11)	5 (3.3-4.1)	6.5 (3.8-10.5)	10 (7.3-12)	6 (5-11)	9 (4-17)	NA
	CI	2.8 (2.3-3.4)	3.4 (3.1-4.1)	2.7 (2.7-3.5)	2.5 (2.0-3.1)	2.9 (2.6-3.3)	2.2 (2-3.2)	NA
	PVR	396 (271-850)	140 (117-171)	335 (275-494)	718 (594-959)	382 (268-640)	515 (292-702)	NA
	PCWP	12 (8-14)	8.5 (7-11.8)	9.5 (7-13)	11 (8-13)	8 (7-12)	9.5 (9-15.3)	NA

Abbreviations: n – number; IQ – interquartile; M/F – Male/Female; WHO FC – World health organisation functional class; COPD – chronic obstructive pulmonary disease; AF – Atrial fibrillation; A Flutter – Atrial flutter; ILD – interstitial lung disease; OSA – Obstructive sleep apnoea; VTE – Venous thromboembolism; PFT – Pulmonary function tests; ISWT – incremental shuttle walking test; RHC – right heart catheter

In determining their PH disease groups of interest, our research collaborators at Actelion pharmaceutical were interested in studying patients with SSc both with and without PAH, with particular interest in a “clean” SSc-PAH cohort, with a specific criterion that these patients should have no element of interstitial lung disease. Patients with SSc-PAH were therefore split out into SSc-PAH (without ILD) and SSc-PAH-ILD (patients with some ILD, but still treated and managed as group 1 PAH rather than group 3 PH). Further groups were requested for comparison including patients with IPAH and PH related to interstitial lung disease (PH-ILD; group 3 PH), PAH but with other underlying CTDs (PAH-Other CTD) and healthy volunteers (HV).

My overarching research question pertains only to the classification of PAH in patients with SSc, and therefore for the purposes of classifying within these disease groups our analyses focussed predominantly on the data from patients with SSc.

3.2 Aim

Through univariate, and multivariable modelling to determine which proteins or combinations of proteins can classify PAH in SSc.

3.3 Methods

The full patient cohort (Table 3.1) was first assessed using principle component analysis (PCA) to determine whether clear differences between all phenotype groups could be demonstrated using the full protein dataset.

PCA was then repeated on only the specific test cohort (SSc-PAH vs SSc-no PH) to reduce noise and to look for identifiable differences between only these groups using the full protein dataset.

Variable selection was performed using the methods previously described and reports univariate statistics, manual panel building, random forest and lasso models. Performance of the final derivation model is then assessed using the model prediction against the known patient classes in a confusion matrix to report diagnostic statistics.

3.4 Results: Analysis Part 1

The test cohort selected (Table 3.2) included a disease group, consisting of SSc patients with “clean” PAH (Group 1 PH), against a control group of patients with SSc without PH at RHC. In this analysis, in parallel to the cohorts studied by our collaborators, the disease group was carefully phenotyped to minimize the risk of contamination within the group from PH of other aetiologies such as lung disease as the driving pathological mechanism behind the PH (Group 3 PH, see Table 1.1).

3.4.1 Demographics

Table 3.2: Patient Demographics for Analysis Part 1.

		SSC-PAH	SSC-no PH	p-value
n		22	22	
Age (IQ range)		69 (63.3 - 72)	61.5 (56.3 - 67.8)	0.034
Gender (M/F)		7/15	2/20	0.134
Deaths		15	4	0.002
WHO FC	**1**	0	1	0.04
	2	3	8	
	3	19	13	
	4	0	0	
Co-Morbidity	**COPD (%)**	4 (18)	1 (5)	0.34
	Haemolysis (%)	0 (0)	0 (0)	NA
	Myeloproliferative (%)	1 (5)	0 (0)	1
	AF (%)	2 (9)	1 (5)	1
	A Flutter (%)	0 (0)	0 (0)	NA
	ILD (%)	0 (0)	9 (41)	0.003
	Asthma (%)	1 (5)	1 (5)	NA
	Sarcoidosis (%)	0 (0)	1 (5)	1
	OSA (%)	2 (9)	0 (0)	0.47
	VTE (%)	2 (9)	2 (9)	NA
PFTs (median + IQ))	**FEV1 Percent**	85.9 (82.3-97.6)	77.3 (68.6 - 101.6)	0.34
	FVC Percent	100.8 (90.9 - 111.5)	94.7 (73.4 - 107.9)	0.13
	TLCO Percent	43.6 (39.6 - 47)	55.1 (46.7 - 67.3)	0.002
ISWT (median + IQ)	**Distance**	115 (73 - 248)	210 (75 - 338)	0.24
RHC (median + IQ)	**mPAP**	38 (30 - 52)	21 (19 - 22)	<0.001
	RA pressure	9 (6 - 11)	5 (3.3 - 6.8)	0.005
	CI	2.8 (2.3 - 3.4)	3.4 (3.1 - 3.6)	0.004
	PVR	396 (271 - 850)	139 (117 - 171)	<0.001
	PCWP	12 (8 - 14)	8.5 (7 - 11.8)	0.07

Abbreviations: n – number; IQ – interquartile; M/F – Male/Female; WHO FC – World health organisation functional class; COPD – chronic obstructive pulmonary disease; AF – Atrial fibrillation; A Flutter – Atrial flutter; ILD – interstitial lung disease; OSA – Obstructive sleep apnoea; VTE – Venous thromboembolism; PFT – Pulmonary function tests; ISWT – incremental shuttle walking test; RHC – right heart catheter

3.4.2 Exploratory data analysis

3.4.2.1 Principle Component Analysis

PCA was initially performed on the full patient cohort using all protein variables in the dataset (Figure 3.1).

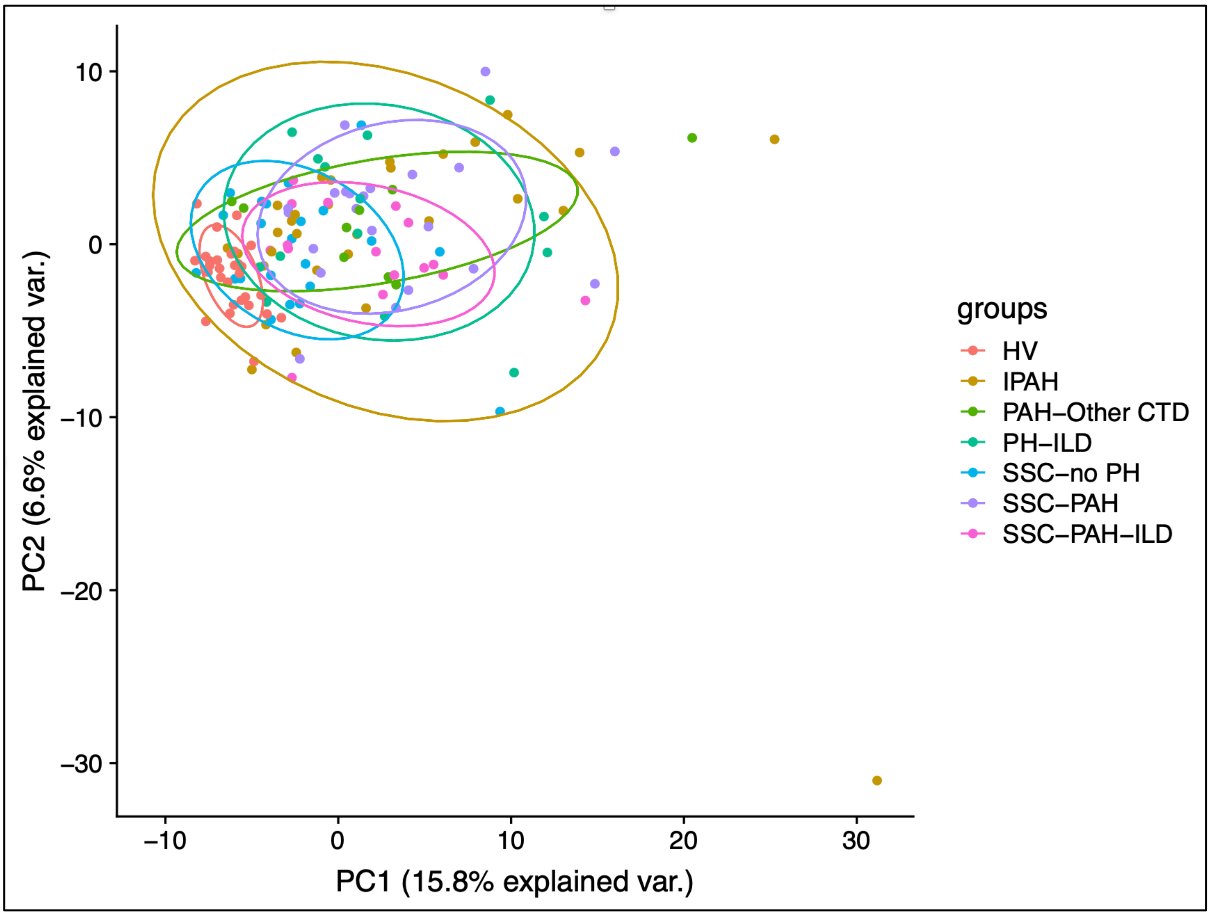


Figure 3.1: Principle component analysis of all patient classes in the initial cohort (Table 3.1).

Abbreviations: PC1 – principle component 1, PC2 – principle component 2, var. – variance.

This shows distinct clustering of healthy volunteers from all other disease groups, but no other significant clustering between the disease cohorts demonstrable in the first two principle components. Only 22.4% of the variance is explained by the first two principle components, suggesting a high level of complexity in the dataset, and low level of data redundancy.

PCA was also performed on the test cohort (SSc-PAH vs SSc-no PH) again using all proteins available in the dataset (Figure 3.2).

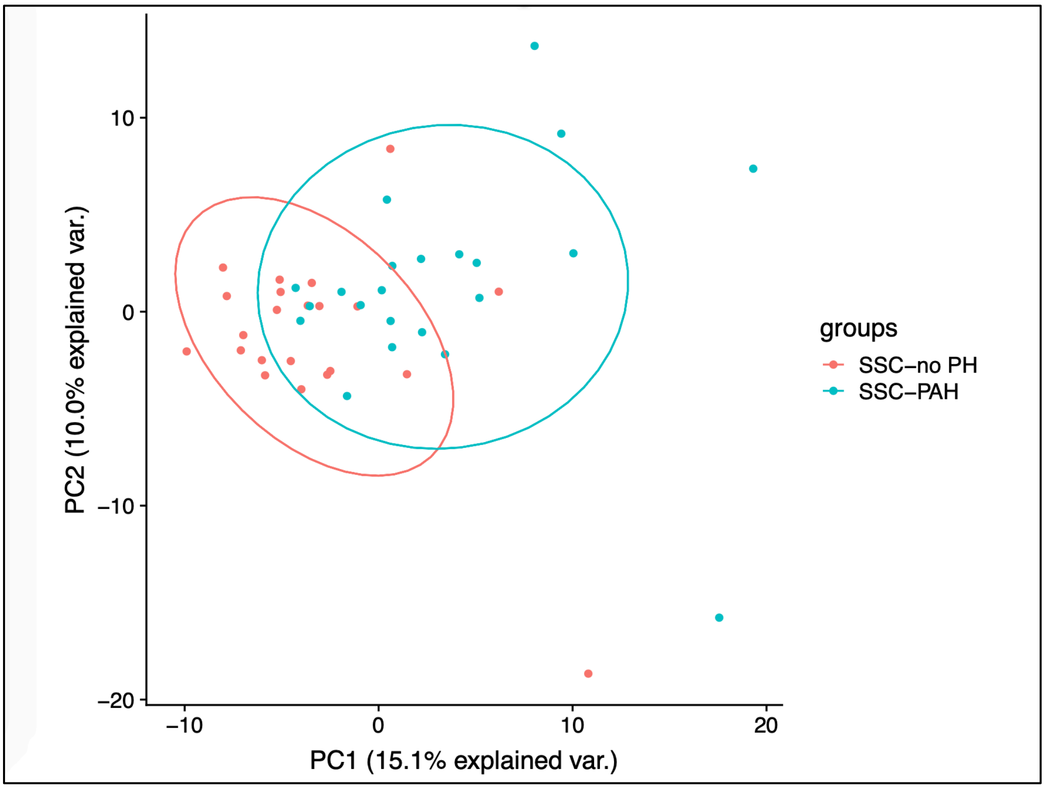


Figure 3.2: PCA of SSc-PAH and SSc-no PH.

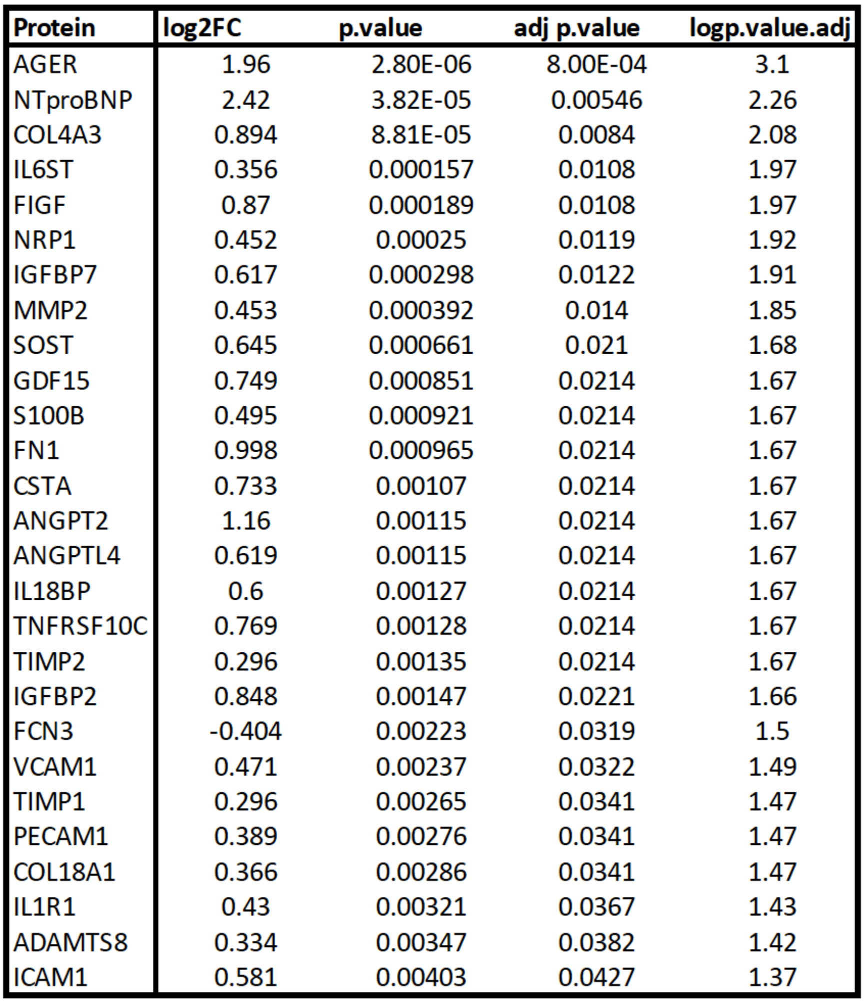
Abbreviations: PC1 – principle component 1, PC2 – principle component 2, var. – variance.

This PCA demonstrates some separation of the two patient classes based on the full protein dataset, but a significant overlap remains. Outliers in the dataset are noted, in keeping with variability in the general population, and more apparent due to the small number of patients in this analysis. For the purposes of our study, outliers were treated as variants within a population arm and retained in the study. Without the inclusion of these outliers, we could not claim that any model generated was applicable to the general patient cohort arm.

Again, only a relatively small proportion of the total variance is explained by the first two principle components suggesting a complicated relationship between protein profile the patients in each arm. By this I mean that the variance in protein expression between patients in the two groups cannot readily be defined by reducing the dataset into just two principle components, requiring multiple components to explain the variance in the dataset. This serves as an indicator that we are unlikely to find a simple statistical model which can perfectly classify between these two patient groups. At this level however, the partial clustering of data shown suggests that it may be possible to develop a model to differentiate the groups based on protein information alone, however this is unlikely to produce a simple or perfect classifying model, and any model is likely to require the inclusion of multiple variables.

3.4.2.2 Protein fold change

Fold change was explored for all proteins between the disease and control groups (Table 3.3) using a Mann-Whitney test to determine significance between the groups, p-values are adjusted using the false discovery rate formula due to the high number of statistical tests involved.



Protein	log2FC	p.value	adj p.value	logp.value.adj
AGER	1.96	2.80E-06	8.00E-04	3.1
NTproBNP	2.42	3.82E-05	0.00546	2.26
COL4A3	0.894	8.81E-05	0.0084	2.08
IL6ST	0.356	0.000157	0.0108	1.97
FIGF	0.87	0.000189	0.0108	1.97
NRP1	0.452	0.00025	0.0119	1.92
IGFBP7	0.617	0.000298	0.0122	1.91
MMP2	0.453	0.000392	0.014	1.85
SOST	0.645	0.000661	0.021	1.68
GDF15	0.749	0.000851	0.0214	1.67
S100B	0.495	0.000921	0.0214	1.67
FN1	0.998	0.000965	0.0214	1.67
CSTA	0.733	0.00107	0.0214	1.67
ANGPT2	1.16	0.00115	0.0214	1.67
ANGPTL4	0.619	0.00115	0.0214	1.67
IL18BP	0.6	0.00127	0.0214	1.67
TNFRSF10C	0.769	0.00128	0.0214	1.67
TIMP2	0.296	0.00135	0.0214	1.67
IGFBP2	0.848	0.00147	0.0221	1.66
FCN3	-0.404	0.00223	0.0319	1.5
VCAM1	0.471	0.00237	0.0322	1.49
TIMP1	0.296	0.00265	0.0341	1.47
PECAM1	0.389	0.00276	0.0341	1.47
COL18A1	0.366	0.00286	0.0341	1.47
IL1R1	0.43	0.00321	0.0367	1.43
ADAMTS8	0.334	0.00347	0.0382	1.42
ICAM1	0.581	0.00403	0.0427	1.37

Table 3.3: Statistically significant fold changes.

Abbreviations: log2FC – Log$_2$ fold change, p.value – unadjusted p-value, adj p.value – adjusted p-value, logp.value.adj – log$_2$ adjusted p-value.

Figure 3.3 show these amongst all other proteins in a volcano plot demonstrating the extent to which the proteins in the dataset are altered between the disease and control group.

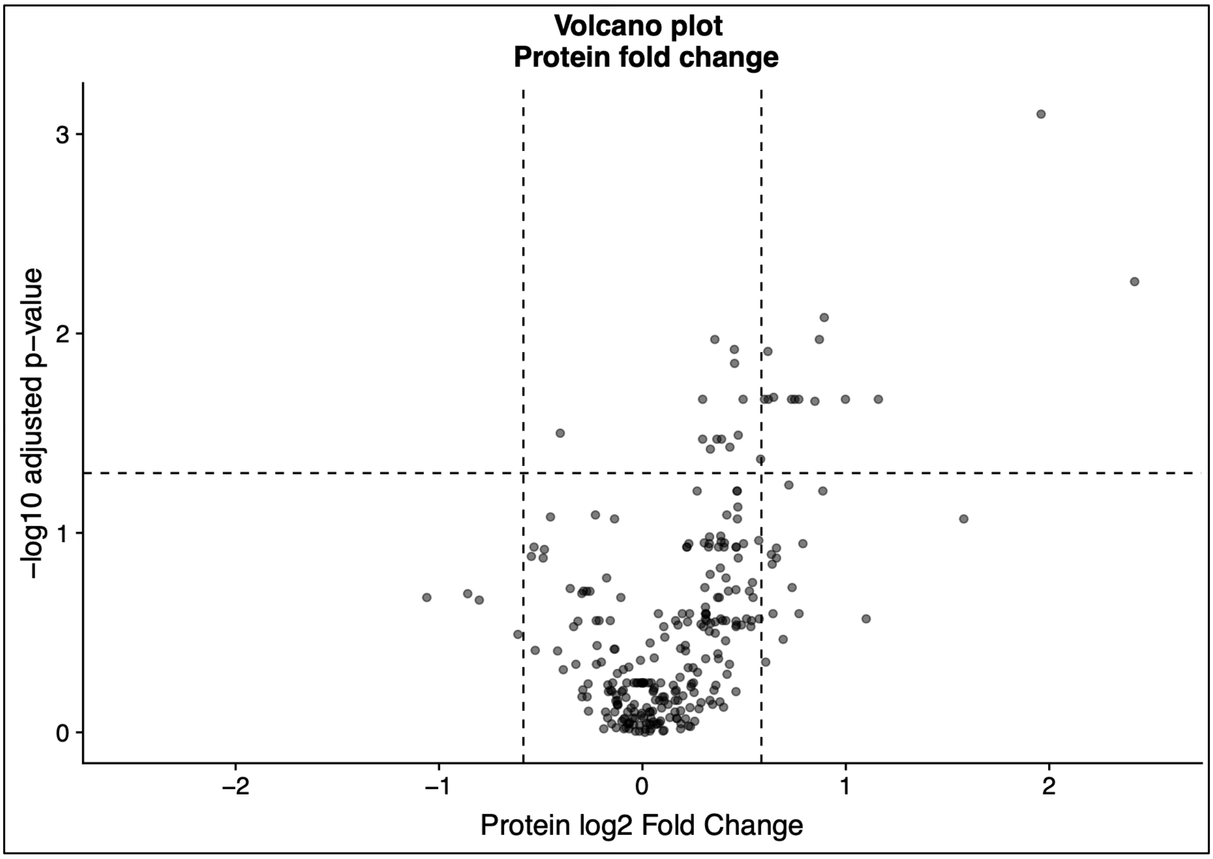


Figure 3.3: Volcano plot for protein fold change: all proteins shown.

All proteins in the dataset included in this analysis. Each point is plotted partially transparent to prevent obscuration of overlapping proteins. Horizontal dotted line corresponds to adjusted p-value of 0.05, those proteins above this demonstrating statistical significance on this inverted axis. Horizontal dotted lines correspond to 1.5 fold change in protein concentration.

3.4.3 Variable pre-processing

296 protein analytes were measured using the Myriad RBM discovery platform. 1 protein variable (HGF) was immediately removed as reported ‘insufficient sample’ and returned no data. For the purposes of classification between SSc with PAH and SSc without PH, the dataset was reduced to these patient groups only (n=22 and 22 respectively) (Table 3.2). Datapoints falling outside the limits of detection were transformed as previously described. After removing variables which had little or no variance (i.e. all datapoints at same value and therefore lacking any useful information) 279 proteins remain. After reducing the dataset further according to those proteins with a univariate classifying AU-ROC >0.7, data for 57 proteins remained. Examining for significant collinearity with a correlation matrix (Figure 3.5)

and variance inflation factors identified 11 proteins with high levels of collinearity, these were considered and removed from further analysis, leaving 46 proteins of interest.

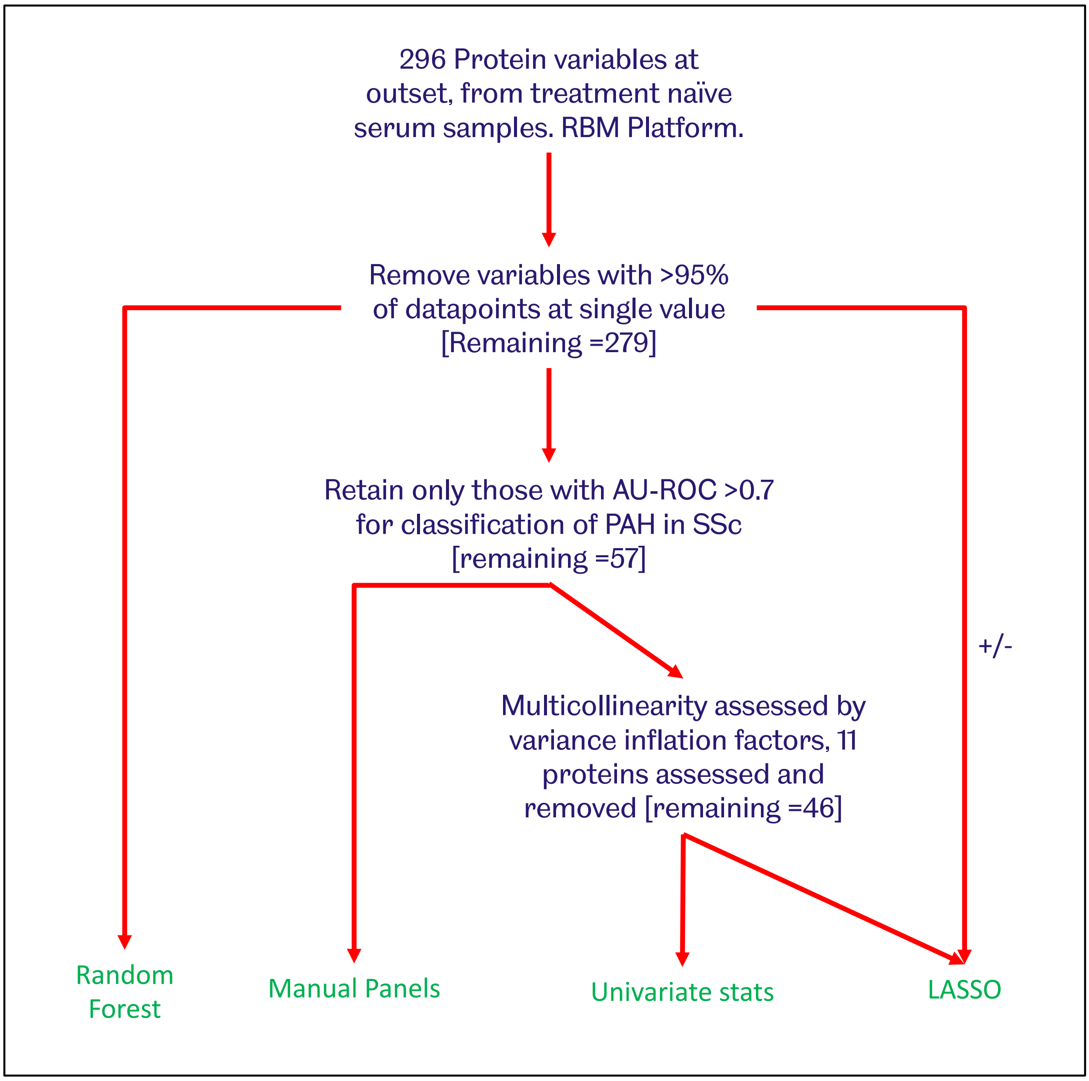

296 Protein variables at outset, from treatment naïve serum samples. RBM Platform.
Remove variables with >95% of datapoints at single value [Remaining =279]
Retain only those with AU-ROC >0.7 for classification of PAH in SSc [remaining =57]
Multicollinearity assessed by variance inflation factors, 11 proteins assessed and removed [remaining =46]
+/-
Random Forest
Manual Panels
Univariate stats
LASSO

Figure 3.4: Data pre-processing flow diagram

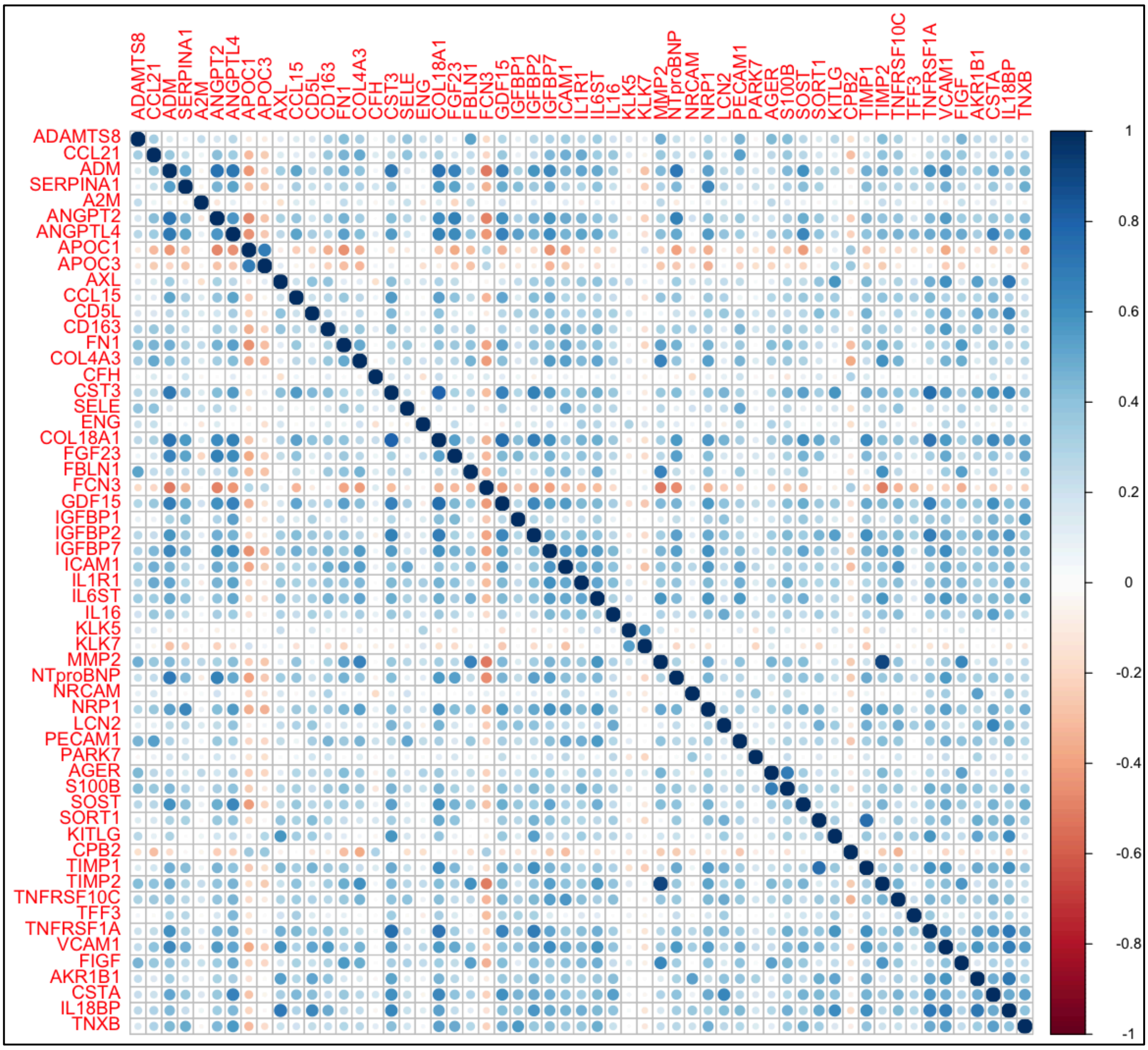
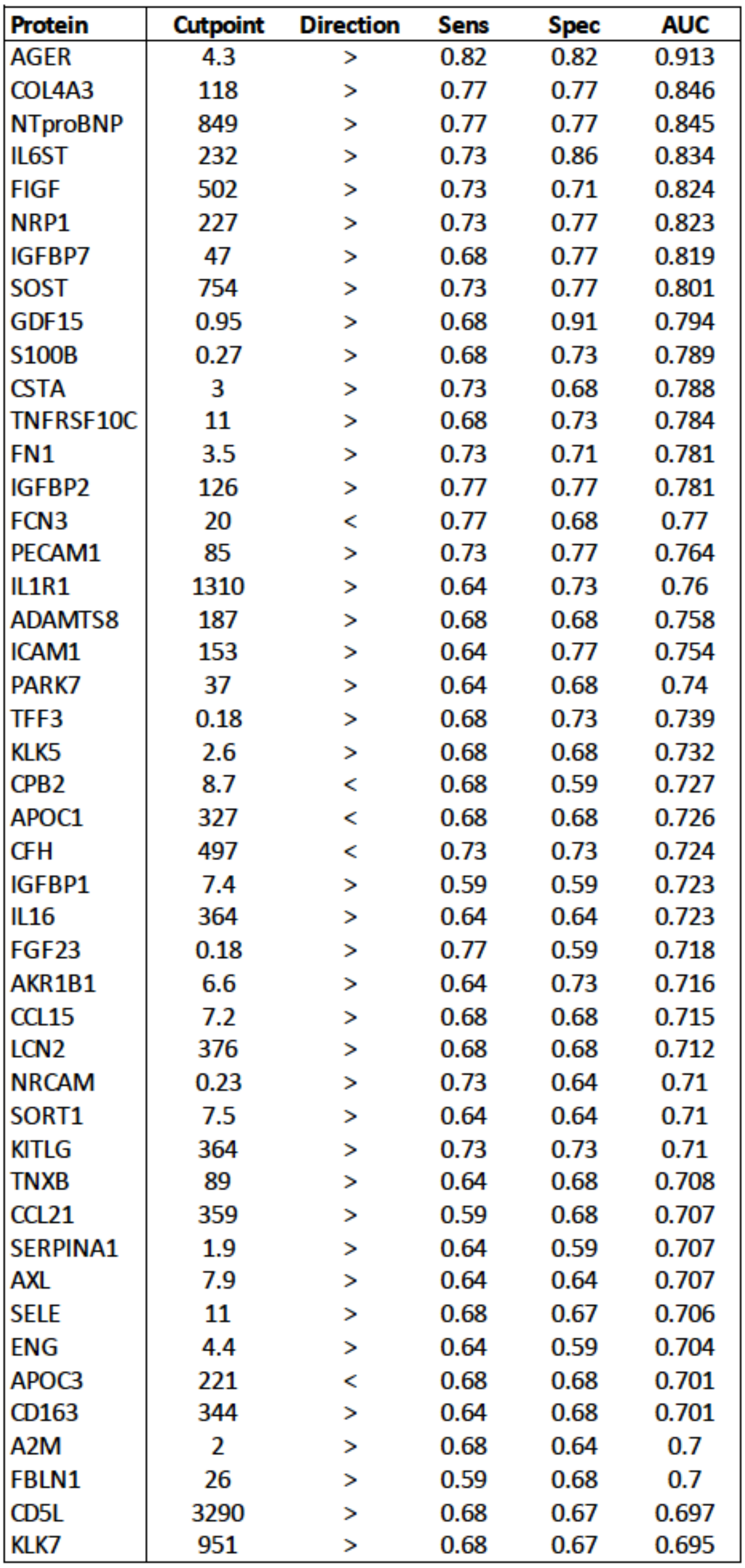


Figure 3.5: Correlation matrix for 57 remaining proteins. Correlation matrix showing Spearman correlations

3.4.4 Univariate classifying statistics

The remaining 46 proteins were assessed individually for their univariate classifying utility (Table 3.4) and ranked according to the AU-ROC for classifying between the disease and control cohort.

Table 3.4: Individual protein thresholds from ROC analysis and univariate diagnostic statistics



Protein	Cutpoint	Direction	Sens	Spec	AUC
AGER	4.3	>	0.82	0.82	0.913
COL4A3	118	>	0.77	0.77	0.846
NTproBNP	849	>	0.77	0.77	0.845
IL6ST	232	>	0.73	0.86	0.834
FIGF	502	>	0.73	0.71	0.824
NRP1	227	>	0.73	0.77	0.823
IGFBP7	47	>	0.68	0.77	0.819
SOST	754	>	0.73	0.77	0.801
GDF15	0.95	>	0.68	0.91	0.794
S100B	0.27	>	0.68	0.73	0.789
CSTA	3	>	0.73	0.68	0.788
TNFRSF10C	11	>	0.68	0.73	0.784
FN1	3.5	>	0.73	0.71	0.781
IGFBP2	126	>	0.77	0.77	0.781
FCN3	20	<	0.77	0.68	0.77
PECAM1	85	>	0.73	0.77	0.764
IL1R1	1310	>	0.64	0.73	0.76
ADAMTS8	187	>	0.68	0.68	0.758
ICAM1	153	>	0.64	0.77	0.754
PARK7	37	>	0.64	0.68	0.74
TFF3	0.18	>	0.68	0.73	0.739
KLK5	2.6	>	0.68	0.68	0.732
CPB2	8.7	<	0.68	0.59	0.727
APOC1	327	<	0.68	0.68	0.726
CFH	497	<	0.73	0.73	0.724
IGFBP1	7.4	>	0.59	0.59	0.723
IL16	364	>	0.64	0.64	0.723
FGF23	0.18	>	0.77	0.59	0.718
AKR1B1	6.6	>	0.64	0.73	0.716
CCL15	7.2	>	0.68	0.68	0.715
LCN2	376	>	0.68	0.68	0.712
NRCAM	0.23	>	0.73	0.64	0.71
SORT1	7.5	>	0.64	0.64	0.71
KITLG	364	>	0.73	0.73	0.71
TNXB	89	>	0.64	0.68	0.708
CCL21	359	>	0.59	0.68	0.707
SERPINA1	1.9	>	0.64	0.59	0.707
AXL	7.9	>	0.64	0.64	0.707
SELE	11	>	0.68	0.67	0.706
ENG	4.4	>	0.64	0.59	0.704
APOC3	221	<	0.68	0.68	0.701
CD163	344	>	0.64	0.68	0.701
A2M	2	>	0.68	0.64	0.7
FBLN1	26	>	0.59	0.68	0.7
CD5L	3290	>	0.68	0.67	0.697
KLK7	951	>	0.68	0.67	0.695

Statistics for individual proteins. Threshold: diagnostic cut-off level at Youden index with value for individual protein. Direction: Demonstrating whether the disease group has increased or suppressed protein expression. AUC: Area under the curve on receiver operating curve. Protein abbreviations: see appendix.

3.4.5 Panels based on univariate statistics

57 proteins were selected for the assessment of their classifying utility in combination panels. As all panels are assessed, and each protein contributes to the score on a univariate basis, collinearity was considered not to compromise this analysis. All possible combinations of between 2 and 5 protein length were generated, totalling 4,612,972 combinations in total, with total uninterrupted processing time at 3.2 days to completion.

Table 3.5: Diagnostic panels (top 20)

Protein Panel	TP	TN	FN	FP	Accuracy	Sens	Spec	PPV	NPV
CD5L|CFH|COL18A1|ICAM1|KITLG	22	20	0	2	0.95	0.91	1.00	1.00	0.92
CFH|IGFBP7|KLK7|PECAM1|TIMP1	21	21	1	1	0.95	0.95	0.95	0.95	0.95
GDF15|IGFBP2|IL6ST|PECAM1|TIMP1	20	22	2	0	0.95	1.00	0.91	0.92	1.00
AGER|CFH|COL18A1|PECAM1|TIMP1	21	21	1	1	0.95	0.95	0.95	0.95	0.95
AGER|APOC3|CD5L|GDF15|ICAM1	21	21	1	1	0.95	0.95	0.95	0.95	0.95
AKR1B1|CFH|ICAM1|SOST|TIMP1	20	22	2	0	0.95	1.00	0.91	0.92	1.00
AGER|CFH|IL6ST|SELE|SORT1	21	21	1	1	0.95	0.95	0.95	0.95	0.95
CD5L|FN1|IL6ST|KITLG|PARK7	20	22	2	0	0.95	1.00	0.91	0.92	1.00
AGER|AXL|CFH|COL18A1|PECAM1	20	22	2	0	0.95	1.00	0.91	0.92	1.00
CD5L|ICAM1|NTproBNP|PARK7|SELE	21	21	1	1	0.95	0.95	0.95	0.95	0.95
AKR1B1|CD5L|CFH|ICAM1|KITLG	20	22	2	0	0.95	1.00	0.91	0.92	1.00
AGER|APOC3|FN1|ICAM1|SOST	21	21	1	1	0.95	0.95	0.95	0.95	0.95
CD5L|FN1|IL6ST|PARK7|TNFRSF10C	20	22	2	0	0.95	1.00	0.91	0.92	1.00
AGER|CD5L|ICAM1|IGFBP7|PECAM1	21	21	1	1	0.95	0.95	0.95	0.95	0.95
AGER|APOC3|ICAM1|SELE|SOST	21	21	1	1	0.95	0.95	0.95	0.95	0.95
AXL|CD5L|GDF15|PARK7|PECAM1	20	22	2	0	0.95	1.00	0.91	0.92	1.00
COL18A1|FCN3|IL6ST|PARK7|PECAM1	20	22	2	0	0.95	1.00	0.91	0.92	1.00
AGER|CD5L|ICAM1|NTproBNP|PECAM1	21	21	1	1	0.95	0.95	0.95	0.95	0.95
AGER|AXL|NRCAM|PARK7|PECAM1	20	22	2	0	0.95	1.00	0.91	0.92	1.00
AGER|CFH|FN1|IL6ST|TIMP1	21	21	1	1	0.95	0.95	0.95	0.95	0.95

Top 20 panels (based on diagnostic accuracy) from total 4,612,972 panels generated, with diagnostic statistics. TP: True positive; TN: True negative; FP: False positive; FN: False negative; Sens: Sensitivity; Spec: Specificity; PPV: Positive predictive value; NPV: Negative predictive value.

Table 3.6: Diagnostic panels (bottom 20)

Protein Panel	TP	TN	FN	FP	Accuracy	Sens	Spec	PPV	NPV
CCL21|ENG|FBLN1|IL16|SERPINA1	14	12	8	10	0.59	0.55	0.64	0.60	0.58
A2M|CPB2|ENG|IL1R1|KLK7	13	13	9	9	0.59	0.59	0.59	0.59	0.59
CD163|FBLN1|IGFBP1	12	14	10	8	0.59	0.64	0.55	0.58	0.60
APOC1|CCL21|ENG|KLK5|SERPINA1	13	13	9	9	0.59	0.59	0.59	0.59	0.59
CPB2|IGFBP1	8	18	14	4	0.59	0.82	0.36	0.56	0.67
ENG|FGF23|IL16|NRCAM|SERPINA1	15	11	7	11	0.59	0.50	0.68	0.61	0.58
APOC3|ENG|ICAM1|IL16|SERPINA1	13	13	9	9	0.59	0.59	0.59	0.59	0.59
CCL21|CPB2|ENG|ICAM1|SERPINA1	14	12	8	10	0.59	0.55	0.64	0.60	0.58
A2M|APOC3|CCL21|CPB2|IGFBP1	13	13	9	9	0.59	0.59	0.59	0.59	0.59
CD5L|FBLN1|IL16|SELE|SERPINA1	12	14	10	8	0.59	0.64	0.55	0.58	0.60
ADM|AXL|CPB2	14	12	8	10	0.59	0.55	0.64	0.60	0.58
ICAM1|IL16|NRCAM|SERPINA1|TNFRSF1A	12	14	10	8	0.59	0.64	0.55	0.58	0.60
CPB2|ENG|IGFBP1|IL16|VCAM1	14	12	8	10	0.59	0.55	0.64	0.60	0.58
FGF23|IL16|KLK5|SERPINA1	12	14	10	8	0.59	0.64	0.55	0.58	0.60
APOC1|ENG|IL16|SERPINA1|VCAM1	13	12	9	10	0.57	0.55	0.59	0.57	0.57
ENG|FGF23|SERPINA1	15	10	7	12	0.57	0.45	0.68	0.59	0.56
CCL21|ENG|IL16|SERPINA1|VCAM1	14	11	8	11	0.57	0.50	0.64	0.58	0.56
APOC1|CD163|CPB2|ENG|SERPINA1	14	11	8	11	0.57	0.50	0.64	0.58	0.56
APOC1|CD163|CPB2|IL16|SERPINA1	13	12	9	10	0.57	0.55	0.59	0.57	0.57
CCL21|ENG|ICAM1|IL16|SERPINA1	13	12	9	10	0.57	0.55	0.59	0.57	0.57

Lowest 20 panels (based on diagnostic accuracy) from total 4,612,972 panels generated. TP: True positive; TN: True negative; FP: False positive; FN: False negative; Sens: Sensitivity; Spec: Specificity; PPV: Positive predictive value; NPV: Negative predictive value.

The top panels all show a diagnostic accuracy of 95%, with the lowest at 57%, however these panels are generated based on the thresholds in the derivation dataset and are therefore likely to be poorly applicable to the wider population. Processing time on this large number of combinations not sustainable for further consideration in the analysis, apart from to acknowledge the proteins included in the strongest panels during later variable selection.

3.4.6 Random forest modelling

The full dataset was reduced for entry into random forest modelling by selecting only patients with SSc-PAH and SSc-no PH, and reducing the input dataset to remove proteins which lacked variance. The input dataset was therefore composed of 279 proteins and 44 patients. Variable importance data was extracted from the model (Figure 3.6) to determine diagnostic potential and utility of individual proteins within the model.

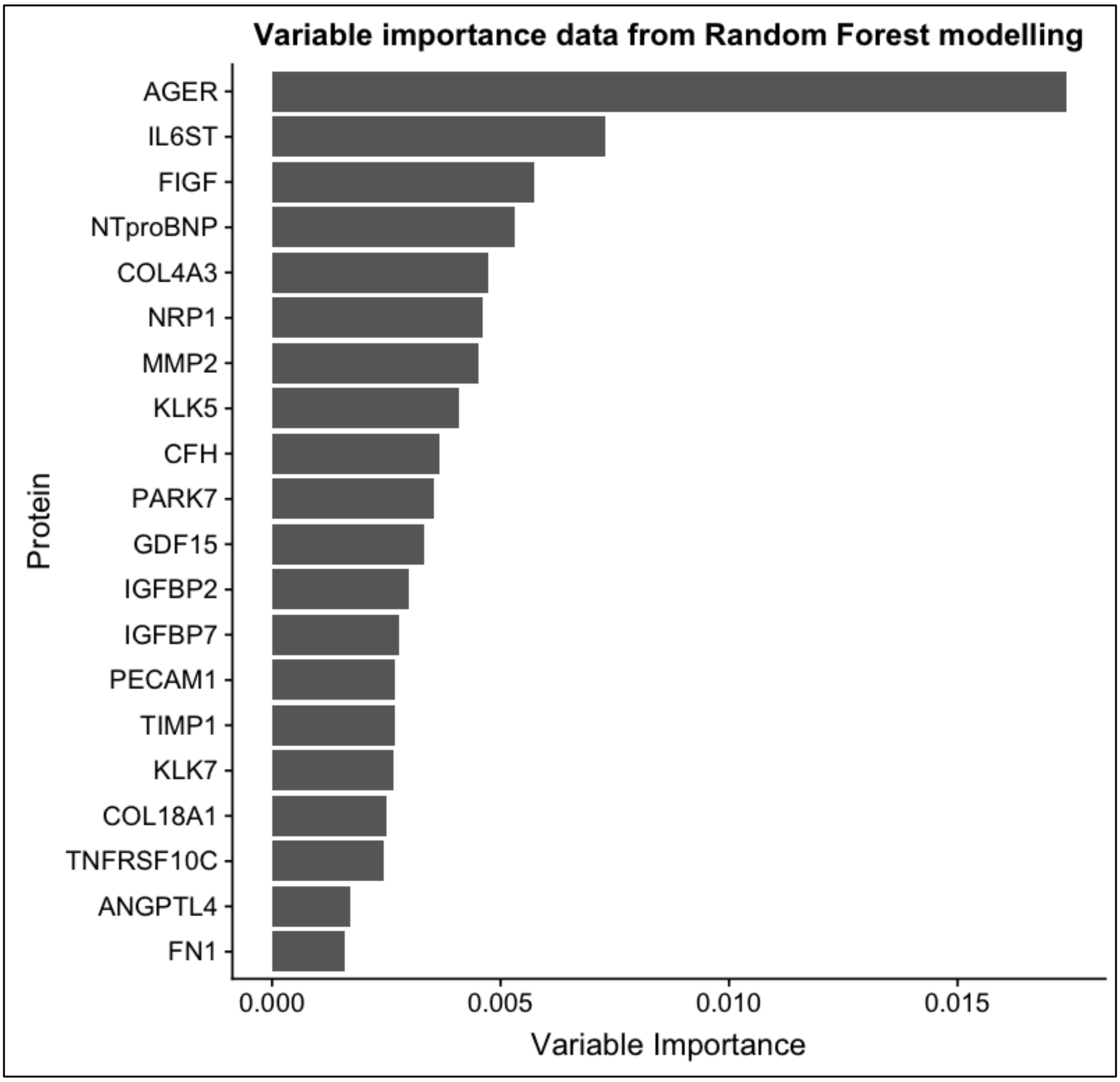


Figure 3.6: Variable importance for classification from Random Forest (SSc-PAH vs SSc-no PH). Graph demonstrates calculated 'importance' of independent variables for univariate classification of patients between the two groups. All variables receive a variable importance, top 20 shown. Abbreviations: see appendix.

AGER returned the most significant variable importance for classification, followed by IL6ST and NT-proBNP.

3.4.7 Least Absolute Shrinkage and Selection Operator (LASSO) modelling

The remaining 46 candidate proteins (Table 3.4) were entered into LASSO modelling. Diagnostic classification was the dependent variable, and protein measurements the independent variables in the training dataset. Cross validation was used to determine the appropriate value for the penalty which yielded the lowest model error, and this penalty was then applied to the dataset using a LASSO model optimised for binomial classification. The selected variables and respective coefficients are shown diagrammatically in Figure 3.7.

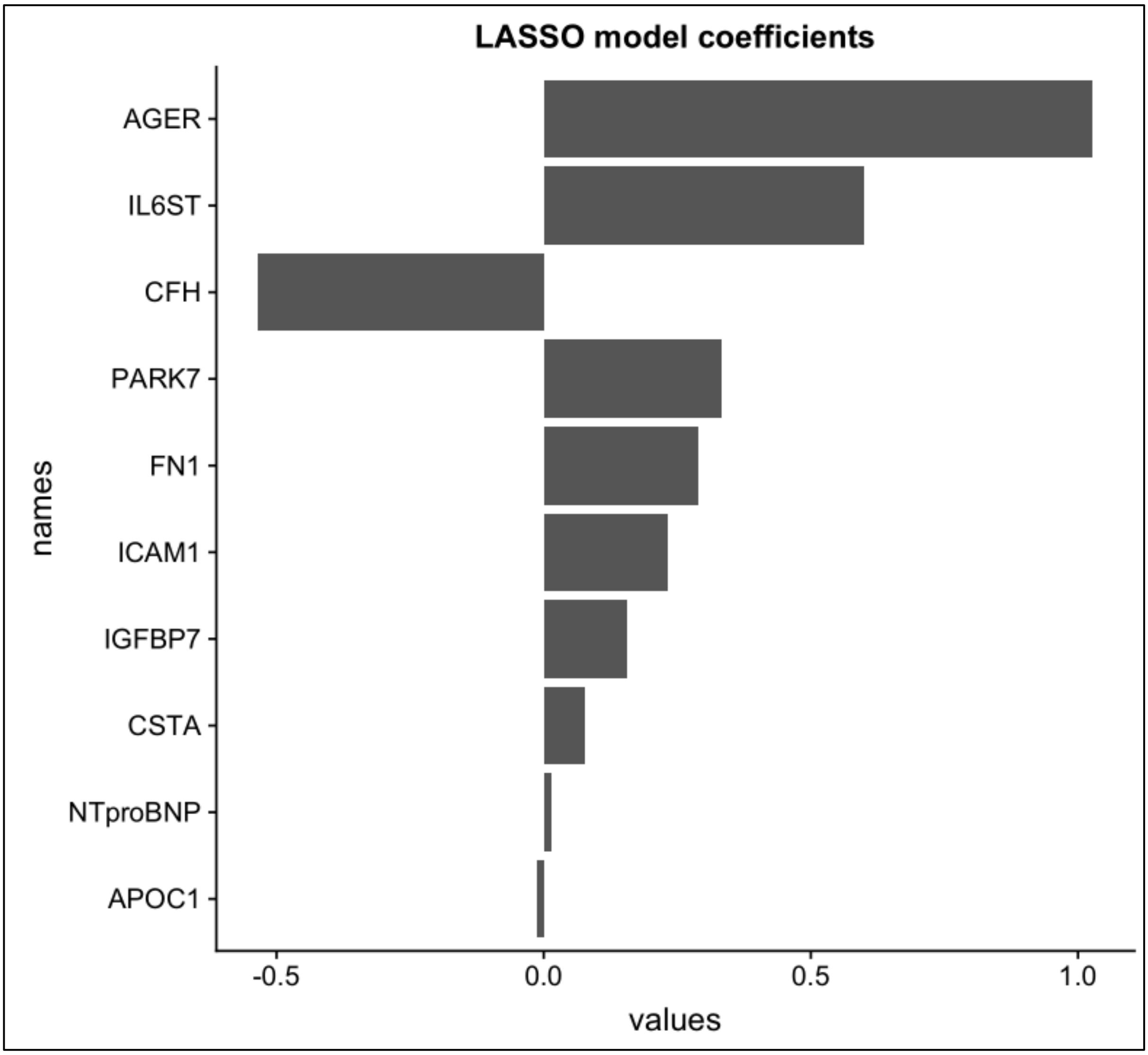


Figure 3.7: Output from LASSO analysis for classification (SSc-PAH vs SSc-no PH)

Bars represent the value of coefficients in the regression model.

LASSO modelling retained 10 proteins in the final model which can be interpreted in the same was as a linear regression formula, with each of the coefficients a multiplier of the corresponding protein concentration.

3.4.8 Model performance

The LASSO model is considered a relatively open model, easily applicable to new data without the need for further complex computing. Applying the penalty term during the development of the model is designed to return a model which can generalise to the true population rather than a model overfit to the derivation dataset only. I therefore considered the LASSO model as the optimal model to adopt as my final classifying model to take forward and to derive performance data against the derivation dataset.

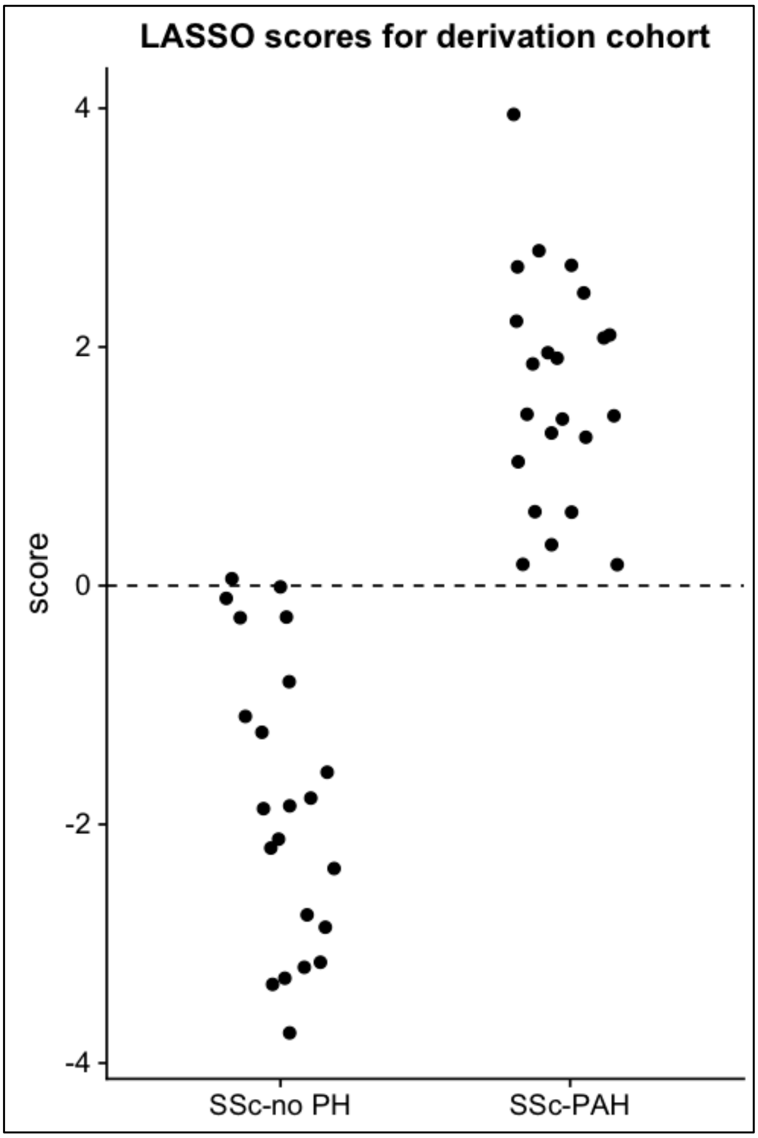


Figure 3.8: LASSO model score for each patient

Jitter plot showing the LASSO models score for each patient, shown by known patient classification.

Figure 3.8 shows the score for each patient in the derivation cohort when the LASSO formula is applied to the known protein concentrations. With the cut-off for classification between the two groups set at 0, we found 1 false positive and no false negatives, giving a model sensitivity 0.95, specificity 1, positive predictive value 1, negative predictive value 0.96, and diagnostic accuracy 98%. The composite ROC curves for both the final model and its individual protein variables (Figure 3.9) demonstrates how the multivariable model outperforms its univariate contributors.

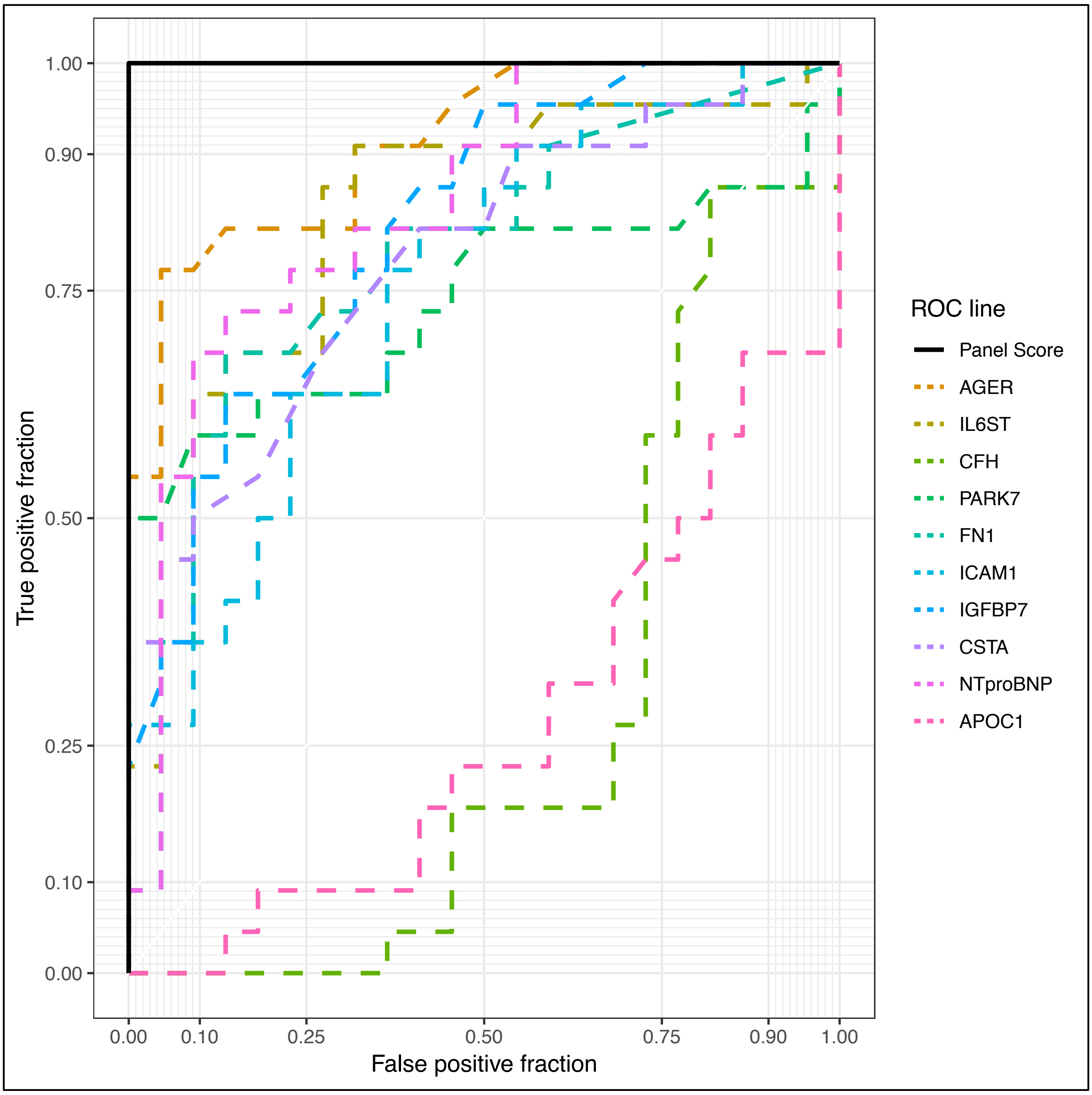


Figure 3.9: Composite ROC curves for individual variables and combined panel. The combined protein panel (solid line) demonstrates superior diagnostic potential to any of its individual proteins (dotted lines)

3.4.9 Comparison with analysis by Actelion pharmaceutical

A similar but independent analysis was conducted on the same dataset by the bioinformatic department of our collaborator Actelion pharmaceutical as part of their separate work. Their statistical preparation and exact methodology were not available to review. Actelion pharmaceutical opted to use random forest as their primary modelling tool, and shared the following result, showing variable importance data, based on what we understand to be the same patient cohort (Figure 3.10).

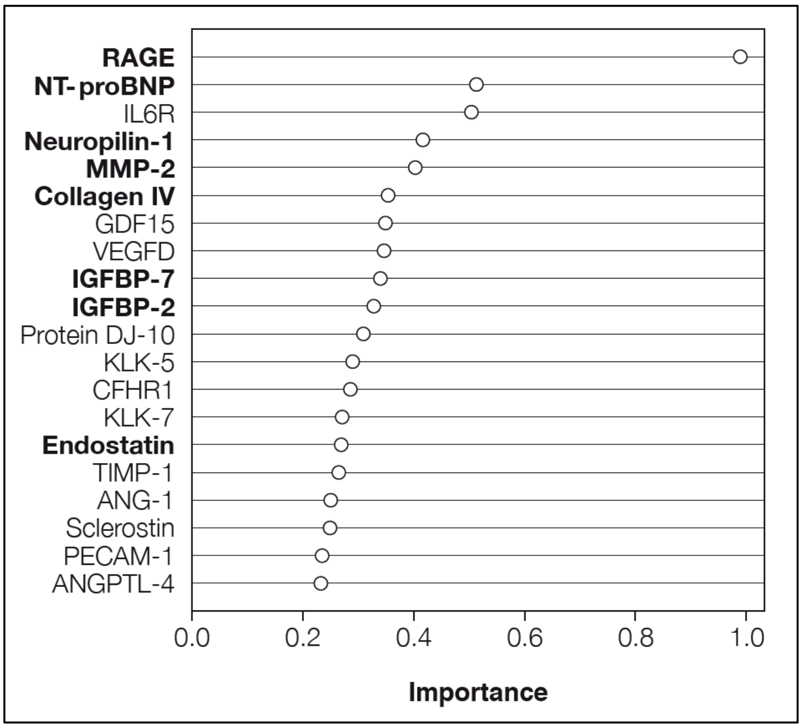


Figure 3.10: Random forest variable importance from Actelion

Data from analysis done by Actelion pharmaceutical. Proteins in bold hold no relevance to our work.

This work helps to validate some of the methodology done in my analysis as the protein identified here also feature in similar order and magnitude of variable importance in my random forest analysis. For the purposes of comparison with my random forest result data in Figure 3.6, the protein descriptors used by Actelion pharmaceutical differ from my own: RAGE = AGER; IL6R = IL6ST; Neuropilin-1 = NRP1; Collagen IV = COL4A3; VEGFD = FIGF; Protein DJ-10 (typing error, should be: Protein DJ-1) = PARK7; CFH-R1 = CFH; Endostatin = COL18A1.

3.4.10 Exploring Influence of Comorbidities

Noting the imbalance the in the comorbidity of interstitial lung disease between the disease and control arms, a key concern was whether this panel, and the chosen protein biomarkers therein, were correctly targeting and classifying the PAH in patients with SSc, or whether the signal was confounded by the imbalance in ILD and classifying patients based on the detection of proteins relevant to this. The imbalance in proportion of ILD in each arm was caused by the careful exclusion of patients with ILD into the SSc-PAH arm of the initial study cohorts (Table 3.1). Similar exclusion was not made for inclusion of patients into the SSc-no PH arm.

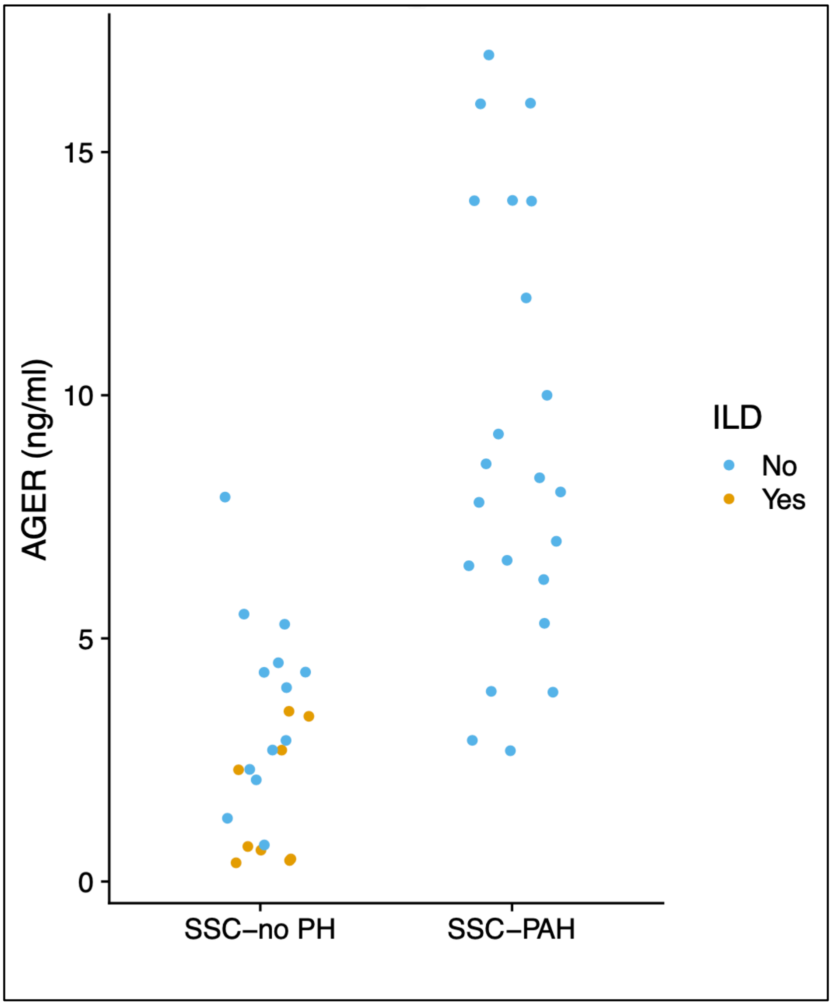


Figure 3.11: Distribution serum AGER concentration labelled for the presence or absence of ILD

Jitter plot showing serum AGER concentrations, grouped by known diagnostic subgroup, and coloured by the presence or absence of ILD

As AGER is by far the strongest predictor from all classifying models explored, and has significantly higher influence in the final model than any other individual protein, I chose to investigate whether this protein is altered in patients with interstitial lung disease. Figure 3.11 shows a simple jitter plot for the concentration of AGER between patients with SSc both with and without PAH, colour identifying patients with evidence of ILD. Our analysis to date suggests that the expression of AGER is significantly increased in patients with SSc-PAH compared to SSc-no PH controls, which this data would support. It is also clear however that the patients with ILD cluster tightly towards the lower end of the protein range, clearly demonstrating that expression of AGER is influenced by the presence of ILD, and may in fact be suppressed in the patients with ILD creating an apparent, but inaccurate, increase in signal in the presence of PAH.

On the basis of this sub-analysis showing that the classifying signal provided by AGER is confounded by the presence of ILD, and the significant influence that AGER has in the classifying models derived to this point, I decided to reassess the test cohorts and reanalyse the dataset.

3.5 Results: Analysis Part 2

Understanding the significant influence of a dataset unbalanced for comorbidities on the outcome of a classifying model I sought to replicate the previous analysis using an updated and more balanced dataset for classification modelling. The error in our initial analysis arose from the decision to model a 'clean' disease cohort, but to allow a cohort mixed for the presence of lung disease for a control cohort which directly influenced the variables selected. Data from a large multinational European database of patients diagnosed with SSc shows that between 34.7% and 53.4% of patients with SSc will have a degree of interstitial lung disease (Walker et al., 2007). The proportion of patients with ILD in our control cohort (SSc-no PH) falls within this range (40.9%). Combining the clean SSc-PAH disease cohort previously analysed with that of the SSc-PAH-ILD cohort shown in Table 3.1 (patients considered to have, and treated as, group 1 PAH with some ILD rather than group 3 PH) yields a disease cohort which is now more balanced for degree of ILD and other recorded comorbidities (Table 3.7). As a screening tool, my classifying model should ideally be applicable to an unselected group of patients with SSc to screen for the presence PAH and as such should be robust to the inclusion of a realistic proportion of patients with ILD.

3.5.1 Demographics

Table 3.7: Demographics for Myriad1b analysis

		SSC-PAH	SSC-no PH	p-value
n		38	22	
Age (IQ range)		69 (63-72)	61.5 (56.3-67.8)	0.03
Gender (M/F)		12/26	2/20	0.1
Deaths		26	4	<0.001
WHO FC	**1**	0	1	0.006
	2	5	8	
	3	31	13	
	4	2	0	
Co-Morbidity	**COPD (%)**	5 (13.2)	1 (4.6)	0.53
	Haemolysis (%)	0 (0)	0 (0)	NA
	Myeloproliferative (%)	1 (2.6)	0 (0)	1
	AF (%)	2 (5.3)	1 (4.6)	1
	A Flutter (%)	0 (0)	0 (0)	NA
	ILD (%)	16 (42.1)	9 (40.9)	1
	Asthma (%)	1 (2.6)	1 (4.6)	1
	Sarcoidosis (%)	0 (0)	1 (4.6)	0.78
	OSA (%)	2 (5.3)	0 (0)	0.73
	VTE (%)	2 (5.3)	2 (9.1)	0.97
PFTs (median + IQ))	**FEV1 Percent**	84.2 (68.3-91.3)	77.3 (68.6-101.6)	0.98
	FVC Percent	93.1 (76.8-107.6)	94.7 (73.4-107.9)	0.99
	TLCO Percent	41.4 (30.1-45.3)	55.1 (46.7-67.3)	<0.001
ISWT (median + IQ)	**Distance**	120 (72.5-220)	210 (75-338)	0.16
RHC (median + IQ)	**mPAP**	35 (29.3-44.8)	21 (19-22)	<0.001
	RA pressure	8 (4-11)	5 (3.3-4.1)	0.012
	CI	2.95 (2.43-3.4)	3.4 (3.1-4.1)	0.003
	PVR	380 (271-600)	140 (117-171)	<0.001
	PCWP	11 (8-14)	8.5 (7-11.8)	0.15

Abbreviations: n – number; IQ – interquartile; M/F – Male/Female; WHO FC – World health organisation functional class; COPD – chronic obstructive pulmonary disease; AF – Atrial fibrillation; A Flutter – Atrial flutter; ILD – interstitial lung disease; OSA – Obstructive sleep apnoea; VTE – Venous thromboembolism; PFT – Pulmonary function tests; ISWT – incremental shuttle walking test; RHC – right heart catheter

The disease and control cohorts are balanced for interstitial lung disease and the included prevalence is consistent with published data on the prevalence of ILD in SSc. Other comorbidities are reasonably balanced. The disease cohort is slightly older than the control, and other expected clinical characteristics are unbalanced, but in keeping with the phenotypical effects of PAH.

3.5.2 Principle Component Analysis

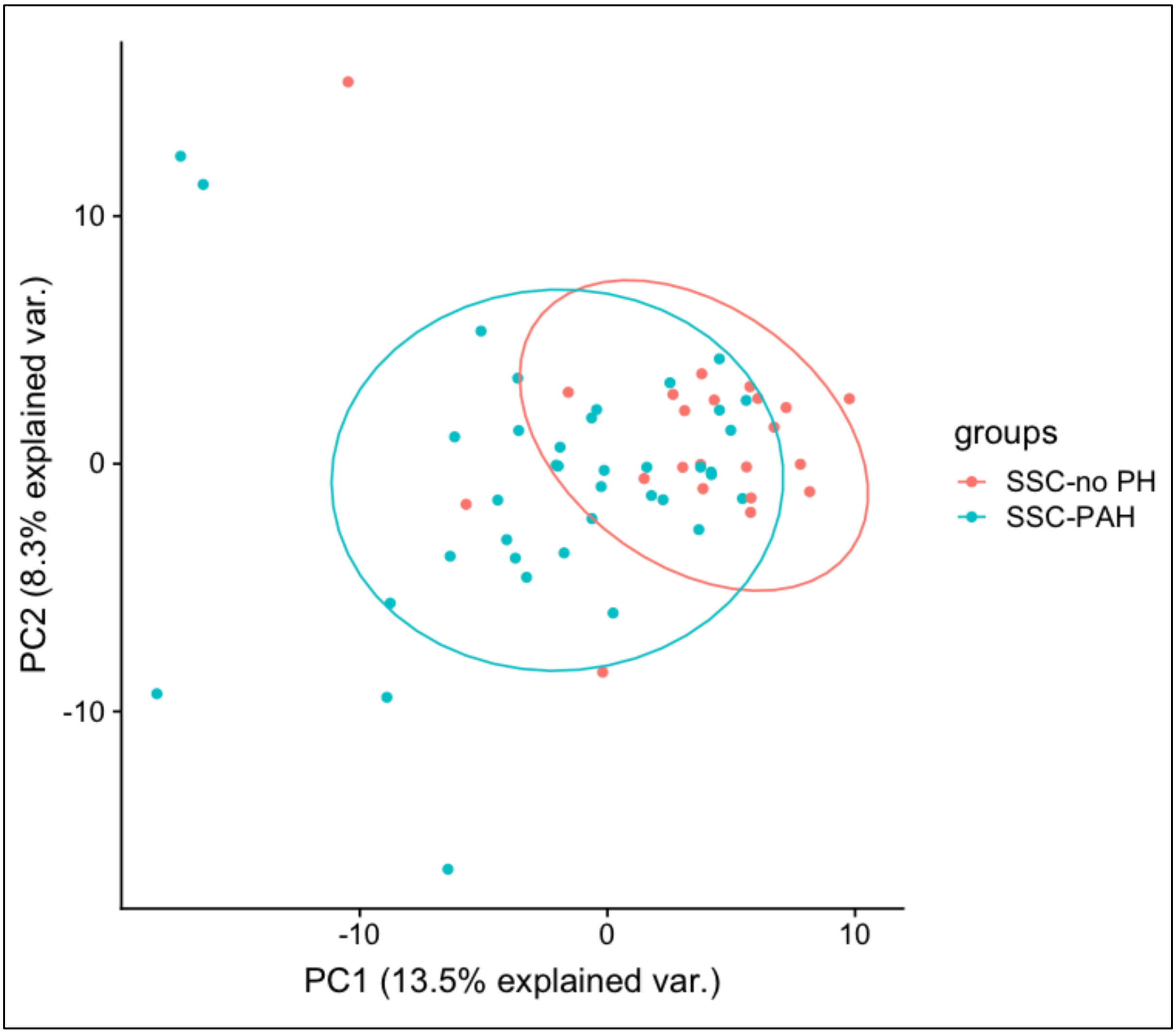


Figure 3.12: PCA of SSc-PAH and SSc-no PH

Abbreviations: PC1 – principle component 1, PC2 – principle component 2, var. – variance.

PCA of this cohort, using all available protein information in the dataset, reveals similar results to that in analysis 1. Only 21.8% of the variance is explained by the first two principle components demonstrating the complexity of the relationship between the two groups.

3.5.3 Variable pre-processing

HGF was removed immediately, and the dataset reduced to only the SSc-PAH (n=38) and SSc-no PH (n=22) cohorts. Datapoints falling outside the limits of detection were transformed, and variables lacking variance removed leaving 281 proteins. After retaining only proteins with AU-ROC >0.7 for univariate classifying potential, 31 protein variables remain. Finally, analysing for collinearity using variance inflation factors found 1 further variable for removal from the dataset leaving 30 proteins. Statistics were performed on the dataset using the relevant subsets of variables as shown in Figure 3.13.

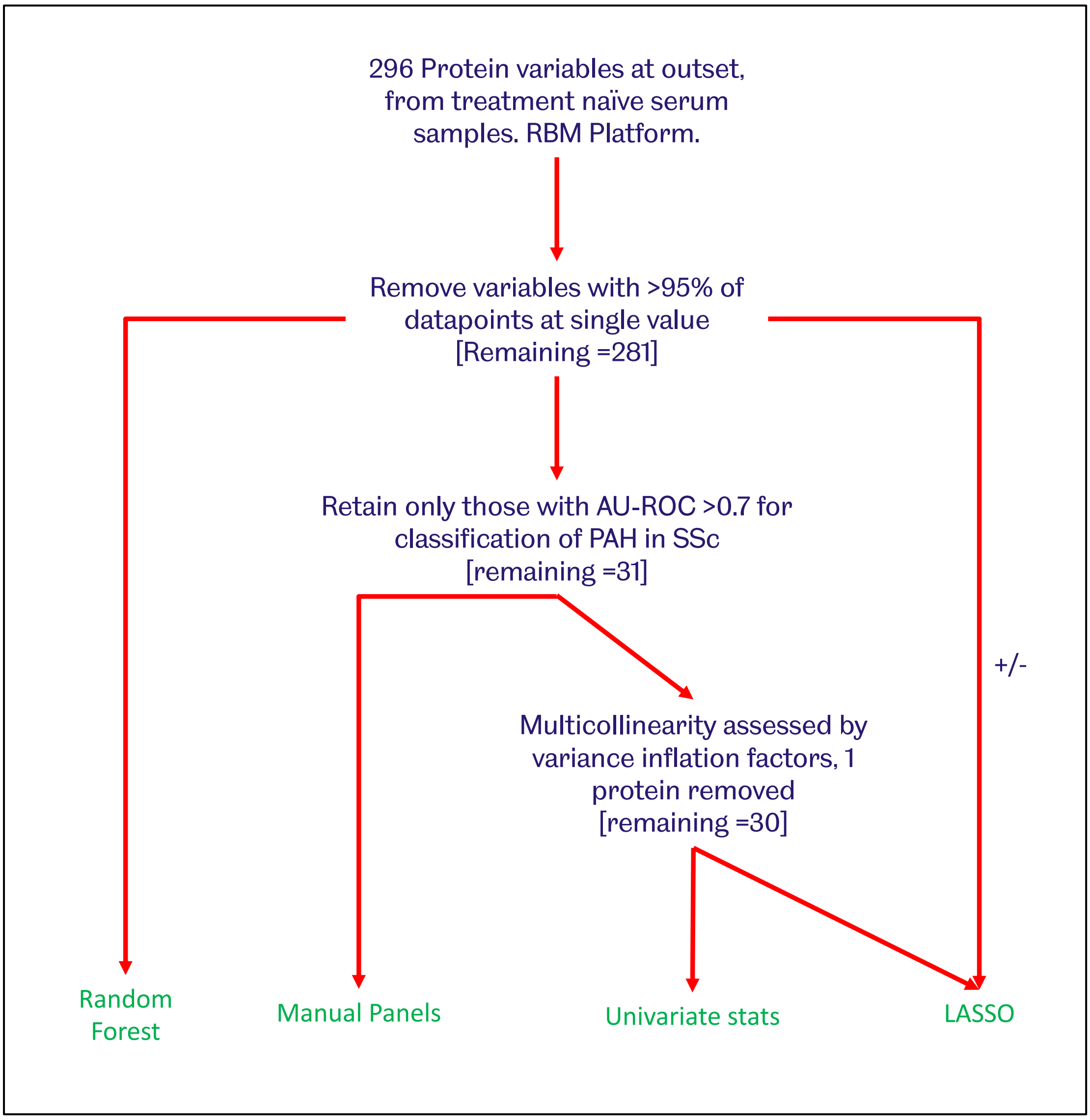

296 Protein variables at outset, from treatment naïve serum samples. RBM Platform.
Remove variables with >95% of datapoints at single value [Remaining =281]
Retain only those with AU-ROC >0.7 for classification of PAH in SSc [remaining =31]
Multicollinearity assessed by variance inflation factors, 1 protein removed [remaining =30]
+/-
Random Forest
Manual Panels
Univariate stats
LASSO

Figure 3.13: Data pre-processing flow diagram

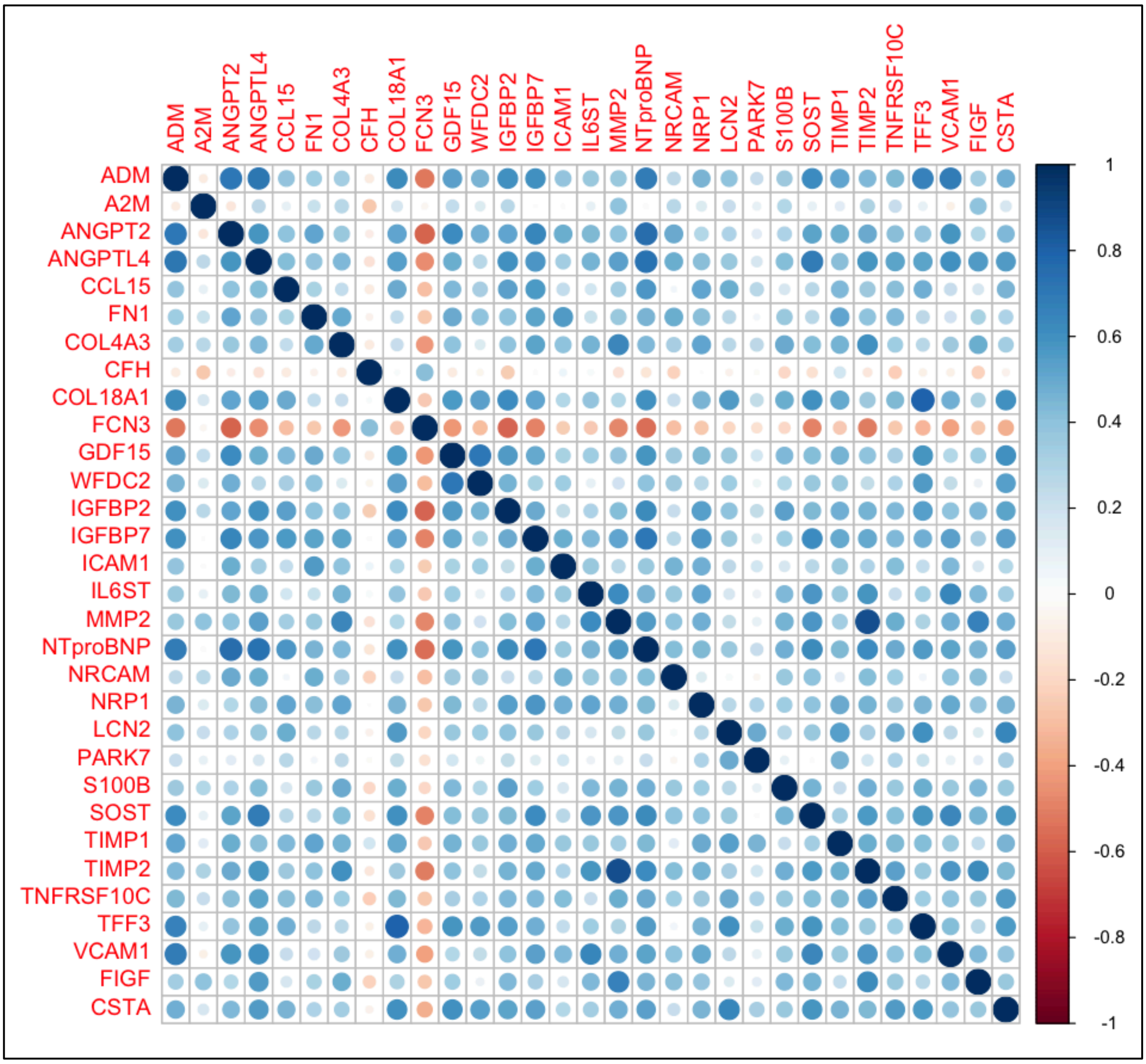


Figure 3.14: Correlation matrix for remaining 31 proteins.

Showing Spearman coefficients.

MMP2 was the most highly correlated variable and was removed from the dataset due to a strong correlation with TIMP2.

3.5.4 Univariate classifying statistics

The remaining 31 proteins were analysed and ranked according to their univariate classifying ability, using AU-ROC as the metric of choice.

Table 3.8:Individual protein thresholds from ROC analysis and univariate diagnostic statistics

Protein	Cutpoint	Direction	Sensitivity	Specificity	AU-ROC
GDF15	1	>	0.73	0.76	0.822
NTproBNP	659	>	0.73	0.71	0.804
FCN3	20	<	0.77	0.71	0.798
MMP2	1740	>	0.68	0.87	0.779
FN1	3.5	>	0.73	0.73	0.775
ANGPT2	5	>	0.73	0.68	0.772
COL4A3	104	>	0.64	0.74	0.772
IGFBP7	47	>	0.68	0.66	0.769
NRP1	227	>	0.73	0.71	0.765
TIMP1	160	>	0.73	0.68	0.765
CSTA	3	>	0.73	0.66	0.763
IGFBP2	126	>	0.77	0.76	0.76
TIMP2	85	>	0.73	0.68	0.76
IL6ST	232	>	0.73	0.71	0.753
TNFRSF10C	11	>	0.68	0.68	0.749
SOST	696	>	0.68	0.74	0.748
CFH	497	<	0.73	0.71	0.747
LCN2	376	>	0.68	0.71	0.735
ICAM1	153	>	0.64	0.74	0.73
ANGPTL4	147	>	0.73	0.68	0.728
CCL15	6.8	>	0.64	0.71	0.728
ADM	3.6	>	0.68	0.66	0.724
TFF3	0.16	>	0.64	0.66	0.721
COL18A1	82	>	0.73	0.68	0.72
WFDC2	1260	>	0.64	0.65	0.719
FIGF	475	>	0.64	0.65	0.717
S100B	0.27	>	0.68	0.61	0.713
NRCAM	0.23	>	0.73	0.63	0.711
VCAM1	886	>	0.64	0.66	0.71
A2M	2	>	0.68	0.66	0.709
PARK7	37	>	0.64	0.66	0.703

Statistics for individual proteins. Threshold: diagnostic cut-off level at Youden index with value for individual protein. Direction: Demonstrating whether the disease group has increased or suppressed protein expression. AUC: Area under the curve on receiver operating curve. Protein abbreviations: see appendix.

3.5.5 Panels based on univariate statistics

The univariate statistics calculated in Table 3.8 were used to compute all possible combinations of multivariable panels using the method described in methods section 2.4.4.1. Due to the high processor requirement this was done on the higher power Mac Pro. All combination of between 2 and 5 protein length were assessed, totalling 206,337 combinations, with total processing time at 5.3mins.

Table 3.9: Diagnostic panels (top 20)

Protein Panel	TP	TN	FN	FP	Accuracy	Sens	Spec	PPV	NPV
CFH|COL18A1|CSTA|GDF15|MMP2	17	38	5	0	0.92	1.00	0.77	0.88	1.00
GDF15|NRP1|PARK7|SOST|TIMP2	17	38	5	0	0.92	1.00	0.77	0.88	1.00
ANGPTL4|GDF15|NRP1|PARK7|TIMP2	17	38	5	0	0.92	1.00	0.77	0.88	1.00
GDF15|IGFBP2|LCN2|NRP1|TIMP2	17	38	5	0	0.92	1.00	0.77	0.88	1.00
FCN3|GDF15|IL6ST|NRP1|TIMP2	17	38	5	0	0.92	1.00	0.77	0.88	1.00
GDF15|IGFBP2|NRP1|PARK7|TIMP2	17	38	5	0	0.92	1.00	0.77	0.88	1.00
CFH|COL18A1|FCN3|MMP2|PARK7	20	35	2	3	0.92	0.92	0.91	0.95	0.87
IGFBP2|IL6ST|NRP1|NTproBNP|TIMP2	17	37	5	1	0.90	0.97	0.77	0.88	0.94
CFH|NRP1|NTproBNP|SOST|TIMP2	16	38	6	0	0.90	1.00	0.73	0.86	1.00
CFH|GDF15|MMP2|NTproBNP|TIMP1	16	38	6	0	0.90	1.00	0.73	0.86	1.00
CFH|IL6ST|NRP1|NTproBNP	18	36	4	2	0.90	0.95	0.82	0.90	0.90
CFH|FN1|ICAM1|MMP2|VCAM1	19	35	3	3	0.90	0.92	0.86	0.92	0.86
FCN3|GDF15|NRP1|SOST|TIMP2	17	37	5	1	0.90	0.97	0.77	0.88	0.94
GDF15|NRP1|PARK7|TIMP2	17	37	5	1	0.90	0.97	0.77	0.88	0.94
CFH|FN1|IL6ST|NRP1|PARK7	20	34	2	4	0.90	0.89	0.91	0.94	0.83
A2M|CFH|FN1|ICAM1|VCAM1	20	34	2	4	0.90	0.89	0.91	0.94	0.83
CFH|GDF15|MMP2|SOST	16	38	6	0	0.90	1.00	0.73	0.86	1.00
GDF15|LCN2|NRP1|NTproBNP|TIMP1	16	38	6	0	0.90	1.00	0.73	0.86	1.00
IGFBP2|MMP2|PARK7|SOST|TIMP2	17	37	5	1	0.90	0.97	0.77	0.88	0.94
GDF15|LCN2|NRP1|TIMP2	17	37	5	1	0.90	0.97	0.77	0.88	0.94

Top 20 panels (based on diagnostic accuracy) from total 206,337 panels generated, with diagnostic statistics. TP: True positive; TN: True negative; FP: False positive; FN: False negative; Sens: Sensitivity; Spec: Specificity; PPV: Positive predictive value; NPV: Negative predictive value.

Table 3.10: Diagnostic panels (bottom 20)

Protein Panel	TP	TN	FN	FP	Accuracy	Sens	Spec	PPV	NPV
FIGF|VCAM1	19	18	3	20	0.62	0.47	0.86	0.86	0.49
LCN2|WFDC2	20	17	2	21	0.62	0.45	0.91	0.89	0.49
FN1|PARK7	22	15	0	23	0.62	0.39	1	1	0.49
CSTA|FIGF|NRCAM|PARK7|WFDC2	15	22	7	16	0.62	0.58	0.68	0.76	0.48
NRCAM|TFF3	21	16	1	22	0.62	0.42	0.95	0.94	0.49
FIGF|ICAM1	19	18	3	20	0.62	0.47	0.86	0.86	0.49
CCL15|TFF3	17	20	5	18	0.62	0.53	0.77	0.8	0.49
COL4A3|TNFRSF10C	19	18	3	20	0.62	0.47	0.86	0.86	0.49
CSTA|FIGF|WFDC2	15	22	7	16	0.62	0.58	0.68	0.76	0.48
CCL15|COL4A3	17	20	5	18	0.62	0.53	0.77	0.8	0.49
FIGF|TFF3|VCAM1|WFDC2	16	21	6	17	0.62	0.55	0.73	0.78	0.48
CCL15|TFF3|VCAM1|WFDC2	16	21	6	17	0.62	0.55	0.73	0.78	0.48
A2M|ICAM1	19	18	3	20	0.62	0.47	0.86	0.86	0.49
IGFBP7|PARK7	21	15	1	23	0.6	0.39	0.95	0.94	0.48
CFH|TFF3	20	16	2	22	0.6	0.42	0.91	0.89	0.48
CCL15|FIGF	18	18	4	20	0.6	0.47	0.82	0.82	0.47
COL4A3|PARK7	19	17	3	21	0.6	0.45	0.86	0.85	0.48
ICAM1|WFDC2	18	18	4	20	0.6	0.47	0.82	0.82	0.47
VCAM1|WFDC2	19	16	3	22	0.58	0.42	0.86	0.84	0.46
PARK7|WFDC2	19	16	3	22	0.58	0.42	0.86	0.84	0.46

Bottom 20 panels (based on diagnostic accuracy) from total 206,337 panels generated, with diagnostic statistics. TP: True positive; TN: True negative; FP: False positive; FN: False negative; Sens: Sensitivity; Spec: Specificity; PPV: Positive predictive value; NPV: Negative predictive value.

These protein panels, generated only from proteins with some univariate classifying potential, show a range of accuracy between 92% for the top panels and 58% for the lowest.

3.5.6 Random forest modelling

Random forest modelling was performed on the dataset with 281 proteins remaining, as this gives a good overview of the variable importance of each protein among the larger dataset. Having previously determined that the “black box” result gained from this method is less suitable for a clinically applicable model, we use the results for informative purposes on protein importance while allowing for the likelihood of a more overfit model given the high number of variables computed.

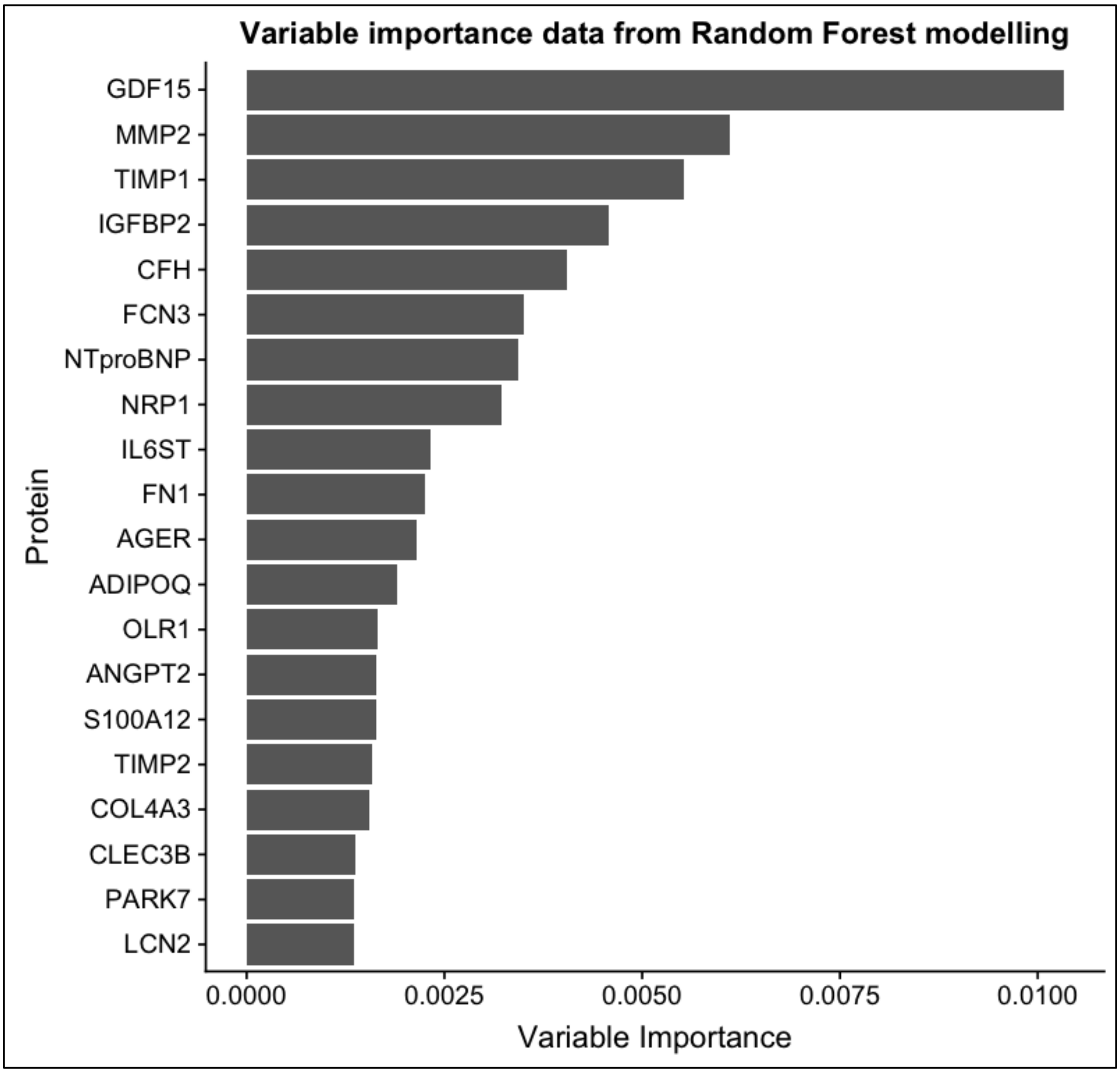


Figure 3.15: Variable importance for classification from random forest.

Graph demonstrates calculated 'importance' of independent variables for univariate classification of patients between the two groups. All variables receive a variable importance, top 20 shown. Abbreviations: see appendix.

This analysis ranks GDF15 as the most frequently selected variable with the purest downstream class split. AGER, ranked the highest important in Chapter 1, now falls down the list, but remains among the top 20 proteins.

3.5.7 LASSO modelling

As previously described, LASSO modelling will perform both variable selection and regression modelling. The LASSO process will only add variables to the model if doing so yields information gain. Due to this, collinear variables are dropped within the process, however there is little user choice as to the protein dropped once at this stage. Due to this, the input variables have been reduced as strictly as possible before entry into this analysis (Figure 3.13).

The 30 protein variables entered into this analysis are those shown in Figure 3.14, with MMP2 removed due to collinearity.

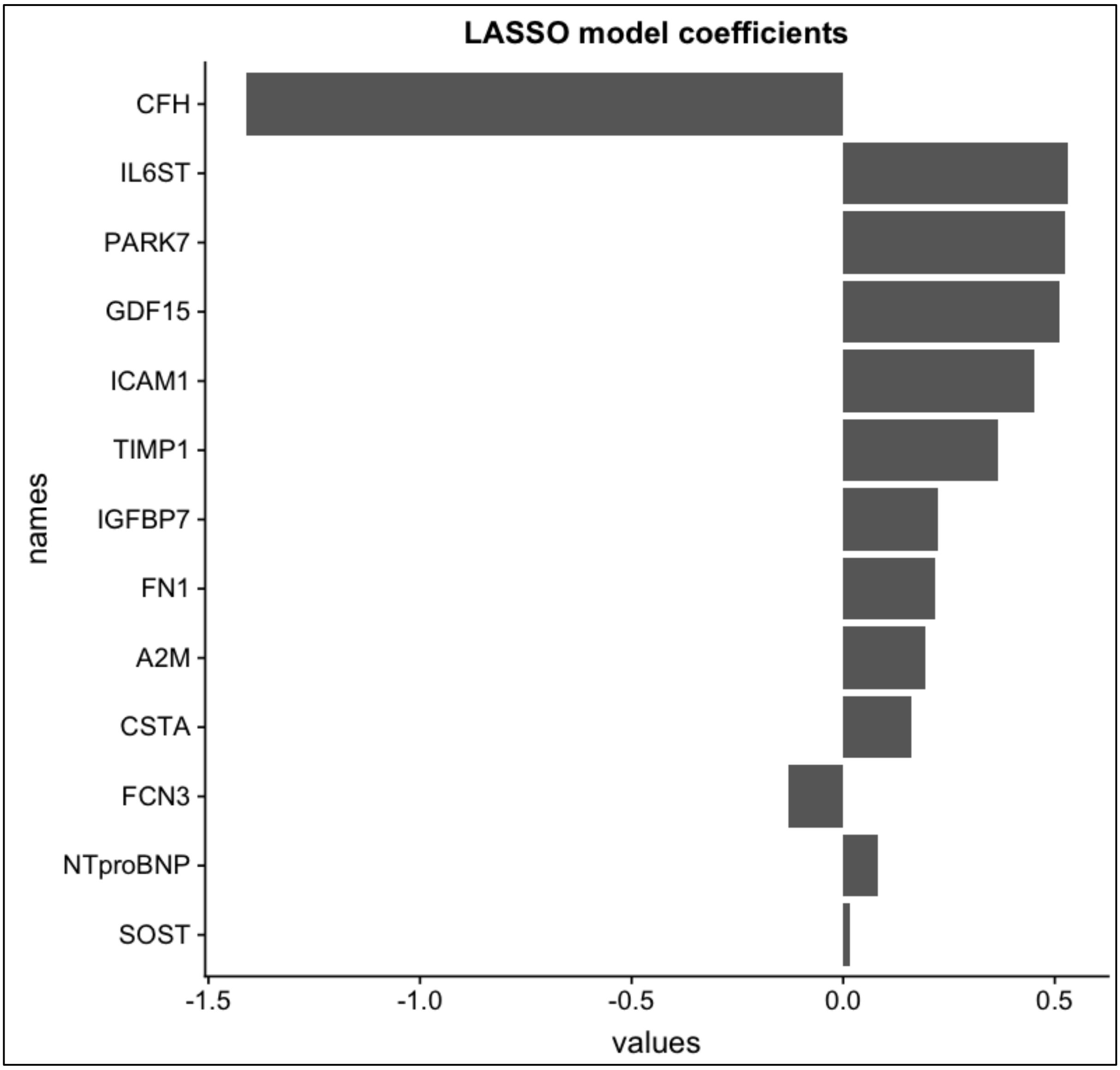


Figure 3.16: Output from LASSO analysis for classification (SSc-PAH vs SSc-no PH)

Bars represent the value of coefficients in the regression model.

Figure 3.16 shows the result of LASSO modelling, returning a regression formula with 13 protein variables, CFH as the strongest predictor in the model, with serum expression down-regulated in patients with PAH compared to controls.

3.5.8 Model performance

As for part 1 of this analysis, it was considered that the LASSO model represents the most open and useable model, which could be translated into clinical practice. It is on this basis that this model was taken as the most appropriate final classifying model for further study.

Applying the model to the derivation dataset gives the results shown in Figure 3.17.

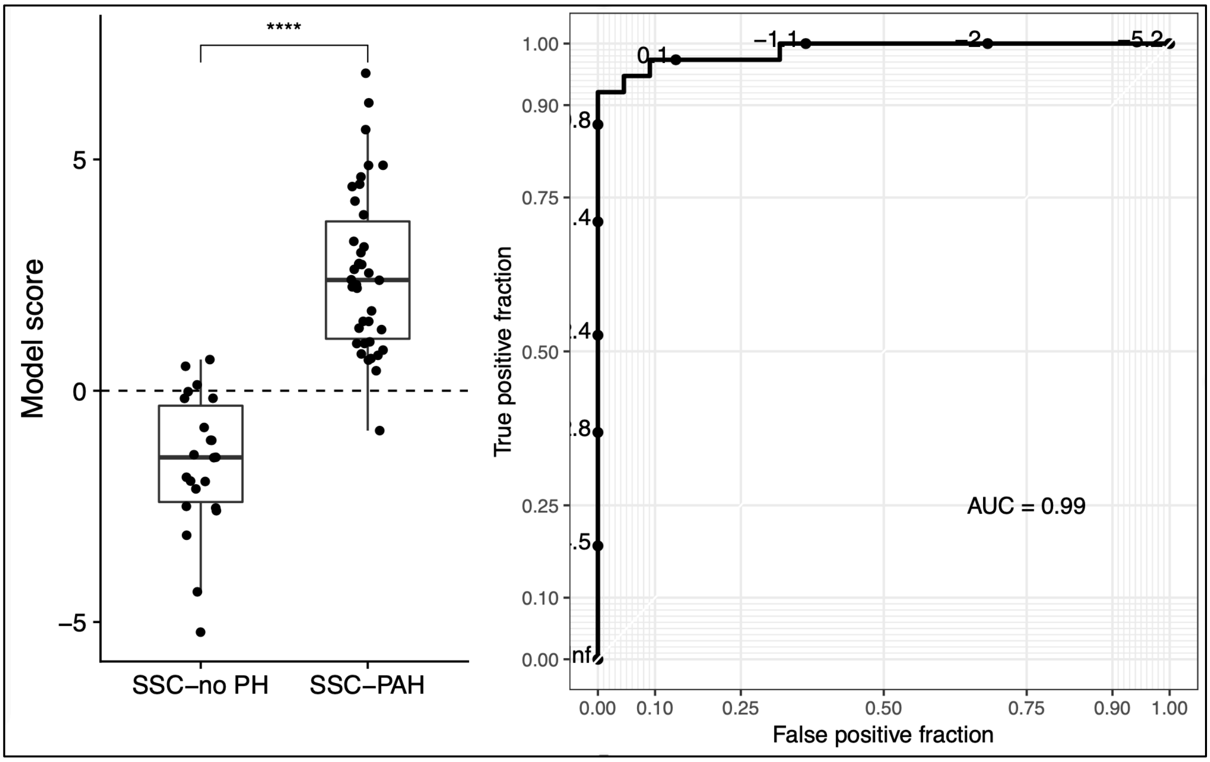


Figure 3.17: LASSO model applied to derivation dataset

Showing scores derived from LASSO model and derivation dataset (left), with cut-off classifying threshold (dotted line), and AU-ROC for classification of SSc-PAH from SSc-no PH on the right.

At a classifying threshold of 0, the following model statistics are as follows: Sensitivity 0.97, Specificity 0.86, Positive Predictive Value 0.93, Negative Predictive Value 0.95, False positive rate 13.6%, False negative rate 2.6% and Diagnostic accuracy 93.3%.

3.5.9 Validation of classifying model

With the intention of increasing the samples size in the derivation cohort we sent samples from a further 28 baseline, treatment naïve patients from the Sheffield biorepository (SSc-PAH n=20; SSc-no PH n=8) to Myriad RBM for identical analysis as were the original samples. Quality and batch control were assessed internally by Myriad RBM and are discussed further in section 4.5.1.

As an interim test of our methodology so far, we initially used these samples as we would for a validation cohort to test the model. The protein concentration data from these new patients were entered into the classifying model with the resulting scores shown in Figure 3.18.

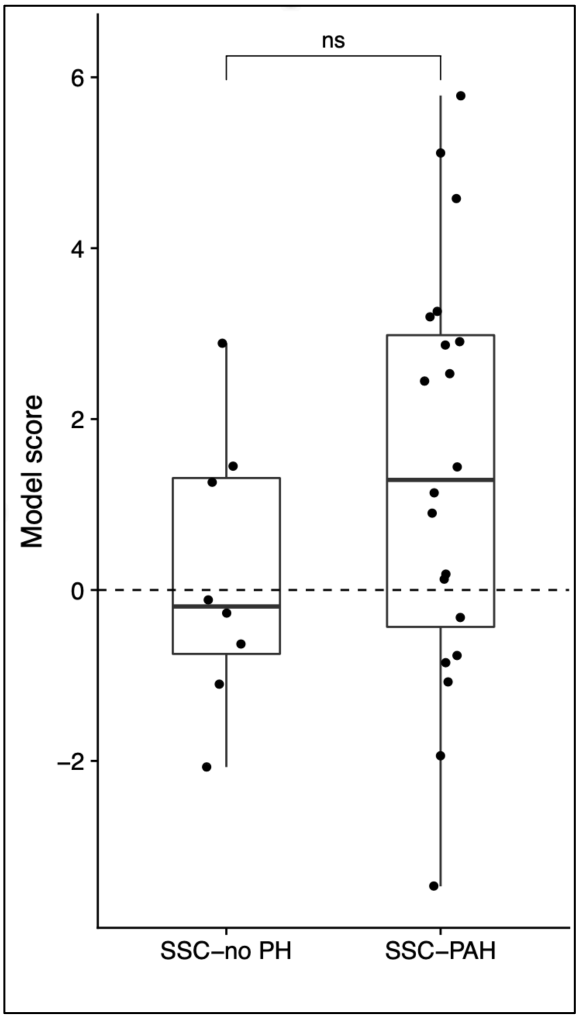


Figure 3.18: Scores for new samples entered into the classifying model

Split according to known diagnostic classification.

These results from a new cohort of patients demonstrate that the current model lacks any true classifying potential, and given the results from the derivation cohort (Figure 3.17), despite the level of variable pre-processing done, the model is significantly overfit to the derivation cohort only, and has no true general classifying potential.

In retrospect, a model with 13 protein variables to distinguish between only 60 patients is likely to represent an overfit model.

3.6 Discussion

In this section I set out with the aim to develop a statistical model which can use the available protein concentration dataset, or a subset of it, to accurately classify patients with SSc into those with or without PAH.

The dataset that we initially received from our collaborating partners included 296 protein concentrations for patients in the phenotype groups in Table 3.1. For our analyses, using only the data associated to patients with SSc, this presented the issue of modelling a very small number of samples with a very large number of predictors. In order to overcome this we used a multimodal approach to statistical modelling, exploring the application of both univariate statistics and a variety of multivariable modelling systems including machine learning tools.

Univariate statistics identified the proteins which could best classify disease on a single protein level, however the high number of repeated statistical tests, and the small sample size makes the risk of error high when planning to apply a model to a general population. A combination of protein biomarkers would provide a more secure model, less exposed than a single protein to confounding change in patient status.

Multivariable modelling in this dataset is not straightforward. As I require a binary outcome – classification – the most appropriate statistical approach would be with logistic regression using the protein concentration data as the predictors. This approach is not possible on this dataset as the model becomes saturated with too few patient samples and too many predictors in the derivation dataset and returns no statistical data. Agresti (2007) suggests that in building a logistic regression model a minimum of 10 samples for each predictor is required to produce a reliable model, which to model our large number of predictors would require a much larger patient sample.(Agresti, 2007) Machine learning models such as random forest and LASSO have been developed specifically to handle this type of dataset with a much greater number of predictors to samples available.

High level analysis of the dataset through PCA, both as a whole (Figure 3.1), and more specifically looking at the targeted groups (Figure 3.2 & Figure 3.12) demonstrate the complexity in classifying between these patient cohorts. There is significant overlap between the phenotype groups when shown on the first two principle components which suggests a complex relationship between the groups.

My first statistical analysis produced a classifying model which identified AGER as the most significant predictor of disease class. Post-hoc sub-analysis of an imbalance in ILD in the test arms demonstrated the significant risk of modelling in an unbalanced cohort, as expression of AGER protein is demonstrated to be significantly influenced by the presence or absence of this comorbidity. While Meloche et al. have demonstrated a clear increase in AGER expression in PASMCs from patients with PAH in vitro, it is also established that AGER expression is significantly reduced in patients with interstitial lung disease.(Manichaikul et al., 2017, Meloche et al., 2013) Our data would support both findings, however in a cohort of patients at high risk of both PAH and ILD, AGER is therefore an unreliable predictor of PAH in this disease group. These findings were shared with our collaborating partners at Actelion pharmaceutical and Myriad RBM who shared their particular interest in AGER following their analysis on their separate derivation cohort (to which we did not have any access, our dataset was their validation dataset). My findings were of interest as it is my understanding that beyond the matched phenotype group categories (same as Table 3.1), they do not have access to the level of detailed phenotype information regarding the presence or absence of ILD in their derivation dataset which we have provided for the Sheffield cohort.

Analysis of the dataset to account for ILD as a confounding co-morbidity required balancing the disease and control arms for this condition. I considered balancing the cohorts by removing patients with ILD from the control arm to maintain a very “clean” PAH cohort, but this approach resulted in a much smaller overall cohort with only 13 patients in the control arm. The dataset was already small at the outset compared to the number of predictors, and reducing the dataset further significantly increases risking the accuracy of a final model, and in particular reduces control over the risk of model overfitting. As previously discussed, there is a significant proportion of interstitial lung disease in the general population of patients with SSc in whom this type of screening tool would be placed. As such I felt it more appropriate to balance the cohort of patients by combining the SSc-PAH and SSc-PAH-ILD cohorts to produce a larger dataset which is now statistically balanced for presence of ILD, and other comorbidities. The proportion of ILD in each arm is also consistent with data on the prevalence of ILD in the SSc population, meaning that any classifying model is likely to be more applicable to this target population.

Repeat analysis of this cohort with balance proportions of ILD altered the final model. Again the LASSO model (Figure 3.16) was selected as the final classifying model as this represents a truly multivariable model which is open and transparent and can be manually calculated, in contrast to a random forest model which, although also a very good classifier, depends on highly complicated calculations of new data in a “black box” system in order to generate a predicted class. The proteins selected and coefficients thereof are significantly altered from the previous analysis, raising CFH as the most significant predictor, while AGER drops out entirely. Despite LASSO modelling within which one of the key advantages is the penalty term applied to shrink the number of protein variables with the key objective of addressing risk of model overfitting, the validation data (Figure 3.18) demonstrates that this model remains significantly overfit to the derivation dataset and is of little value when applied to an external cohort.

Model fit has proven the most significant challenge, particularly when dealing with a dataset with such significantly rotated dimensions. Model fit refers to how well a model describes a set of observations, with good fit a compromise between underutilisation of the derivation dataset to derive the classifying threshold – underfitting – to highly complicated modelling which describes the derivation dataset perfectly, but cannot generalise to accurately classify a new datapoint – overfitting (Figure 3.19).

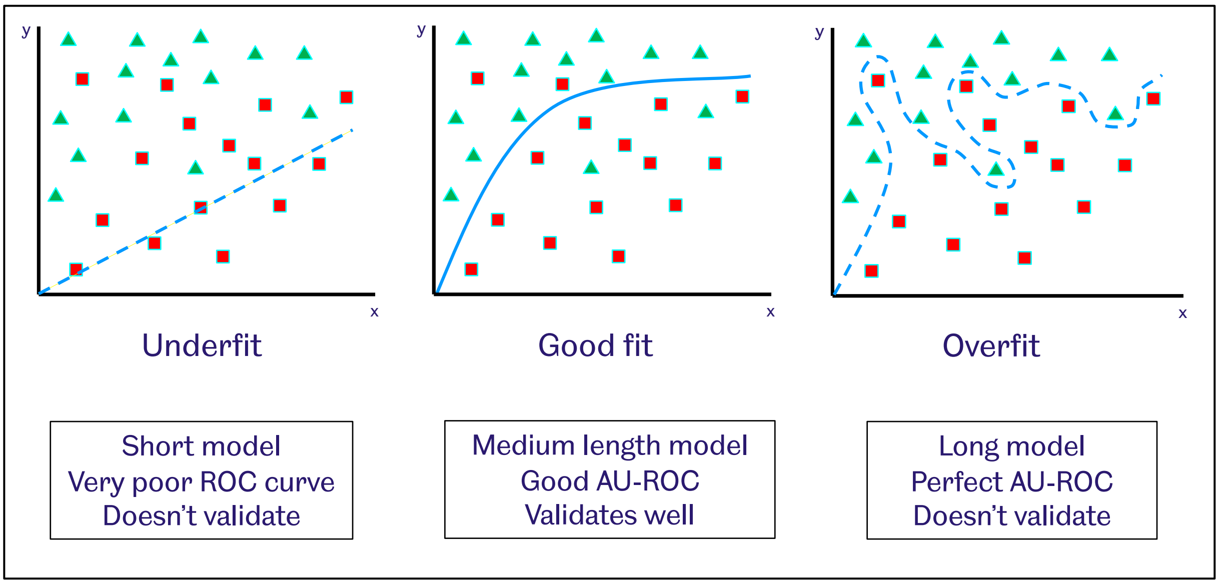


Figure 3.19: Model fit

Representation of the extremes of model fit. Red squares and green triangles represent a derivation dataset consisting of two populations to classify using variables x and y. An underfit model is overly simplistic and poorly represents the derivation dataset, and cannot accurately classify and new datapoint; An overfit model is highly complex and perfectly describes the derivation dataset, but cannot accurately classify a new datapoint. A model with good fit is a compromise between the two, accepting appropriate misclassification in the derivation dataset in order to produce a reasonable classification threshold which can accurately predict the class of a new datapoint.

Our classifying model consists of 13 proteins required to classify disease in a sample of 88 patients. This is a highly complex model which has proven itself overfit when tested against a validation dataset. More stringent constraint of model size is required to produce an accurate predictor.

These analyses have demonstrated the errors that are inherent to this type of classification modelling. Despite significant overfitting in this classifying model, the proteins selected by each of the modelling methods remain significantly altered between the disease and control cohorts at an individual protein level, and therefore the proteins selected remain informative as to the types of proteins likely related to the underlying pathophysiology of PAH and will be allowed some consideration in further modelling.

4 Final classification modelling

4.1 Introduction

The analyses in Chapter 3 identified issues with the effect of imbalanced comorbidities, and the high risk of model overfitting error given the large number of variables and small number of samples in our dataset. For these reasons we decided to increase the samples size for analysis by identifying further treatment naïve patient samples from the Sheffield biorepository and sending these to Myriad RBM for identical analysis. 28 new patients (SSc-PAH n=20; SSc-no PH n=8) were identified to increase the samples size of the derivation cohort.

We then sought to develop a more robust classifying model based on these new combined patient data, with alternative methods to eliminate the errors seen in our earlier work.

4.2 Aim

To develop a classifying model which can accurately classify between SSc-PAH and SSc-no PH, and which can generalise and retain accuracy when validated against an external validation cohort and is not significantly affected by any imbalance in comorbidities.

4.3 Methods

4.3.1 Patient cohort identification and assay

To identify any further suitable patients available to include in a larger analysis we interrogated the Sheffield Pulmonary Vascular Research group biorepository database. This was done using search criteria appropriate to identify any patients not yet included in our analysis, which fit the three major criteria of:

1. Patient with systemic sclerosis.
2. Patients with RHC proven PH considered to be of PAH subtype and subsequently treated as such, or patients fully investigated for PH but found not to have PH at RHC.
3. Patients from which there are treatment naïve baseline serum samples available for analysis.

Further to this, the same detailed demographic information was recovered as for the original patient cohort (Chapter 2.2: Patient Data) from either the biorepository database, or from exhaustive search of patient records and clinical systems. Demographic information, RHC data and imaging investigations were reviewed before accepting patients into the study.

Serum samples from identified patients were shipped on dry ice to Myriad RBM and confirmed to have been received frozen and in good condition. Protein assays were performed by Myriad RBM by the method recorded in 2.3.1.2, and results reported to us for statistical analysis.

4.3.2 Analysis

This analysis includes data from two major assay batches, so initially some work was put into quality control and batch analysis and is reported in section 4.5.1.

Model overfitting to the derivation dataset was the major cause of failure of our earlier analyses, due to small patient sample and relatively large size of the classifying models generated. The increased sample size was obtained to improve on this, however in contrast to previous analyses we will also uncouple variable selection from statistical modelling, allowing for variable selection from a range of machine learning techniques, followed by a modelling procedure designed to be much more stringent in constraining model length. By doing so we will reduce the likelihood that the final model will be overfit to the derivation dataset.

Using an updated derivation dataset, variable selection will be conducted using the same analysis methods described and used in Chapter 3, taking into account the results from random forest and LASSO modelling, along with univariate statistics and manual panel building, and at this stage also considering the limited results available from the parallel analysis done by our collaborating partners at Actelion pharmaceutical. Protein variables identified most frequently across these methods will be taken forward to statistical modelling.

Statistical modelling will be initially done using logistic regression on protein concentration data, with the model subsequently optimized using a recursive backward-step AIC. The use of the logistic regression model allows for a greater transparency for the role of each individual protein included in the model through the reporting of individual component p-values. The model will then manually be further iteratively constrained by removing at each step any protein variable in the model which does not reach statistical significance (p<0.05).

The final statistical model will be used to generate diagnostic statistics based on the score generated for each patient in the derivation cohort.

4.4 Demographics

The new patient cohort includes 28 newly identified patients, 20 SSc-PAH and 8 SSc-no PH, increasing the patient numbers in each of the disease and control groups for analysis (Table 4.1).

Table 4.1: Demographics for Myriad2 analysis

		SSC-PAH	SSC-no PH	p-value
n		58	30	
Age (IQ range)		66.5 (62 - 71.8)	63.5 (57.3 - 67)	0.21
Gender (M/F)		15/43	2/28	0.06
Deaths		36	5	<0.001
WHO FC	1	0	1	<0.001
	2	7	11	
	3	45	18	
	4	6	0	
Co-Morbidity	COPD (%)	7 (12.1)	1 (3.3)	0.34
	Haemolysis (%)	0 (0)	0 (0)	NA
	Myeloproliferative (%)	1 (1.7)	0 (0)	1
	AF (%)	2 (3.5)	1 (3.3)	1
	A Flutter (%)	0 (0)	0 (0)	1
	ILD (%)	21 (36.2)	11 (36.7)	1
	Asthma (%)	2 (3.5)	1 (3.3)	1
	Sarcoidosis (%)	0 (0)	1 (3.3)	0.74
	OSA (%)	3 (5.2)	0 (0)	0.52
	VTE (%)	4 (6.9)	2 (6.7)	1
PFTs (median + IQ))	FEV1 Percent	83.3 (66.1 - 91.2)	87 (70.8 - 102.7)	0.23
	FVC Percent	92.8 (76.7 - 108.8)	98.1 (76.6 - 113.2)	0.4
	TLCO Percent	39.2 (28.3 - 45.1)	53.7 (46.7 - 66.8)	<0.001
ISWT (median + IQ)	Distance	120 (75 - 230)	210 (97.5 - 360)	0.041
RHC (median + IQ)	mPAP	37.5 (30 - 44)	21 (18 - 22)	<0.001
	RA pressure	8 (4 - 11)	4 (3 - 5.8)	<0.001
	CI	2.9 (2.4 - 3.4)	3.3 (3 - 4.1)	<0.001
	PVR	418 (304 - 621)	148 (117 - 192)	<0.001
	PCWP	10 (7 - 13)	8 (6 - 10.8)	0.07

Abbreviations: n – number; IQ – interquartile; M/F – Male/Female; WHO FC – World health organisation functional class; COPD – chronic obstructive pulmonary disease; AF – Atrial fibrillation; A Flutter – Atrial flutter; ILD – Interstitial lung disease; OSA – Obstructive sleep apnoea; VTE – Venous thromboembolism; PFT – Pulmonary function tests; ISWT – incremental shuttle walking test; RHC – right heart catheter

The cohort remains balanced for medical comorbidities and is now better balanced for age. There remains a slight imbalance in gender, however this does not reach statistical significance.

4.5 Results

4.5.1 Quality control between batches

The dataset returned to us from Myriad RBM contained only the protein concentration measurements requested, and did not return any information regarding the assay quality control or standard curves. Batch analysis was therefore limited to analysing for large scale changes in our protein concentration data which could accurately identify which batch they belonged to, done by a combination of PCA analysis, and individual protein distribution analysis. Following this, further assurances regarding data quality and batch control were received from Myriad RBM.

4.5.1.1 Batch effect

Batch effect refers to technical variations in assays specific to each run if the assay was run in batches rather than all on one plate at a single timepoint. For this analysis we have combined the data from two batches of results from Myriad RBM, so an assessment for any evidence of batch effect is important. To do so I used a principle component analysis to identify any clustering of the two batches which might suggest a batch effect (Figure 4.1).

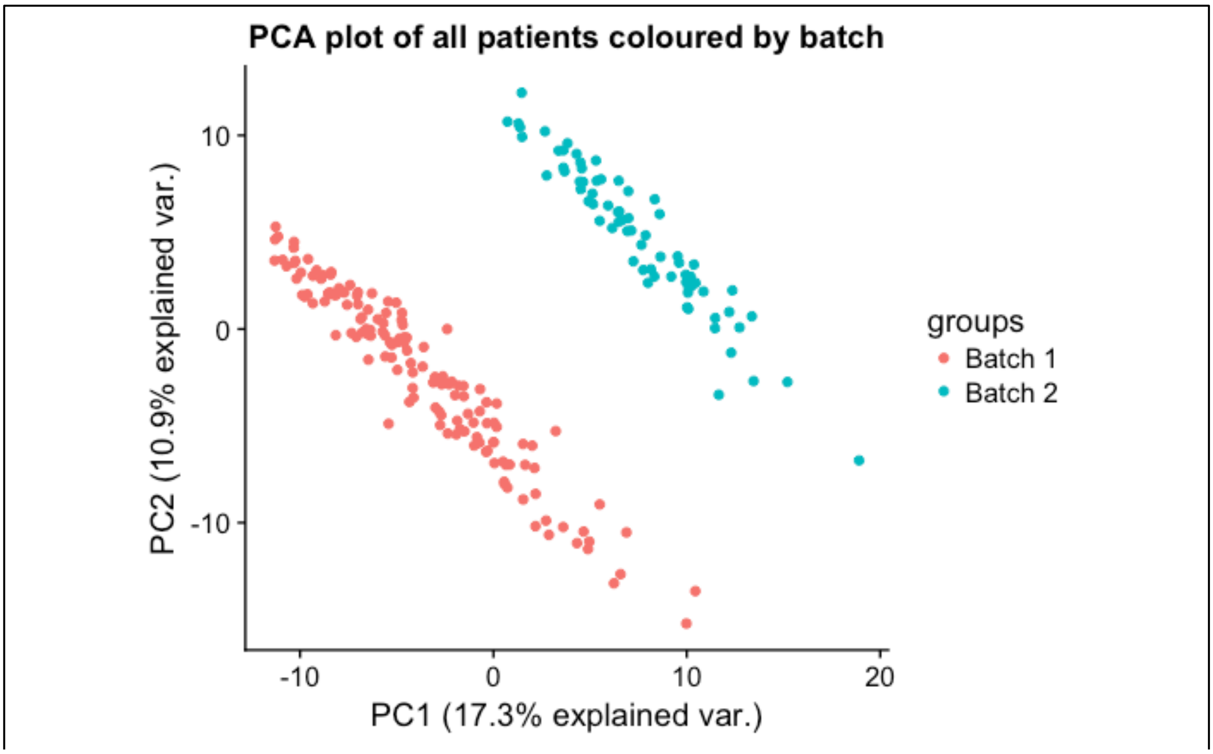


Figure 4.1: PCA of full patient cohort using all protein data.

Batches identified by colours

The initial analysis demonstrated an apparent clear batch effect between the two groups which required further investigation. Review of the full dataset suggested that this effect originated from the limits of detection which were set differently for each protein between the two batches. Protein variables with a sample count outside the limit of detection were therefore creating a clear signal which differentiated the two batches. To minimize this effect, the dataset was reduced to remove any protein for which <10% of values are exactly duplicated (i.e. beyond the limits of detection). Subsequent PCA analysis does not identify patient clustering based on batch alone (Figure 4.2).

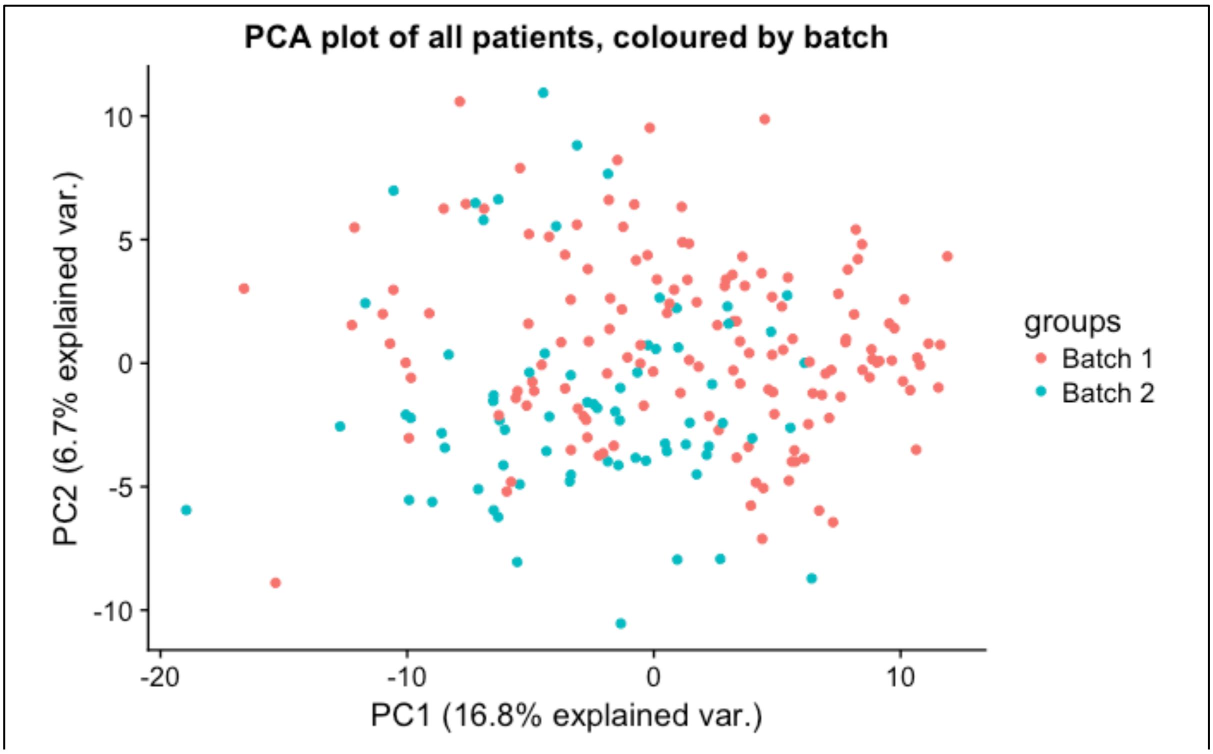


Figure 4.2: PCA analysis after limits of detection addressed

Batches identified by colours

4.5.1.2 Protein distributions

As a second quality control test, we compared the distribution of individual protein concentrations between batches as these should remain similar, allowing for differences attributable to differences in disease proportions between the two batches. Biological distributions are generally non-parametric, and therefore after reviewing graphical distributions for all individual proteins they were also assessed statistically with Kolmogorov-Smirnov test p-value. All distributions were examined individually and allowing for reasonable variation, these did not reveal evidence of significant batch effect. A representative sample of the top 4 proteins from Myriad1b LASSO model is given in Figure 4.3.

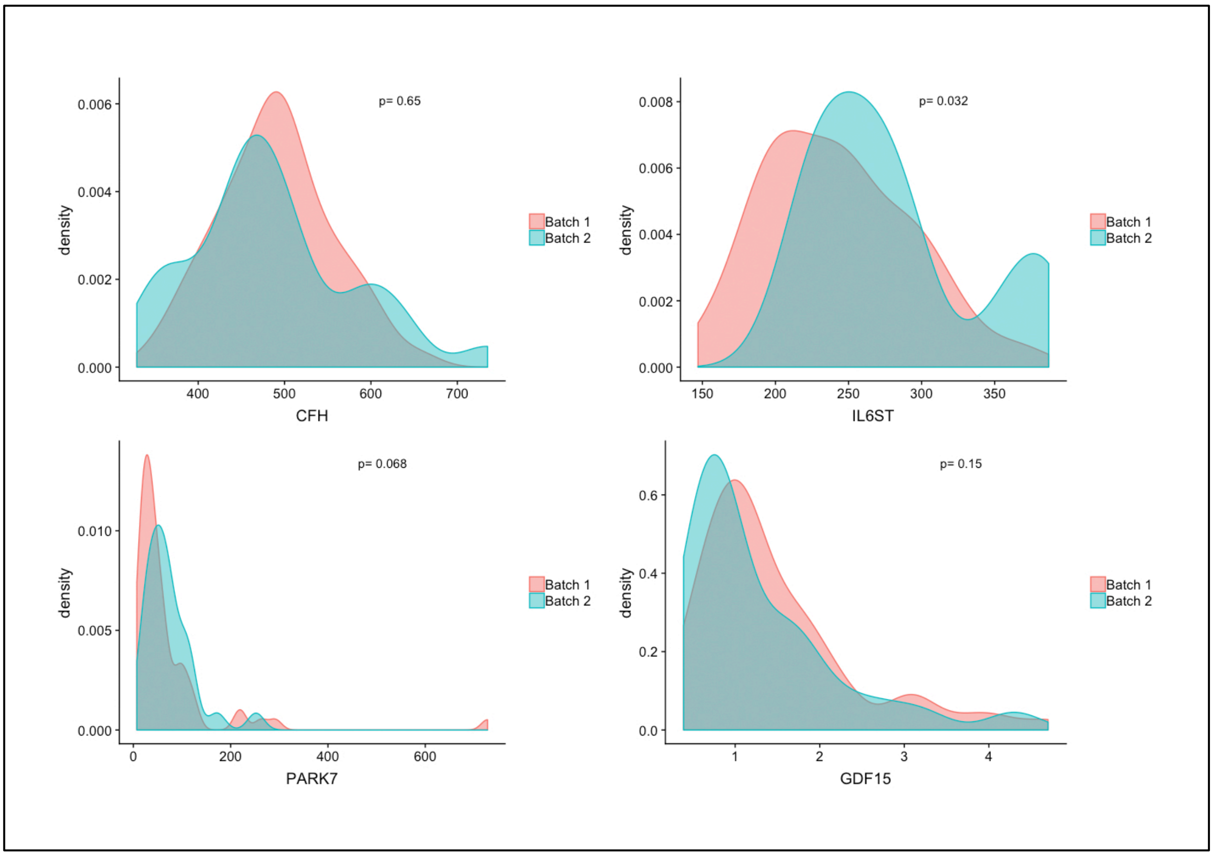


Figure 4.3: Protein concentration distribution plots for batch comparison

Top 4 proteins from Myriad1b analysis shown as representative of whole protein dataset. Batches identified by plot colours. Kolmogorov-Smirnoff p-value for significantly different distribution shown on each graph. Top left: CFH (µg/ml), Top right: IL6ST (ng/ml); Bottom left: PARK7 (ng/ml); Bottom right: GDF15 (ng/ml).

4.5.1.3 Myriad RBM Internal QC

Myriad RBM participates in regular high level quality control audit and testing. The Myriad RBM laboratory holds Clinical Laboratory Improvement Amendment accreditation which is the US federal regulatory standard that approves a lab for direct human clinical testing. A key criterion of this accreditation is calibration and quality control over testing. Each stage of the assay process from kit manufacture to the testing process is highly controlled to reduce variation. Each assay is individually scrutinized for quality, using the standard curves and three level control samples which are plotted across assay runs to enable charting of assay consistency.(Welsh et al.)

In our data, some minor variation was noted during examination of the protein concentration distribution plots, however it was not possible to ascertain from the protein concentration data alone whether this was due to technical batch variation, or whether this is due to a biological signal given the differing disease proportions in each batch. Further to this, given relatively small numbers in each batch, minor biological variations in protein concentration can exert a much larger effect on an average distribution.

Quality control was discussed with Myriad RBM who assured us that their rigorous quality control procedures are in place and followed at all times and on all assays performed in their facility. Batch effect was also analysed between their stored results from each of our two batches and no evidence of any batch effect was found between the two.

4.5.2 Protein variable pre-processing and Exploratory data analysis

Variable pre-processing varied slightly from previous iterations due to the different absolute limits of detection given for each analysis batch, requiring each batch to be analysed independently for the proportion of datapoints falling outside the limit prior to combining the batches for further pre-processing and subsequent variable selection (Figure 4.4). Any protein variable if in either analysis batch >90% of protein datapoints fall outside the limit of detection then the variable is excluded. 67 protein variables were excluded by this criterion, each of which was individually examined for any predictive relationship to disease class, however no additional information gain was found. Within the remaining 229 protein variables for 88 patient samples, there were 22 missing datapoints from a total of 20152 (0.11%) due to insufficient sample at assay. These were deemed to be missing at random and imputed using MissForest. Datapoints remaining outside the limits of detection were revalued at the corresponding absolute limit of detection.

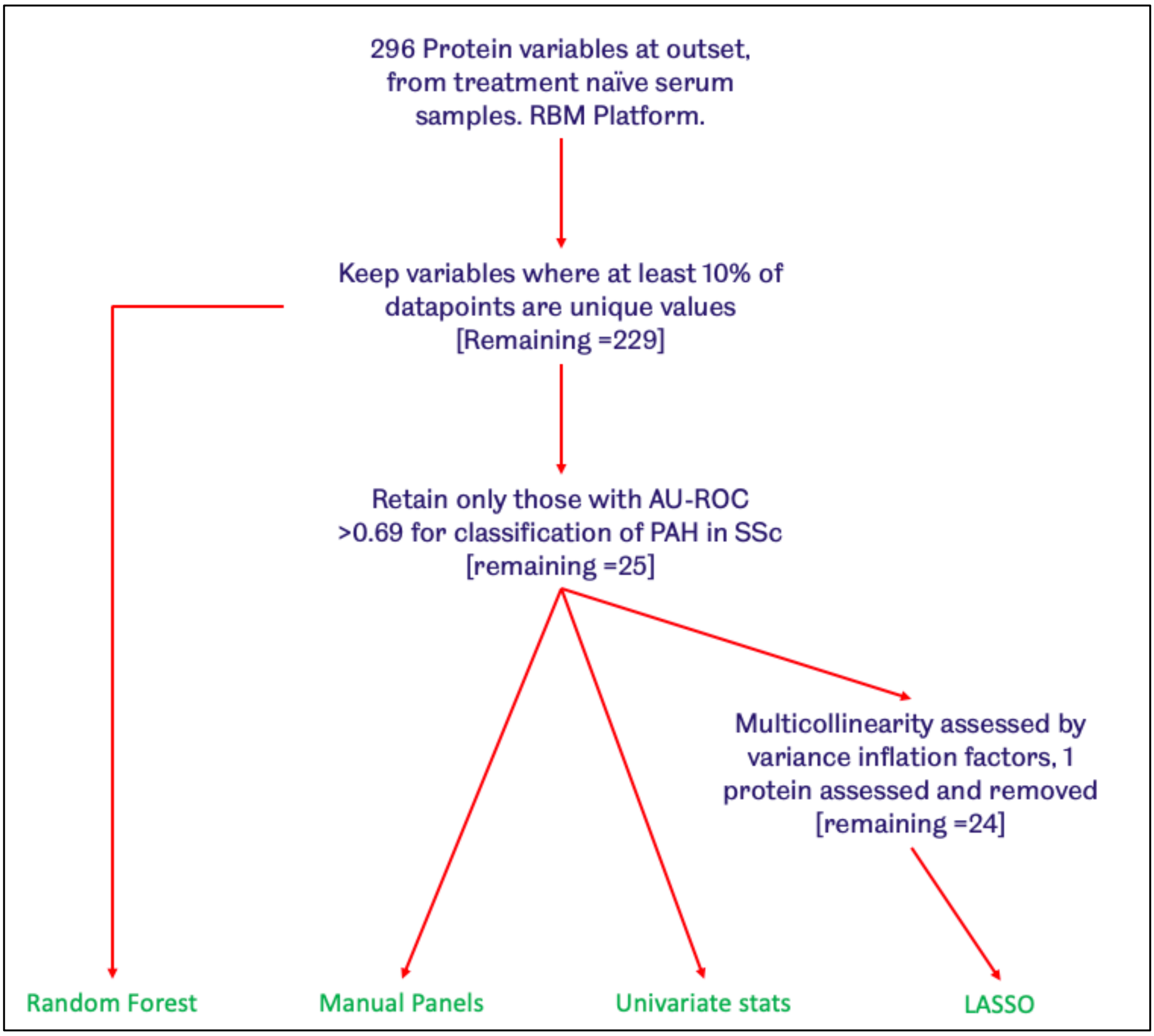

296 Protein variables at outset, from treatment naïve serum samples. RBM Platform.
Keep variables where at least 10% of datapoints are unique values [Remaining =229]
Retain only those with AU-ROC >0.69 for classification of PAH in SSc [remaining =25]
Multicollinearity assessed by variance inflation factors, 1 protein assessed and removed [remaining =24]
Random Forest
Manual Panels
Univariate stats
LASSO

Figure 4.4: Data pre-processing flow diagram

High level exploratory data analysis of this cohort with all 229 proteins with PCA shows some weak evidence of clustering of patients but with very high level of overlap between the two groups (Figure 4.5).

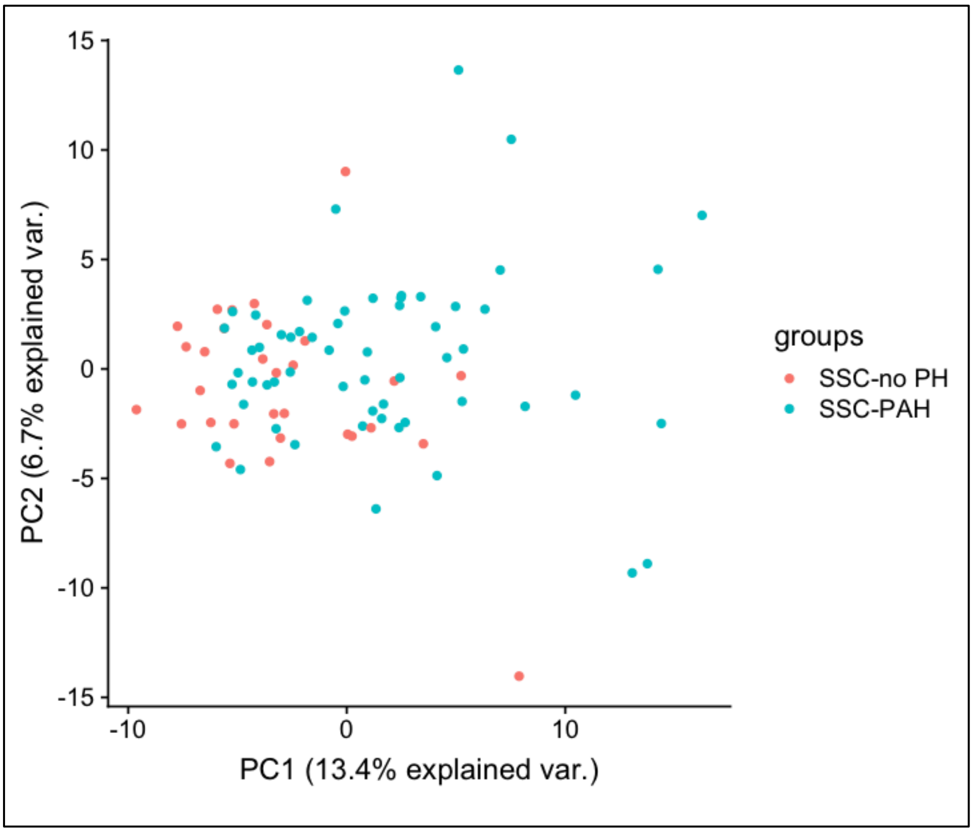


Figure 4.5: PCA of all 88 patients in Myriad2 analysis

PCA of all patients using all 229 proteins at this point. Points coloured according to known patient classification.

Basic fold change analysis of 229 proteins is shown in Figure 4.6. This demonstrates the number of proteins which are significantly altered between the disease and control group, and the magnitude of change.

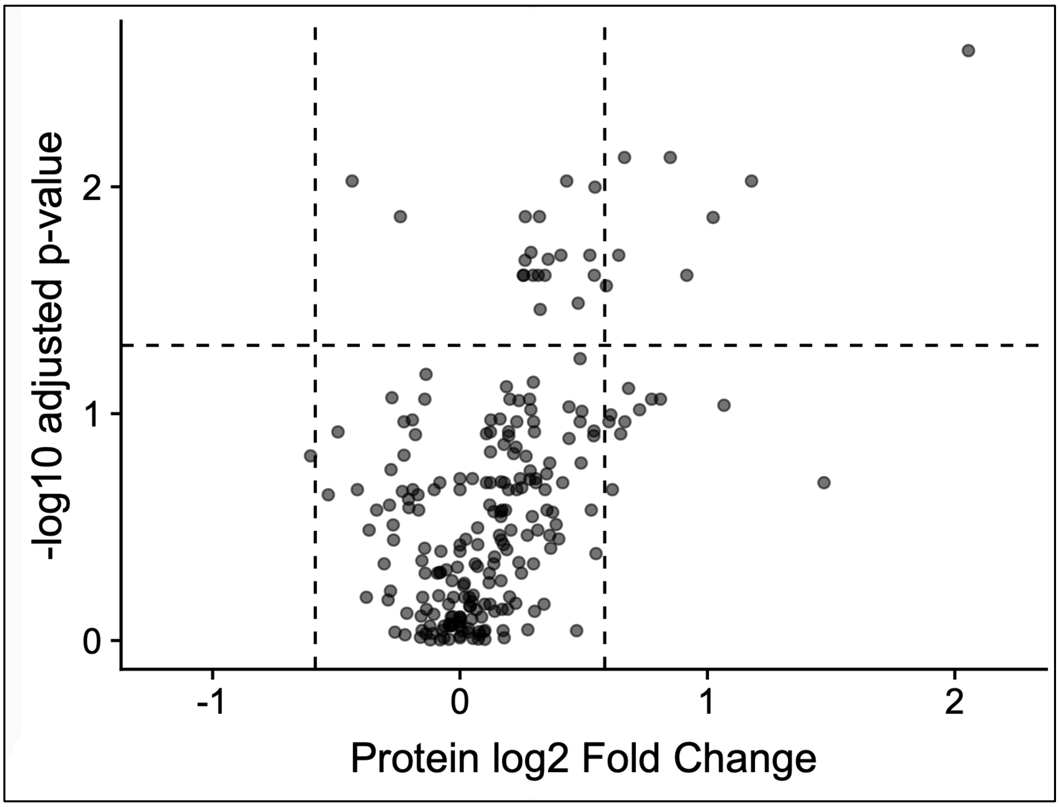


Figure 4.6: Fold change analysis of all 229 proteins

Fold change between SSc-PAH and SSc-no PH for all 229 proteins. Each point is translucent to prevent obscuration of overlapping proteins. Horizontal line corresponds to p=0.05 threshold, and vertical lines represent 1.5 fold change each way.

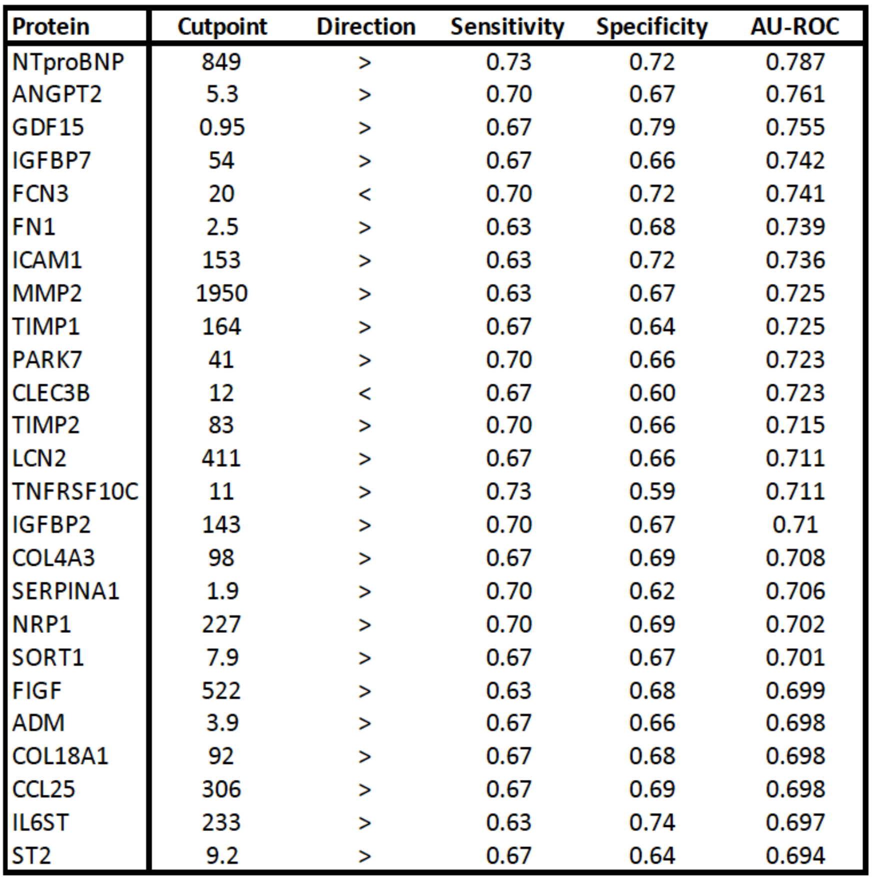
4.5.3 Variable selection

Due to the previously described issues with overfitting seen with the combined variable selection and modelling techniques, we decided to uncouple variable selection from final classification modelling. Variable selection was conducted using the results of all four previously described methods; univariate classifying statistics; combination panels based on these statistics; random forest models and the results of LASSO. At this point we also included in our considerations the limited data and variable selection shared with us from parallel analysis done by the collaborating team at Actelion pharmaceutical (3.4.9).

4.5.3.1 Univariate classifying statistics

The remaining 25 proteins were analysed and ranked according to their univariate classifying potential, using AU-ROC as the metric of choice (Table 4.2).

Table 4.2: Individual protein thresholds from ROC analysis and univariate diagnostic statistics



Protein	Cutpoint	Direction	Sensitivity	Specificity	AU-ROC
NTproBNP	849	>	0.73	0.72	0.787
ANGPT2	5.3	>	0.70	0.67	0.761
GDF15	0.95	>	0.67	0.79	0.755
IGFBP7	54	>	0.67	0.66	0.742
FCN3	20	<	0.70	0.72	0.741
FN1	2.5	>	0.63	0.68	0.739
ICAM1	153	>	0.63	0.72	0.736
MMP2	1950	>	0.63	0.67	0.725
TIMP1	164	>	0.67	0.64	0.725
PARK7	41	>	0.70	0.66	0.723
CLEC3B	12	<	0.67	0.60	0.723
TIMP2	83	>	0.70	0.66	0.715
LCN2	411	>	0.67	0.66	0.711
TNFRSF10C	11	>	0.73	0.59	0.711
IGFBP2	143	>	0.70	0.67	0.71
COL4A3	98	>	0.67	0.69	0.708
SERPINA1	1.9	>	0.70	0.62	0.706
NRP1	227	>	0.70	0.69	0.702
SORT1	7.9	>	0.67	0.67	0.701
FIGF	522	>	0.63	0.68	0.699
ADM	3.9	>	0.67	0.66	0.698
COL18A1	92	>	0.67	0.68	0.698
CCL25	306	>	0.67	0.69	0.698
IL6ST	233	>	0.63	0.74	0.697
ST2	9.2	>	0.67	0.64	0.694

Statistics for individual proteins. Threshold: diagnostic cut-off level at Youden index with value for individual protein. Direction: Demonstrating whether the disease group has increased or suppressed protein expression. AUC: Area under the curve on receiver operating curve. Protein abbreviations: see appendix.

4.5.3.2 Panels based on univariate statistics

Combination panels were generated based on the univariate statistics given in Table 4.2.

Table 4.3: Diagnostic panels (top 20)

Protein panel	TP	TN	FN	FP	Accuracy	Sens	Spec	PPV	NPV
CLEC3B|GDF15|IGFBP7|IL6ST|PARK7	25	53	5	5	0.89	0.91	0.83	0.91	0.83
CCL25|CLEC3B|COL4A3|FCN3|FIGF	24	54	6	4	0.89	0.93	0.80	0.90	0.86
CLEC3B|FCN3|FIGF|GDF15|ICAM1	25	53	5	5	0.89	0.91	0.83	0.91	0.83
CLEC3B|FCN3|FIGF|ICAM1|TIMP2	25	53	5	5	0.89	0.91	0.83	0.91	0.83
CLEC3B|FIGF|GDF15|LCN2|NRP1	24	54	6	4	0.89	0.93	0.80	0.90	0.86
CLEC3B|FIGF|FN1|GDF15|IL6ST	25	53	5	5	0.89	0.91	0.83	0.91	0.83
CLEC3B|FN1|GDF15|IL6ST|PARK7	24	53	6	5	0.88	0.91	0.80	0.90	0.83
CLEC3B|COL4A3|GDF15|IL6ST|PARK7	25	52	5	6	0.88	0.90	0.83	0.91	0.81
CLEC3B|FCN3|FIGF|FN1|IL6ST	24	53	6	5	0.88	0.91	0.80	0.90	0.83
CCL25|CLEC3B|COL4A3|FCN3|SERPINA1	25	52	5	6	0.88	0.90	0.83	0.91	0.81
CLEC3B|FCN3|FIGF|FN1|ICAM1	24	53	6	5	0.88	0.91	0.80	0.90	0.83
CCL25|FIGF|GDF15|ICAM1|IL6ST	25	52	5	6	0.88	0.90	0.83	0.91	0.81
CLEC3B|FIGF|GDF15|IGFBP7|SERPINA1	23	54	7	4	0.88	0.93	0.77	0.89	0.85
CCL25|GDF15|ICAM1|IL6ST|MMP2	25	52	5	6	0.88	0.90	0.83	0.91	0.81
CCL25|FIGF|GDF15|IL6ST|TNFRSF10C	25	52	5	6	0.88	0.90	0.83	0.91	0.81
CCL25|COL4A3|FCN3|FIGF|ICAM1	26	51	4	7	0.88	0.88	0.87	0.93	0.79
CCL25|CLEC3B|FCN3|FIGF|ICAM1	25	52	5	6	0.88	0.90	0.83	0.91	0.81
CCL25|CLEC3B|FIGF|GDF15|SERPINA1	24	53	6	5	0.88	0.91	0.80	0.90	0.83
CCL25|CLEC3B|IL6ST|LCN2|NTproBNP	25	52	5	6	0.88	0.90	0.83	0.91	0.81
CCL25|CLEC3B|FIGF|IL6ST|LCN2	25	52	5	6	0.88	0.90	0.83	0.91	0.81

Top 20 panels (based on diagnostic accuracy) from total 68,380 panels generated, with diagnostic statistics. TP: True positive; TN: True negative; FP: False positive; FN: False negative; Sens: Sensitivity; Spec: Specificity; PPV: Positive predictive value; NPV: Negative predictive value.

Table 4.4: Diagnostic panels (bottom 20)

Protein panel	TP	TN	FN	FP	Accuracy	Sens	Spec	PPV	NPV
CLEC3B|TIMP2	26	28	4	30	0.61	0.48	0.87	0.88	0.46
ADM|SORT1	26	28	4	30	0.61	0.48	0.87	0.88	0.46
COL4A3|FIGF	26	28	4	30	0.61	0.48	0.87	0.88	0.46
COL4A3|IGFBP7	26	28	4	30	0.61	0.48	0.87	0.88	0.46
CLEC3B|LCN2	26	28	4	30	0.61	0.48	0.87	0.88	0.46
CCL25|IGFBP7	26	28	4	30	0.61	0.48	0.87	0.88	0.46
CLEC3B|FN1	26	28	4	30	0.61	0.48	0.87	0.88	0.46
ICAM1|ST2	26	28	4	30	0.61	0.48	0.87	0.88	0.46
FN1|LCN2	27	27	3	31	0.61	0.47	0.90	0.90	0.47
FCN3|MMP2	23	31	7	27	0.61	0.53	0.77	0.82	0.46
CLEC3B|SERPINA1	27	27	3	31	0.61	0.47	0.90	0.90	0.47
FN1|TIMP1	27	27	3	31	0.61	0.47	0.90	0.90	0.47
FIGF|FN1	27	26	3	32	0.60	0.45	0.90	0.90	0.46
CCL25|FIGF	26	27	4	31	0.60	0.47	0.87	0.87	0.46
ICAM1|MMP2	26	27	4	31	0.60	0.47	0.87	0.87	0.46
FN1|ICAM1	23	30	7	28	0.60	0.52	0.77	0.81	0.45
MMP2|ST2	25	28	5	30	0.60	0.48	0.83	0.85	0.45
CLEC3B|SORT1	25	28	5	30	0.60	0.48	0.83	0.85	0.45
FN1|MMP2	26	26	4	32	0.59	0.45	0.87	0.87	0.45
MMP2|SORT1	26	26	4	32	0.59	0.45	0.87	0.87	0.45

Lowest 20 panels (based on diagnostic accuracy) from total 68,380 panels generated, with diagnostic statistics. TP: True positive; TN: True negative; FP: False positive; FN: False negative; Sens: Sensitivity; Spec: Specificity; PPV: Positive predictive value; NPV: Negative predictive value.

68,380 panels in total were tested, with the highest returning a diagnostic accuracy of 89% and the lowest returning 59%. The frequency of occurrence of proteins in the top panels was used to inform the variable selection process to follow.

4.5.3.3 Random forest modelling

For this Myriad2 analysis, 229 proteins were entered into the random forest algorithm for classification, and all proteins were then ranked according to their variable importance (Figure 4.7).

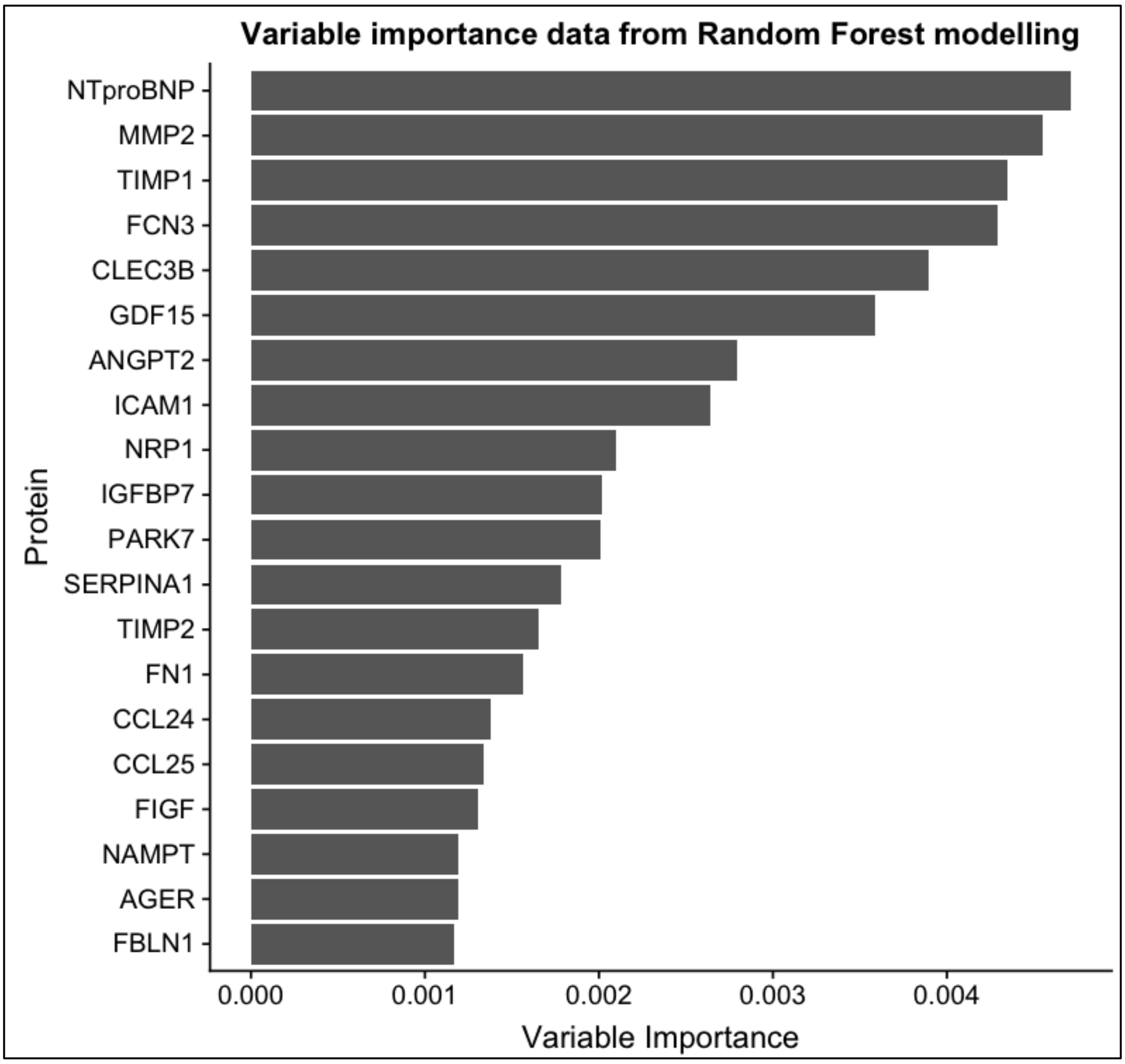


Figure 4.7: Variable importance for classification from random forest

Graph demonstrates calculated 'importance' of independent variables for univariate classification of patients between the two groups. All variables receive a variable importance, top 20 shown. Abbreviations: see appendix.

4.5.3.4 LASSO modelling

Protein concentration data for 25 proteins were entered into LASSO modelling. For this analysis, data were log transformed, centred and scaled prior to analysis. Analysis for multicollinearity with variance inflation factors identified one protein (MMP2) as highly correlated with other included proteins and this was excluded. The model returned by LASSO is given in (Figure 4.8).

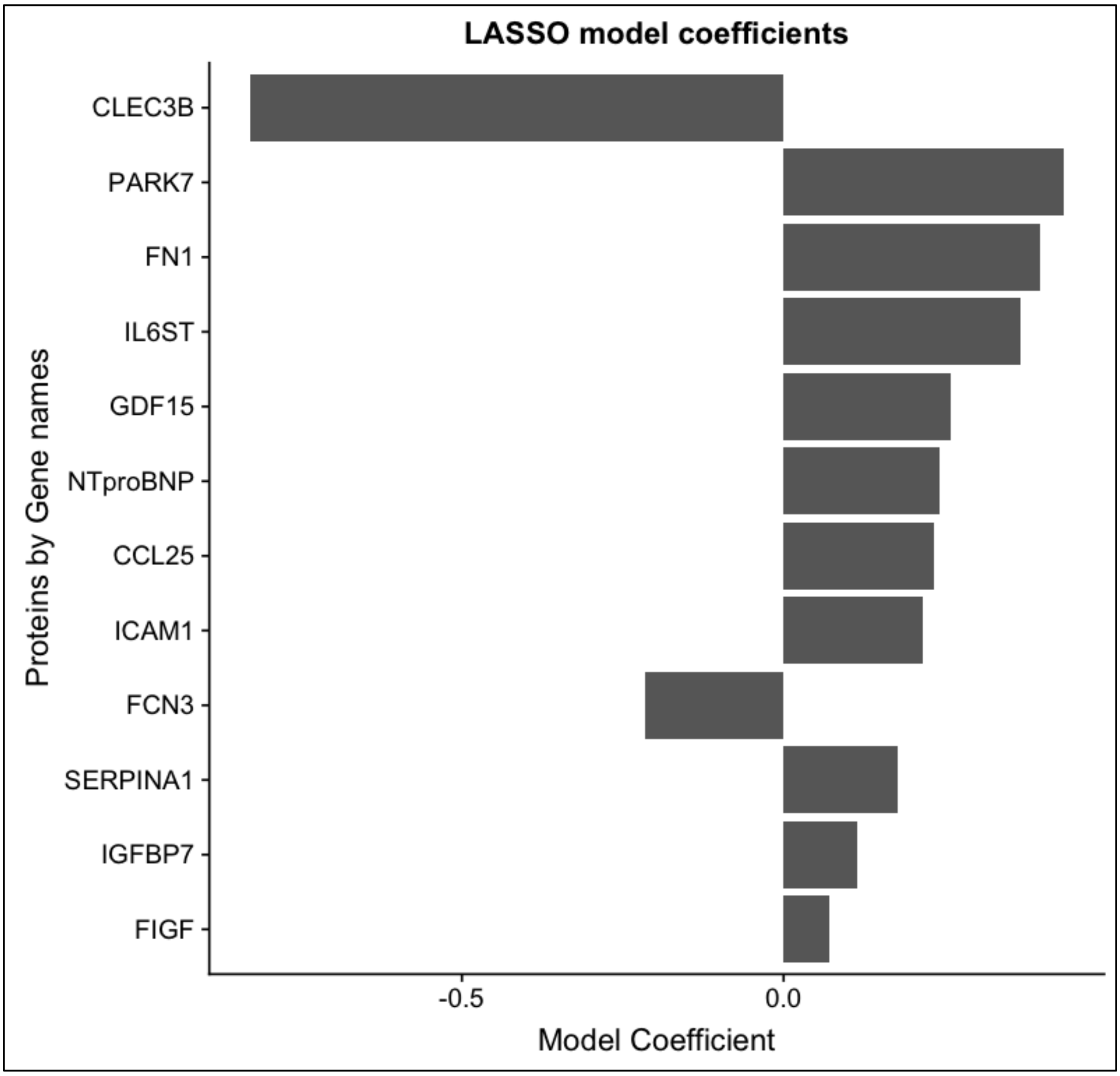


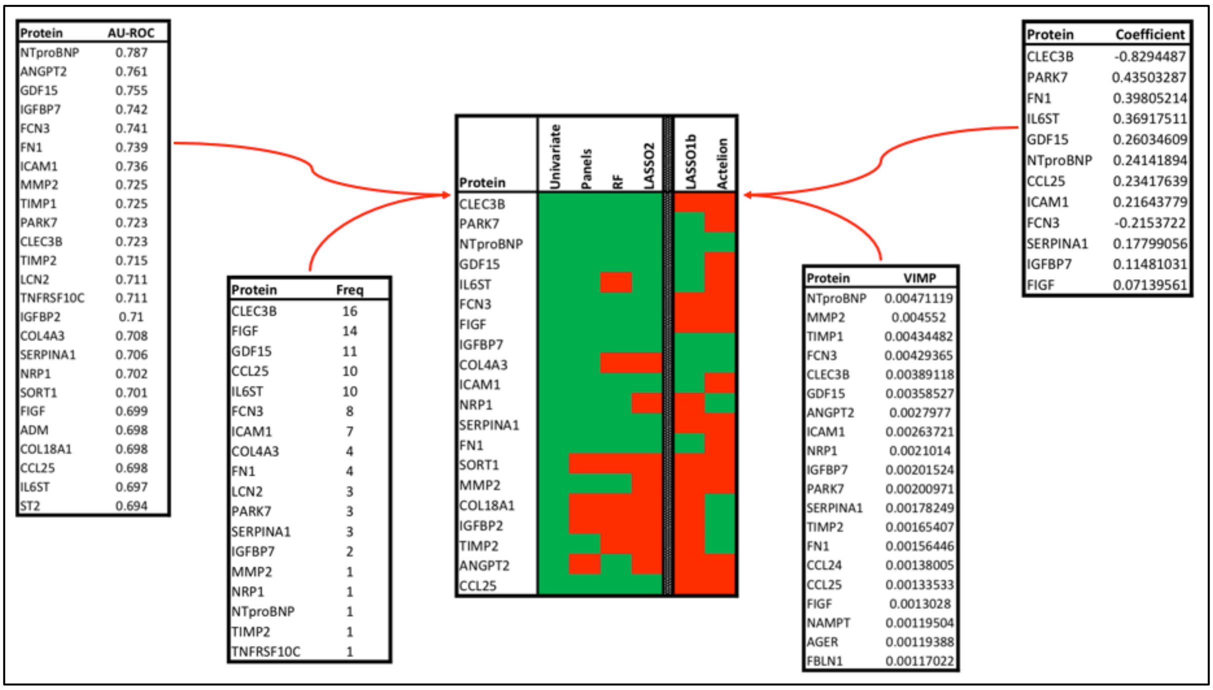
Figure 4.8: LASSO model for Myriad2 analysis

LASSO model returned from 24 proteins entered into analysis for 88 patient samples.

The analysis returned a model 12 proteins in length, which for a larger sample set is likely to represent a model with reduced chance of overfitting.

4.5.3.5 Final protein selection

The results of the above analyses were considered to decide which proteins to take forward to classification modelling. The results were considered as they are displayed above, with the frequency of occurrence of proteins in the top panel building result taken from that method. We also considered the previous results from our Myriad1b analysis, and the results shared with us which had been produced by collaborators at Actelion pharmaceutical (Figure 4.9).



Protein	AU-ROC
NTproBNP	0.787
ANGPT2	0.761
GDF15	0.755
IGFBP7	0.742
FCN3	0.741
FN1	0.739
ICAM1	0.736
MMP2	0.725
TIMP1	0.725
PARK7	0.723
CLEC3B	0.723
TIMP2	0.715
LCN2	0.711
TNFRSF10C	0.711
IGFBP2	0.71
COL4A3	0.708
SERPINA1	0.706
NRP1	0.702
SORT1	0.701
FIGF	0.699
ADM	0.698
COL18A1	0.698
CCL25	0.698
IL6ST	0.697
ST2	0.694

Protein	Freq
CLEC3B	16
FIGF	14
GDF15	11
CCL25	10
IL6ST	10
FCN3	8
ICAM1	7
COL4A3	4
FN1	4
LCN2	3
PARK7	3
SERPINA1	3
IGFBP7	2
MMP2	1
NRP1	1
NTproBNP	1
TIMP2	1
TNFRSF10C	1

Protein	Univariate	Panels	RF	LASSO2	LASSO1b	Actelion
CLEC3B	🟩	🟩	🟩	🟩	🟥	🟥
PARK7	🟩	🟩	🟩	🟩	🟩	🟥
NTproBNP	🟩	🟩	🟩	🟩	🟩	🟩
GDF15	🟩	🟩	🟩	🟩	🟩	🟥
IL6ST	🟩	🟩	🟥	🟩	🟩	🟥
FCN3	🟩	🟩	🟩	🟩	🟥	🟥
FIGF	🟩	🟩	🟩	🟩	🟥	🟥
IGFBP7	🟩	🟩	🟩	🟩	🟩	🟩
COL4A3	🟩	🟩	🟥	🟥	🟩	🟩
ICAM1	🟩	🟩	🟩	🟩	🟩	🟥
NRP1	🟩	🟩	🟩	🟥	🟥	🟩
SERPINA1	🟩	🟩	🟩	🟩	🟥	🟥
FN1	🟩	🟩	🟩	🟩	🟩	🟥
SORT1	🟩	🟥	🟥	🟥	🟥	🟥
MMP2	🟩	🟩	🟩	🟥	🟥	🟥
COL18A1	🟩	🟥	🟥	🟥	🟥	🟩
IGFBP2	🟩	🟥	🟥	🟥	🟥	🟩
TIMP2	🟩	🟩	🟥	🟥	🟥	🟩
ANGPT2	🟩	🟥	🟩	🟥	🟥	🟥
CCL25	🟩	🟩	🟩	🟩	🟥	🟥

Protein	VIMP
NTproBNP	0.00471119
MMP2	0.004552
TIMP1	0.00434482
FCN3	0.00429365
CLEC3B	0.00389118
GDF15	0.00358527
ANGPT2	0.0027977
ICAM1	0.00263721
NRP1	0.0021014
IGFBP7	0.00201524
PARK7	0.00200971
SERPINA1	0.00178249
TIMP2	0.00165407
FN1	0.00156446
CCL24	0.00138005
CCL25	0.00133533
FIGF	0.0013028
NAMPT	0.00119504
AGER	0.00119388
FBLN1	0.00117022

Protein	Coefficient
CLEC3B	-0.8294487
PARK7	0.43503287
FN1	0.39805214
IL6ST	0.36917511
GDF15	0.26034609
NTproBNP	0.24141894
CCL25	0.23417639
ICAM1	0.21643779
FCN3	-0.2153722
SERPINA1	0.17799056
IGFBP7	0.11481031
FIGF	0.07139561

Figure 4.9: Protein selection

Representation of variable selection. Middle panel shows heat map, green indicating the presence of protein in that variable selection method. Four white panels show results of the analyses on Myriad2 dataset. Abbreviations: Protein names: see appendix; Freq: frequency of occurrence of protein in panel results; VIMP: variable importance from random forest.

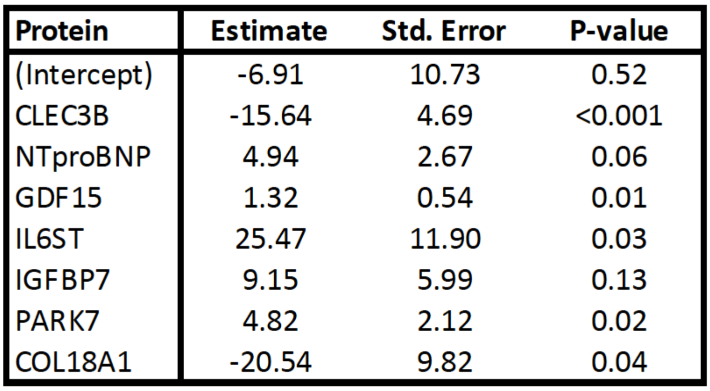
Major consideration was given to proteins identified by variable selection methods based on the Myriad2 dataset. Minor consideration was allowed for proteins identified in previous analyses and to results shared from Actelion pharmaceutical. Proteins taken forward to final classification modelling were CLEC3B; PARK7; NTproBNP; GDF15; IL6ST; FCN3; FIGF; IGFBP7; COL4A3; ICAM1; NRP1; SERPINA1; SORT1; MMP2; COL18A1; IGFBP2; TIMP2; ANGPT2; CCL25.

FN1 was identified but excluded due to a high level of data outside the limits of detection (25%).

4.5.4 Modelling

Prior to statistical modelling all protein variables were log transformed and scaled. The 19 selected variables were then entered into logistic regression, with immediate entry of the model into backward step-AIC optimisation. The AIC metric gives a composite of an estimation of model "goodness of fit" to the derivation data, but penalises increasing model length, designed to favour a model more likely to generalise. A model of increased size with very good fit to the derivation data is more likely to be over fit and to closely describe the derivation data only and less likely to generalise and function as a true predictor. The lower the AIC metric, the better the predicted model performance. The step-AIC process reduced the model to from 19 to 7 protein length (Table 4.5).

Table 4.5: Logistic regression model after step-AIC optimization

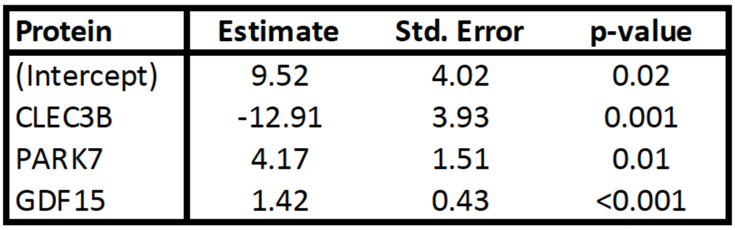


Protein	Estimate	Std. Error	P-value
(Intercept)	-6.91	10.73	0.52
CLEC3B	-15.64	4.69	<0.001
NTproBNP	4.94	2.67	0.06
GDF15	1.32	0.54	0.01
IL6ST	25.47	11.90	0.03
IGFBP7	9.15	5.99	0.13
PARK7	4.82	2.12	0.02
COL18A1	-20.54	9.82	0.04

This 7 protein model suggested by AIC optimisation is the best possible using the AIC metric, as when shortened further, the detrimental effect on "goodness of fit" exceeds the gains from reducing model length and the AIC metric increases.

The model given retains several variables which do not reach statistical significance when modelled with the other proteins included. It is likely therefore that these proteins yield minimal extra information gain in the model and as such a further round of constraint was conducted, removing non-significant protein variables and remodelling to re-evaluate the contribution of those remaining. After this process a final model consisting of only 3 proteins remained (Table 4.6).

Table 4.6: Final predictive model for classifying for PAH in SSc



Protein	Estimate	Std. Error	p-value
(Intercept)	9.52	4.02	0.02
CLEC3B	-12.91	3.93	0.001
PARK7	4.17	1.51	0.01
GDF15	1.42	0.43	<0.001

Tetranectin (CLEC3B), Protein DJ-1 (PARK7) and Growth differentiation factor 15 (GDF15) are the component proteins in the final model. This result from logistic regression represents an open, easily understood model that can be applied easily to new patient samples without the need for complex computing methodologies such as would be needed to make predictions from random forest algorithms to predict classification.

4.5.5 Derivation cohort results and classifying statistics

Performance of this model for the prediction of PAH in SSc is demonstrated graphically in (Figure 4.10a).

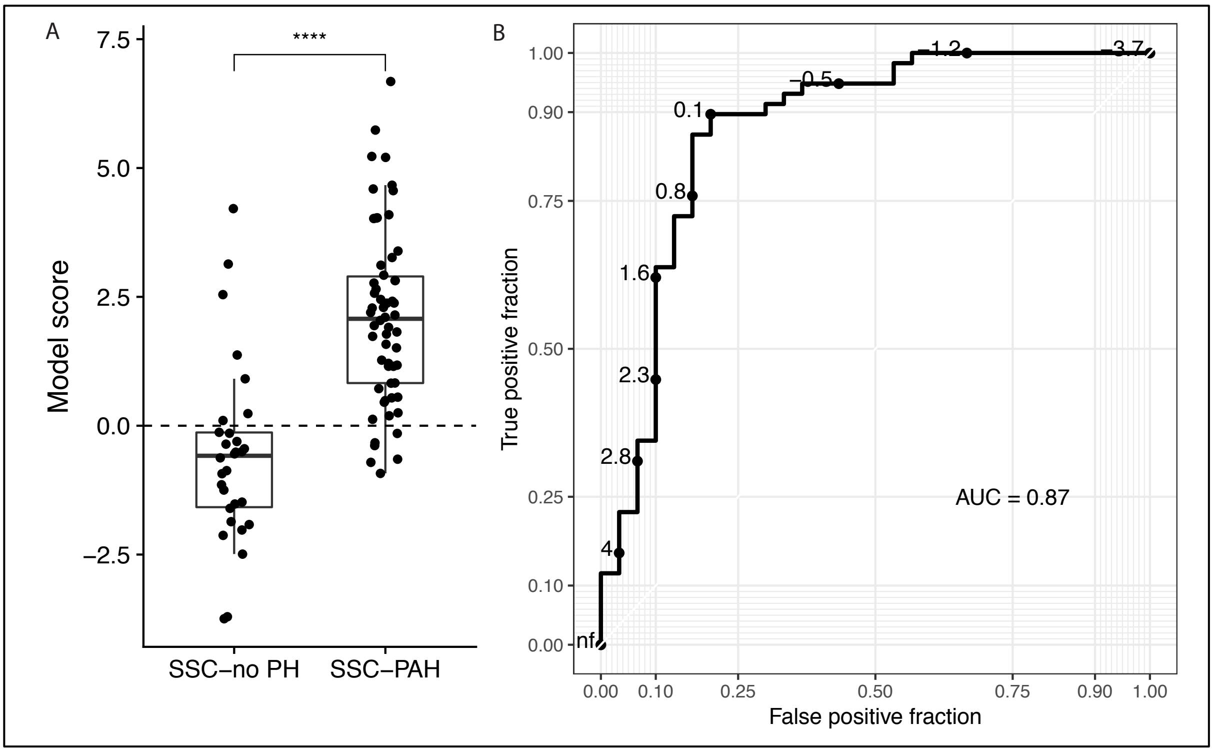
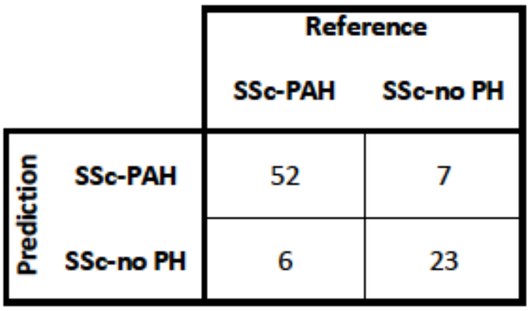


Figure 4.10: Model scores and ROC curve for the classification of PAH in SSc from derivation cohort

A: Model scores plotted for each patient in known diagnostic classes. Horizontal line represents model cut-off for classification into groups to separate true and false predictions. B: ROC curve for classification based on scores in A.

Based on figures from the corresponding confusion matrix (Table 4.7) this model predicts PAH in SSc with sensitivity 0.90, specificity 0.77, positive predictive value 0.88, negative predictive value 0.79 and an AU-ROC for classification of 0.87 (Figure 4.10b).

Table 4.7: Confusion matrix: outcome of derivation modelling



		Reference	
		SSc-PAH	SSc-no PH
Prediction	SSc-PAH	52	7
	SSc-no PH	6	23

Compared by AU-ROC for classification, the panel score is superior to any of its constituent protein variables (Figure 4.11).

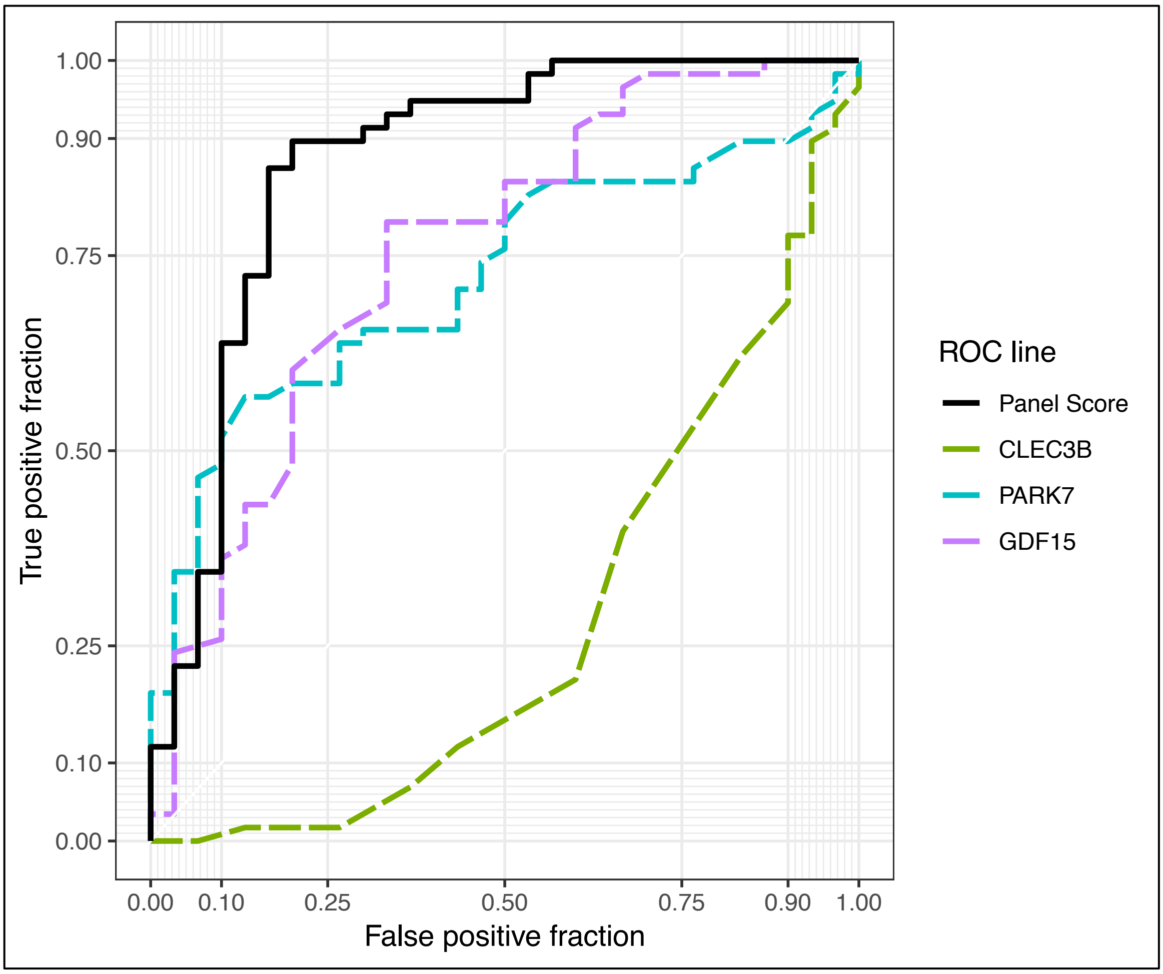


Figure 4.11: Comparison of panel ROC for classification against each component univariate curve

4.5.5.1 Correlation with clinical parameters

The panel score holds a moderate positive correlation with mPAP (r=0.56, p<0.001) and PVR (r=0.48, p<0.001) from RHC, and a moderate negative correlation with percent predicted TL_{CO} (r=-0.44, p<0.001) from PFTs (Figure 4.12).

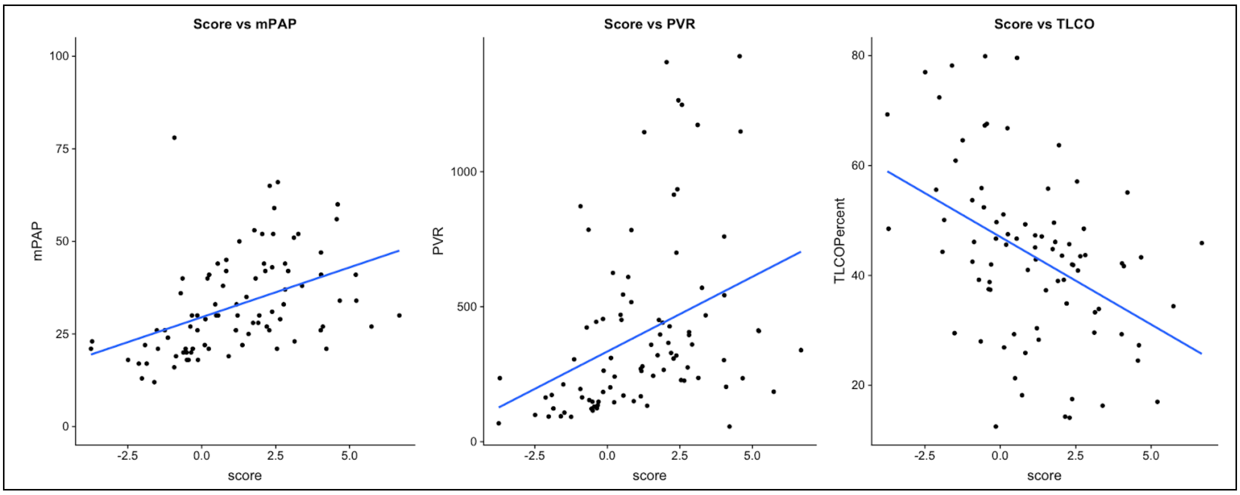


Figure 4.12: Correlation plots with clinical parameters

Line represents the line of best fit for the correlation.

4.5.5.2 Survival analysis

Survival analysis was performed after patient survival information was obtained with a censor point set at the time of the final data update on 29/11/2018. Kaplan Meier analysis of the disease arm only (SSc-PAH)(Figure 4.13), grouped above and below the median panel score shows a trend towards a significant prognostic utility for this panel, however due to low numbers of events this does not achieve statistical significance with a Log-Rank p=0.067.

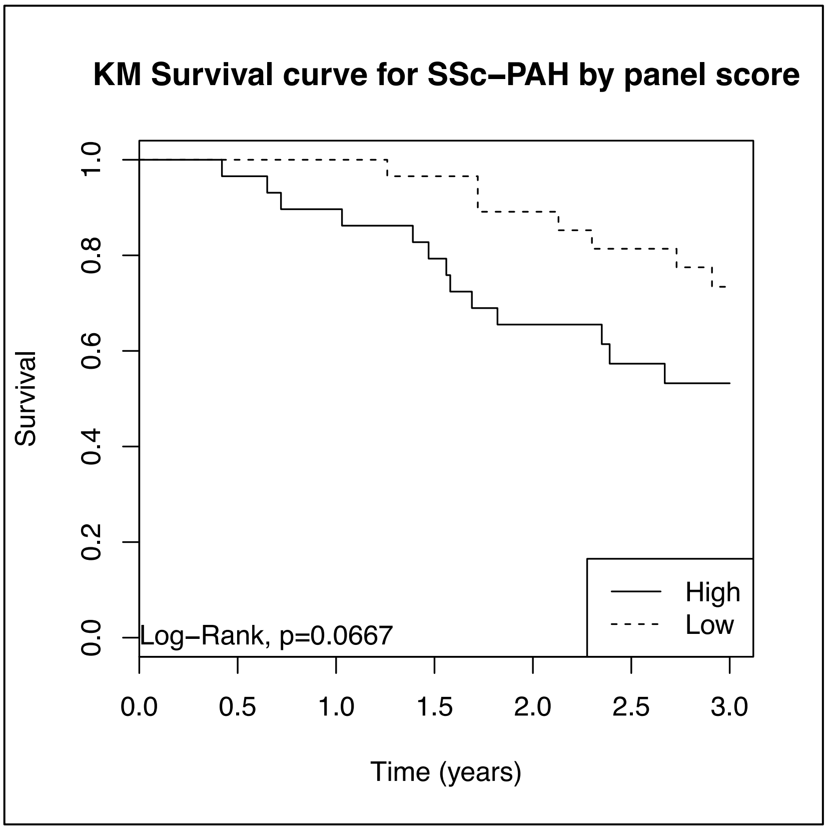


Figure 4.13: Kaplan Meier survival curve for panel score

Survival curve for patients with SSc-PAH split at the median panel score into high and low groups.

Cox regression analysis was used to further assess whether this model score can predict survival, when adjusted for age, gender and ethnicity. This produced a model which significantly predicted survival (p=0.02), with our panel score and gender retained as the significant predictors (Table 4.8).

Table 4.8: Cox regression analysis to include panel score as predictor

Variable	**Co-efficient**	**p value**
Panel score	1.28	0.03
Age	1.03	0.17
Gender (male)	2.61	0.01
Ethnicity	3.4	0.26

Results of Cox regression analysis showing co-efficients and significance levels of variables entered into the model.

4.5.6 Comparison of derivation stats with alternative screening tools

The majority of patients included in this study were recruited to the Sheffield PH biorepository prior to the publication and routine use of the DETECT score, and furthermore as a tertiary referral centre receiving referrals from rheumatology centres rather than making them, not all datapoints needed to complete the DETECT criteria were collected for each patient. To handle missing data the two step DETECT nomogram (Coghlan et al., 2014) was used to assign the most extreme value for each datapoint to favour the known disease classification. By doing so this will potentially lead to a DETECT score for patients which favours the known clinical classification and would lead to improved screening statistics.

When our panel statistics are compared to the statistics calculated from the application of the DETECT protocol and ERS echocardiography screening guidelines to the phenotype data from our derivation cohort, our panel competes favourably in screening for PAH in SSc with a diagnostic accuracy of 0.86, 0.74 and 0.85 for the ERS, DETECT and protein panel respectively (Table 4.9).

Table 4.9: Comparison of protein panel with currently used screening tools

	ERS 2015	**DETECT**	**Sheffield Panel**
True Positive	52	56	52
True Negative	18	7	23
False Positive	7	22	7
False Negative	4	0	6
Sensitivity	0.93	1.00	0.90
Specificity	0.72	0.24	0.77
PPV	0.88	0.72	0.88
NPV	0.82	1.00	0.79
Accuracy	0.86	0.74	0.85

Statistics to compare the results of applying screening models to the derivation dataset. Insufficient data for 7 patients for ERS echocardiography screening and 3 patients for DETECT screening. Abbreviations: PPV: Positive predictive value; NPV: negative predictive value.

4.5.7 Derivation score alone vs Derivation score with DETECT variables included

To investigate whether the variables from the DETECT screening tool could improve on the predictive accuracy of our protein only model, each DETECT variable was added to our protein variables and modelled with logistic regression. For the purposes of statistical modelling it was not practical to assign values to missing datapoints, as this would significantly alter the model coefficient and any statistical significance which is what we are specifically looking for this this process (in contrast to the comparison of statistics in section 4.5.6 where we presented the best possible DETECT model statistics). Due to a high level of missing data, presence or absence of telangiectasia, and serum uric acid levels were not modelled. Due to a low frequency of missing data, but for each arising for different patients, the remaining DETECT clinical variables were modelled individually with the protein data to minimise the loss of samples from the analysis due to those missing data.

Right atrial area, anticentromere antibody (ACA) and right axis deviation were each individually modelled alongside CLEC3B, PARK7 and GDF15 to analyse whether they improve on the predictive accuracy of the model. Right atrial area entered the model with a statistically significant coefficient (0.001, p=0.009) however did not alter the significance level for our protein variables. Neither ACA nor right axis deviation entered the model. The predictive accuracy of the model was not improved beyond that of the protein only model by any of these variables (0.84, 0.87 and 0.87 respectively).

FVC%/TLCO% significantly entered the model (Coefficient 1.82, p=0.004), however again did not replace any protein variables. There was a marginal improvement in the AU-ROC for classification with this model as compared to our protein only model (0.92 vs 0.87 respectively), and a small improvement in predictive accuracy (0.9).

The only variable to significantly alter the predictive accuracy of our protein only model was the TR jet velocity, which when modelled with the protein data proved strongly significant (Coefficient 6.06, p=0.005) and replaced GDF15 and PARK7 as significant variables in the model. CLEC3B concentration remained a significant predictor. The inclusion of TR jet velocity improved the model with AU-ROC (0.97) and predictive accuracy of the model (0.92).

It should be noted that although echocardiographic parameters are used as part of the published DETECT protocol, these more diagnostic metrics should be expected to strongly associate with prediction of the presence of pulmonary hypertension. TR jet velocity in particular is a direct marker of pulmonary hypertension and in areas where gold standard invasive right heart catheterisation is not available, has been suggested as a non-invasive diagnostic alternative.(Parasuraman et al., 2016, Sohail et al., 2019) In progressing the development of our predictive tool, we sought to avoid the inclusion of these more diagnostic variables.

4.6 Discussion

Earlier analyses found issues with an imbalance in comorbidities confounding the predictive model, and then with model overfitting due to a combination of a small sample size and overly complex models. This analysis was designed to overcome these problems by increasing the sample size, keeping the sample balanced for comorbidities, and applying much harsher constraint on model size to produce a model of good fit which can classify an external validation cohort. An additional 28 patient samples were added to the analysis having been identified by an up to date query from the Sheffield PH biorepository. The cohort analysed remains realistic to that which would be encountered in a rheumatology clinic with a balanced 36% of patients with ILD in each arm of the analysis.

Serum samples from the newly identified patients were analysed by Myriad RBM as a second assay batch, but both internal controls at Myriad RBM and batch control analysis on the protein data received found no evidence of a batch error in the dataset beyond differing limits of detection between assays.

Given the results in Chapter 3, it is clear that model overfitting represents the greatest threat to the accuracy of our predictive model with a persisting imbalance between the number of predictors and the sample size. As such we altered our approach to variable selection and classification modelling, uncoupling these and using more rigorous techniques to constrain the final model size. Variable selection was performed using a composite analysis of the output from machine learning and univariate statistics, with classification modelling performed by logistic regression with several optimisation steps. This approach allows for greater transparency and understanding of the proteins retained in the model, and of those that are subsequently dropped, through the standard availability of a coefficient and statistical significance level for each protein at each stage, data which was not readily available when models generated by machine learning alone were used for prediction.

Our final model consists of only 3 proteins; Growth differentiation factor 15 (GDF15), Tetranectin (CLEC3B) and Protein DJ-1 (PARK7). Published data regarding these proteins will be discussed in the introduction to Chapter 4, however a model made up of these proteins seems acceptable, with each of the three proteins linked to different underlying processes in the pathogenesis of pulmonary vascular disease (Figure 4.14).

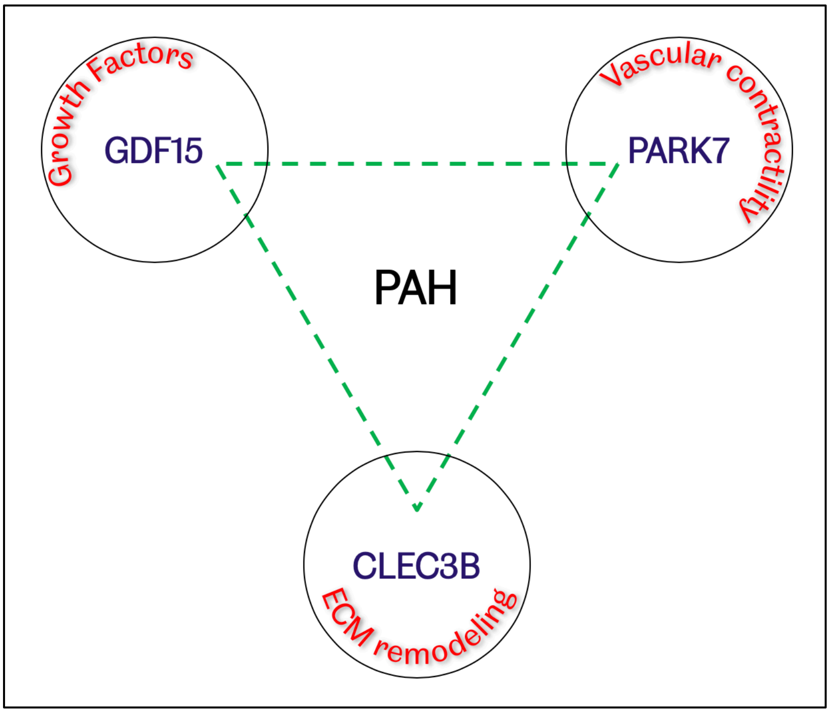


Figure 4.14: Suggested role for model proteins in pathophysiological processes of pulmonary vascular disease

Abbreviations: PAH: Pulmonary arterial hypertension; ECM: Extracellular matrix; Protein names: see appendix.

Scores produced by the model for patients in the disease and control cohorts show a statistically significant difference and a good AU-ROC for classification (0.87) which is superior either to the univariate curve for any of the component proteins within the model, or indeed for any univariate protein from the entire dataset. Furthermore, this protein only model was not significantly improved when remodelled to including the available variable data from the DETECT model, with the exception of TR jet velocity – a direct estimate of PA pressure. My model predicted PAH in SSc from our derivation cohort with a diagnostic accuracy comparable to the alternative methods currently used in clinical practice. It represents an opportunity for screening to be conducted in the SSc patient population based on a single blood draw, which may prove more acceptable to patients, and may produce a screening result within a much shorter timescale when compared to current multimodality testing. It also has the potential to widen access for screening to areas where there may not be ready access to-, or where there may be a significant delay to- other investigation modalities such as pulmonary function tests, or echocardiography.

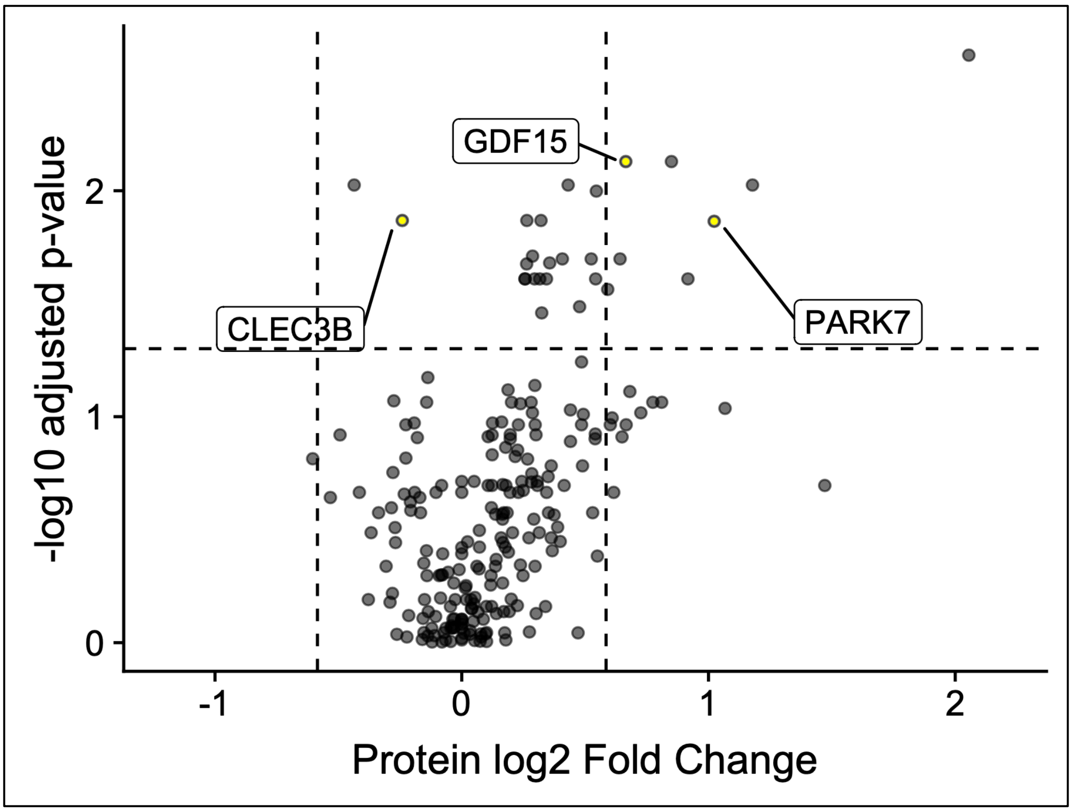


Figure 4.15: Volcano plot demonstrating selected proteins

Figure 4.6 replotted to identify proteins from classification model.

Research question 3 asks whether any predictive model generated could translate from a simple prediction of disease only, to a tool which could function to both classify and demonstrate some prognostic utility. There is a clear difference in survival evident between the disease and control cohorts of my study as would be expected given the significant life limiting nature of a diagnosis of PAH.(Hurdman et al., 2012) In order to assess whether a difference in model score could predict survival, only the scores for patients with SSc-PAH were analysed against time to all-cause mortality. My model score shows a trend possibly suggestive of a relationship to survival, however this did not reach statistical significance and warrants further investigation in a larger patient cohort.

The main limitation of this analysis is that the population being studied are those who have already been referred for investigation for pulmonary hypertension, rather than an unselected general rheumatology SSc population. These patients are therefore preselected, and have a higher pre-test probability of pulmonary hypertension than a general SSc cohort. This is likely to have a disproportionate effect on the control cohort. The control arm in our study consists of a group of patients who have been referred for investigation for some reason, whether this be symptom burden, or abnormal investigations, but this raises the possibility that this group does not accurately represent the wider SSc-no PAH population. The effect of this can be seen in the available demographics, with modest pulmonary pressures existing in this group, and pulmonary function tests showing an unexplained abnormality of TLCO. This could be a cause of a lower true negative fraction in our derivation statistics which could be improved by applying the model to an unselected control cohort.

The aim of the model is to predict PAH in SSc at an early stage to allow for early intervention, however by studying patients who have awaited referral for investigation, we have potentially a cohort of patients with early stage, but already established disease.

4.7 Summary

This concludes the large data analysis, machine learning and variable selection section of my project, having determined that a model with three proteins only from my derivation cohort is the optimal one both for accuracy of detection of PAH in the derivation cohort, and to minimise the risk of overfitting with a small derivation cohort. In the chapters which follow, I will take this model forward to validation both using internal validation methods, and validation in an external cohort of patients. I will then examine some of the cell biology relevant to these protein targets.

5 Validation

5.1 Introduction

We have generated a predictive model consisting of 3 proteins – GDF15; CLEC3B and PARK7. Statistics presented so far have been those based on applying the model to the derivation dataset to predict whether each patient has PAH, and based on this to generate statistics against the known patient disease subgroup. The given patient cohort has therefore both been used to generate and to test the model this is likely to favour our model, and may represent a model overfit to the derivation dataset.

In this chapter we will test whether our model will generalise through three different validation methods: k-fold cross validation using the derivation dataset; validation in an alternative Sheffield PAH subgroup; and validation in an external cohort of patients.

5.2 K-fold cross validation

k-fold cross validation is a technique used to test the performance of a model when only the single dataset used to derive the model is available. It was developed to help distinguish between overfit models which lack stability and models which are likely to likely to represent a true prediction. K-fold cross validation involves splitting the dataset into a number of subsets (k folds), and deriving the model on k-1 folds, and testing the model on the reserved data. At each iteration the same protein variables are used, but the logistic regression coefficients for each protein re-calculated and tested. The model is then repeated until all folds have been used as the reserve and average the result. The cross validation procedure can be repeated a large number of times using a randomised approach to splitting the data each time to enhance model testing.

To test our classification model I entered the derivation dataset (88 samples) and the protein variables (3 proteins) into k-fold cross validation (package: caret), opting for 10 fold, and repeating the cross validation procedure 200 times. The cross validation returned an average diagnostic accuracy of 0.83. This suggests a stable model which therefore has a lower likelihood of model overfitting when compared against the 0.85 diagnostic accuracy given when the model is applied simply back across the whole derivation dataset (Table 4.9).

5.3 Validation in other PAH subtypes

Within the initial cohort of samples sent to Myriad RBM for assay (Table 3.1) were a cohort of patients with idiopathic PAH (IPAH) (n=30) and healthy volunteers (HV) (n=29). Serum samples from these patients were analysed on the same platform, randomised on the same plates, as our final classification model and can therefore act as a control for any variation in assay procedure.

Although derived only in a cohort of patients with SSc, we sought to investigate whether the model would predict PAH in a cohort of patients with IPAH from HV. These alternative patients were not used in the derivation of our classifying model at any stage.

Applying the classifying model to the IPAH/HV cohort classifies PAH in this cohort with an AU-ROC for classification 0.87 (Figure 5.1).

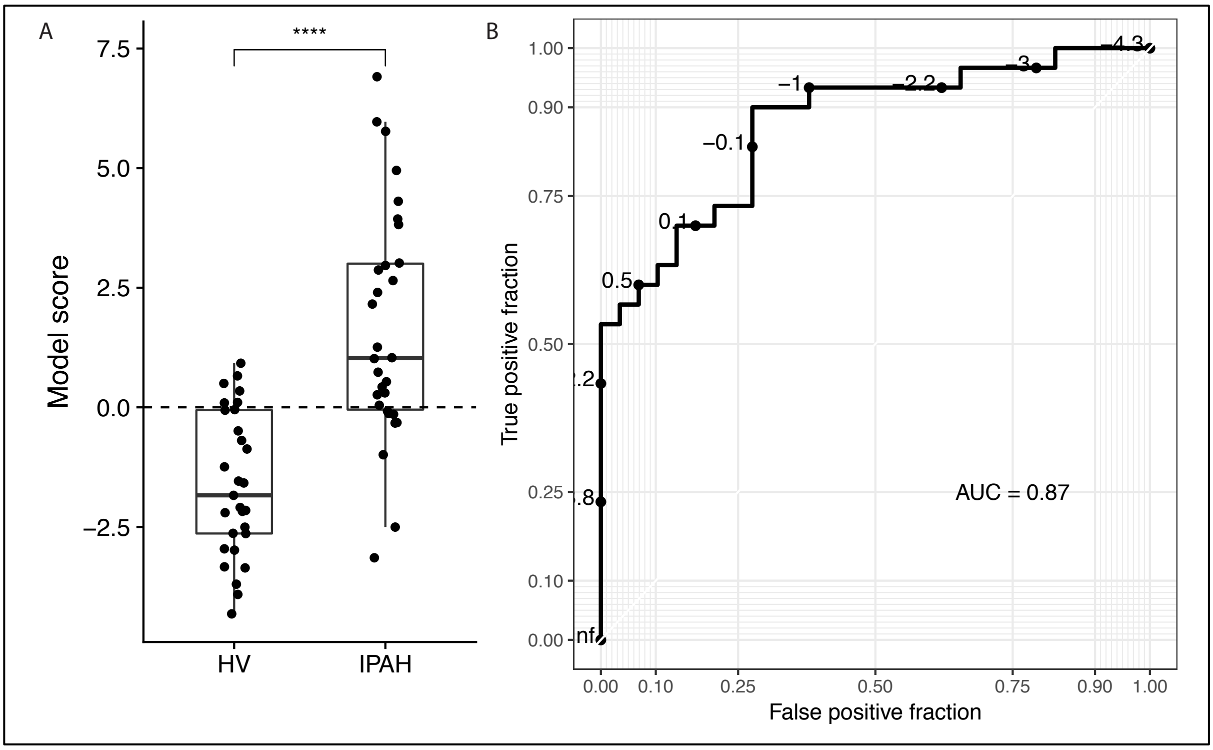


Figure 5.1: Application of classifying model to IPAH/HV cohort.

Result of applying the classifying model, derived in the SSc cohort to a cohort of patients with IPAH and HV.

A: Jitter plot with model scores, B: ROC for classification.

As a validation test of our classifying model, this result supports the model as a classifier of PAH, and furthermore suggests that this model may predict PAH in general, rather than specific only to the SSc specific cohort. This warrants further future investigation to confirm.

5.4 External validation

5.4.1 Validation cohort

An external validation cohort of patients including those with SSc both with and without PAH were identified and received from the Vera Moulton Wall Center for Pulmonary Vascular Disease, Stanford University School of Medicine and the Pulmonary Vascular Center, Vanderbilt University Medical Center. Samples from Stanford were obtained at RHC, and those from Vanderbilt from peripheral blood draw at the time of clinical encounter. All blood samples were clotted, centrifuged and aliquoted and stored at -80°C until required. Serum samples were shipped to us on dry ice and confirmed frozen, in good condition, and uniquely identifiable upon receipt. Samples were stored in our laboratory at -80°C while awaiting validation assay.

Table 5.1: Validation cohort demographics

		SSC-PAH	SSC-no PH	p-value
n		53	25	
Age (IQ range)		61 (54 - 69)	58 (51 - 65)	0.39
Gender (M/F)		5/48	2/23	1
WHO FC	1	3	Unknown	NA
	2	11		
	3	28		
	4	8		
PFTs (median + IQ))	FVC Percent	79 (64.5 - 86.5)	Unknown	NA
	TLCO Percent	64 (39.8 - 88.3)		
6MWT (median + IQ	Distance	458 (313 - 1201)	1425 (1135 - 1800)	<0.001
RHC (median + IQ)	mPAP	46 (34 - 54)	16 (15 - 20)	<0.001
	RA pressure	7 (4.5 - 13)	3 (3 - 7)	<0.001
	CI	2.2 (1.8 - 2.6)	2.4 (1.9 - 3.0)	0.33
	PVR	700 (442 - 986)	131 (107 - 214)	<0.001
	PCWP	10 (8 - 13)	7 (5 - 10)	0.017

Demographics for combined Stanford and Vanderbilt validation cohort.

Abbreviations: n – number; IQ – interquartile; M/F – Male/Female; WHO FC – World health organisation functional class; PFT – Pulmonary function tests; 6MWT – 6-minute walking test; RHC – right heart catheter

5.4.2 Validation assays

To retain the option from a larger group of proteins, we performed validation assays on the external samples targeting the seven proteins identified in the first step of classification modelling (Table 4.5): Tetranectin (CLEC3B); N terminal pro b-type natriuretic peptide (NTproBNP); Growth differentiation factor 15 (GDF15); Interleukin-6 receptor subunit beta (IL6ST); Insulin like growth factor binding protein 7 (IGFBP7); Protein DJ-1 (PARK7); Endostatin (COL18A1).

Validation assays were performed according to the methods described in section 2.5 (Validation assays). As the Myriad RBM assays are based on Luminex technology we initially sought to perform the validation assays using similar technology in house. CLEC3B was not available in a Luminex assay and we therefore performed this analysis using an ELISA assay. NTproBNP did not plex with the other proteins and was performed using a single analyte

Luminex assay. The remaining proteins (GDF15, IL6ST, IGFBP7, PARK7 and COL18A1) were combined in a single Luminex assay.

The Stanford University samples were received first and analysed separately to the Vanderbilt University samples. The Stanford samples required two plates for each assay, and the Vanderbilt samples were all plated on a single plate. For the Stanford assays split across two plates, each plate assay was performed concurrently.

The Luminex NTproBNP assay failed on all plates, so a further ELISA assay was performed for NTproBNP. This further ELISA assay also failed quality control and so as NTproBNP could not be reliably validated, it was not considered for inclusion in the final model.

5.4.3 Quality control and intra-assay batch control for validation assays

All samples were plated in duplicate to assess for any intra-assay variability. Duplicates for all plates and all assays were examined and showed little evidence of significant intra-assay variability when assessed visually (a representative example is shown in Figure 5.2, with all plots in Appendix 3 – External validation QC, Sample replicate plots). Statistical analysis of intra and inter assay variability confirm the quality control of the assay used (Table 5.2).

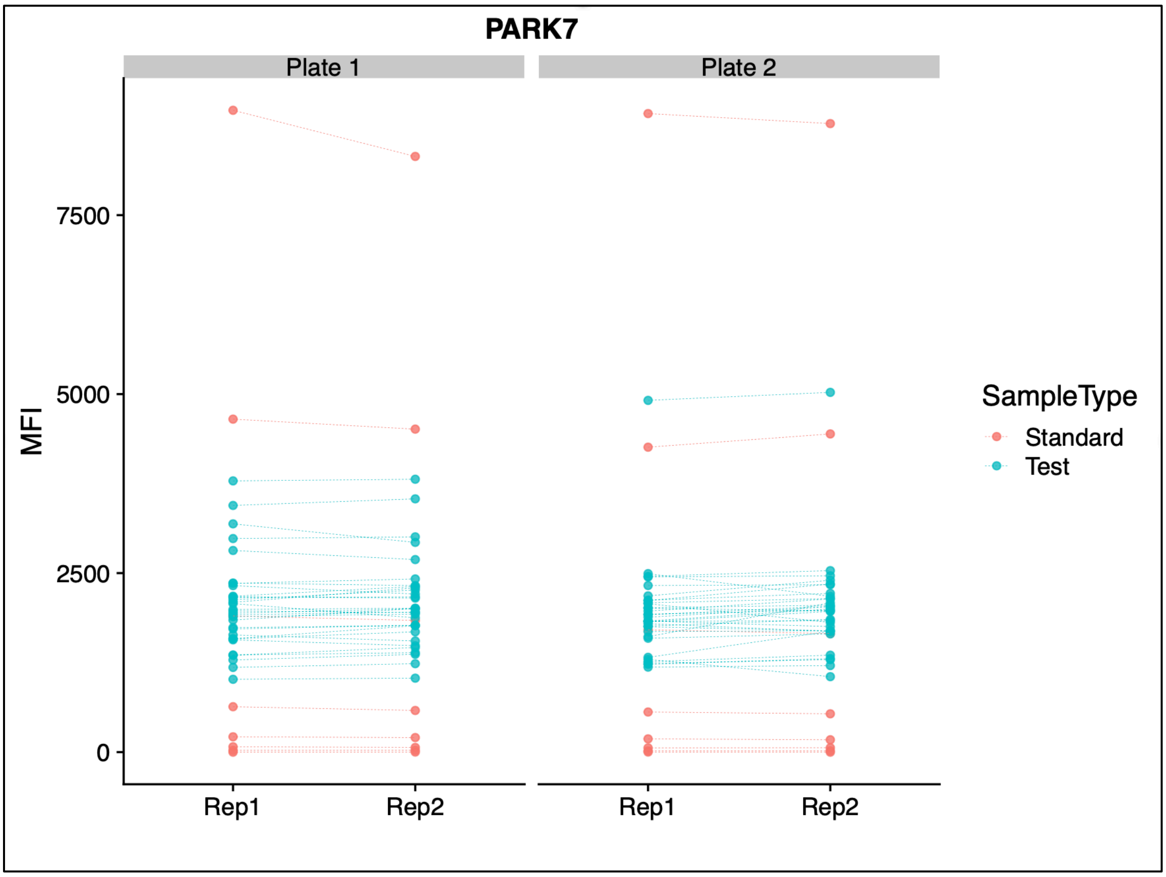


Figure 5.2: Dot plot showing MFI values for each replicate

Dot plot showing raw MFI values for replicates 1 and 2 from Luminex assay for PARK7 in Stanford University samples. Dotted lines join replicates from the same patient.

As the Stanford University samples needed to be analysed on two separate plates, run at the same time under the same environmental condition, this gave the opportunity to examine the standard curves across these plates as a further measure of intra-experiment consistency. The paired standard curves replicated each other precisely suggesting little if any intra-experiment variability. A representative example is given in Figure 5.3 and all standard curve analyses of the Stanford plates shown in Appendix 3 – External validation QC, Stanford assay plate standard curves. Intra- and inter-assay variability co-efficients (CV) are given in Table 5.2.

Table 5.2: Co-efficients of variation (intra and inter-assay) for in-house validation assays

Protein assay	**Intra-assay CV %**	**Inter-assay CV %**
CLEC3B	4.11	1.49
COL18A1	6.94	4.56
GDF15	7.1	8.93
IGFBP7	12.00	6.67
IL6ST	4.83	1.89
PARK7	6.99	3.99
NTproBNP (luminex)	21.84	17.7

Intra-assay CV %: variability between standards and anchor samples across different plates run at the same time. Inter-assay CV %: variability between duplicate values across whole assay.

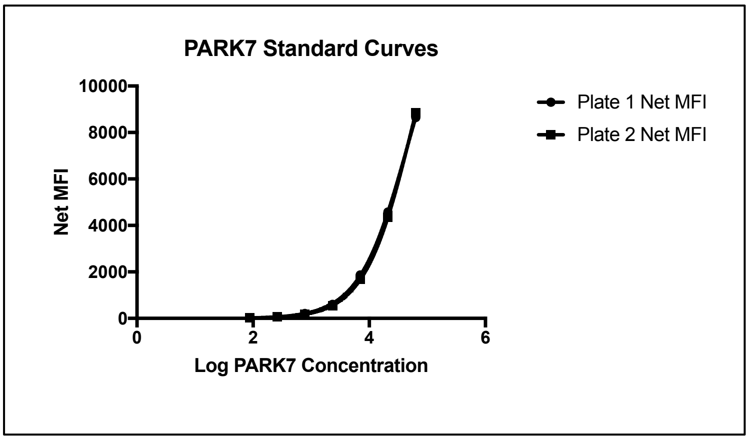


Figure 5.3: Standard curves plotted from the two Stanford Luminex PARK7 assay plates

The standard curves for the Luminex assays suggested the assays to be sub-optimal at detecting throughout the documented range. By this I mean that for each assay, the lower standards tended to be clustered tightly with very little difference in measured MFI signal between each of the analyte concentrations in the lower range of the assay, until the third standard. This decreases the accuracy of calculated analyte concentrations for any measurements falling in this range. In order to examine the likely effect of this on the accuracy of calculated analyte concentrations, I assessed the actual distribution of sample MFI readings against their corresponding standard curve.

To determine the point at which our samples were detect on the standard curves we plotted the distribution of raw readings from ELISA and Luminex assays and demonstrated these against the corresponding standard curve for each plate. A representative example is given in Figure 5.4, with data for all proteins and plates analysed in Appendix 3 – External validation QC, Raw data against standard curves.

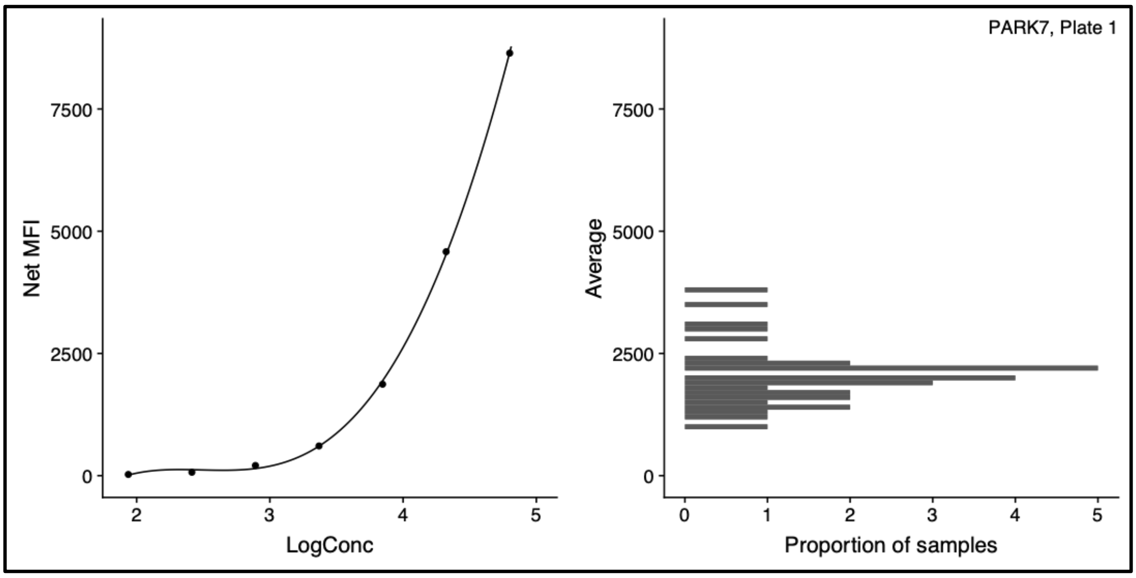


Figure 5.4: Standard curve and raw data histogram from Stanford samples, PARK 7, plate 1.

Standard curve (left) and histogram of raw data values from Luminex detection (right) to demonstrate the spread of data against the corresponding standard curve.

These data show that assays for CLEC3B, GDF15, IGFBP7, IL6ST, and PARK7 fall within the assay detection range. The assay for COL18A1 concentration detected a high proportion of data above the upper limit of the assay. The protein conversion for these data were accordingly revalued at the upper limit of detection. Our Luminex assay for NTproBNP failed on all plates, with assay results returning a net MFI lower than background through the majority of patient samples (Figure 5.5).

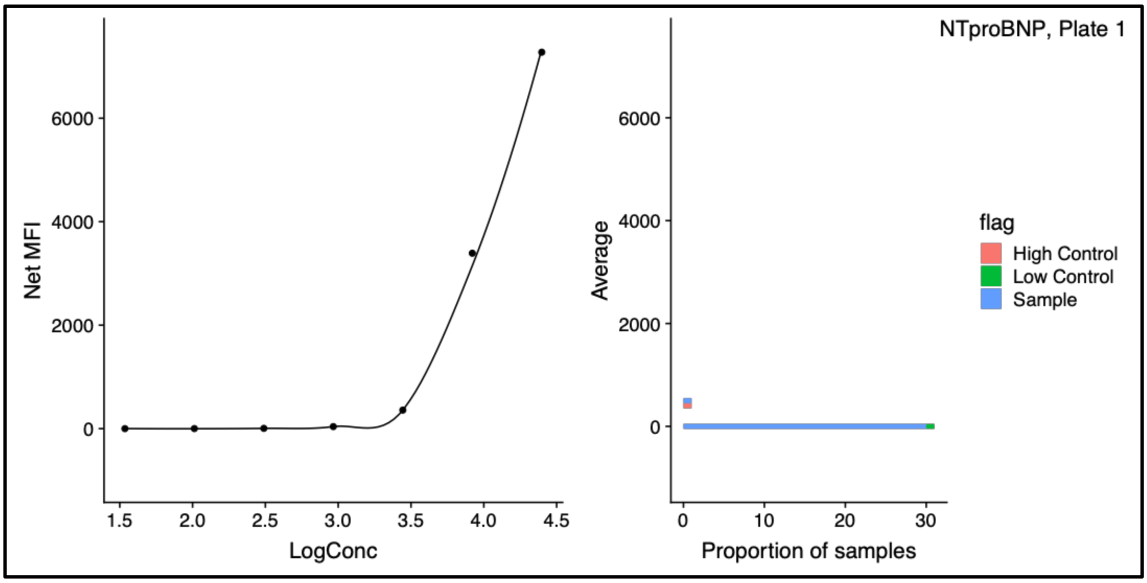


Figure 5.5: Raw MFI data and standard curve for NTproBNP Luminex assay.

Data shown is for Stanford samples, plate 1.

Protein concentrations were then generated from the raw detection values using the plate standard curve data, computed using GraphPad Prism 7.0a using a four parameter curve fit. Protein concentration data for points falling outside the assay detection range were assigned the value of the corresponding limit of detection.

Batch control was assessed using ‘anchor samples’ – Sheffield samples which were analysed on the Myriad RBM platform, also plated on every validation plate to assess for, and to allow for correction of inter-assay variability (Figure 5.6). When assessing protein concentration in the validation assays only (not considering Myriad concentration data) there was consistency demonstrated in the measurement across COL18A1, IGFBP7, IL6ST, NTproBNP and PARK7. There was evidence of a batch error in the measurement of CLEC3B and GDF15. Anchor proteins from Plate 1 Stanford samples for CLEC3B read significantly higher than the other validation plates for this protein. Furthermore, examination of the general protein concentration distribution confirmed the generally higher concentration results for Stanford Plate 1 (Figure 5.7a).

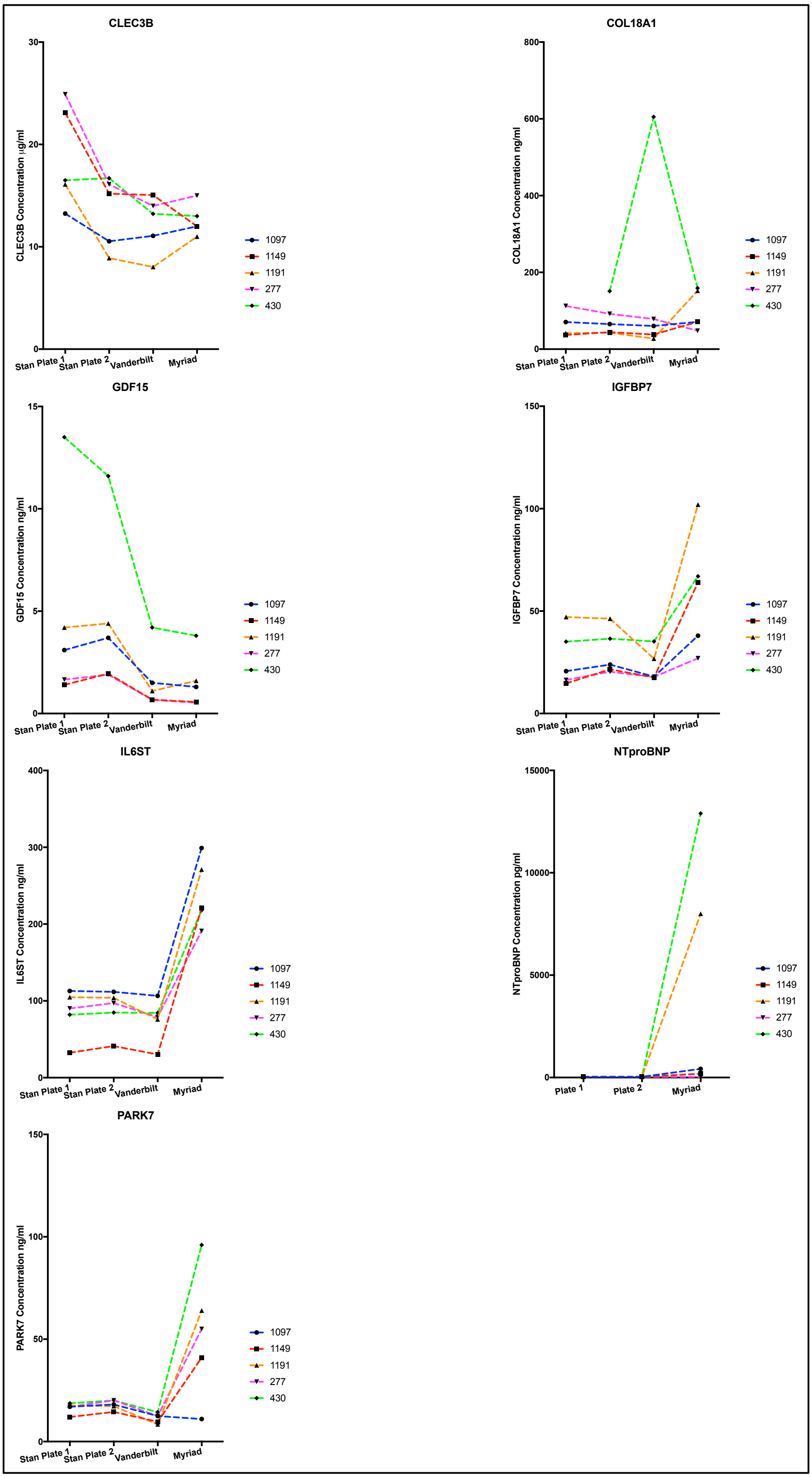


Figure 5.6: Anchor proteins on each plate by serum concentration

Anchor protein concentrations on each plate, and from Myriad dataset. Each coloured line represents an individual patient.

A correction factor was generated for each patient between Stanford plate 1 and the averaged values of Stanford plate 2 and Vanderbilt plate results. From these data a general correction factor was calculated by averaging the individual patient correction factors. After correction there was no significant difference between protein concentration distributions across the three validation plates (Figure 5.7b).

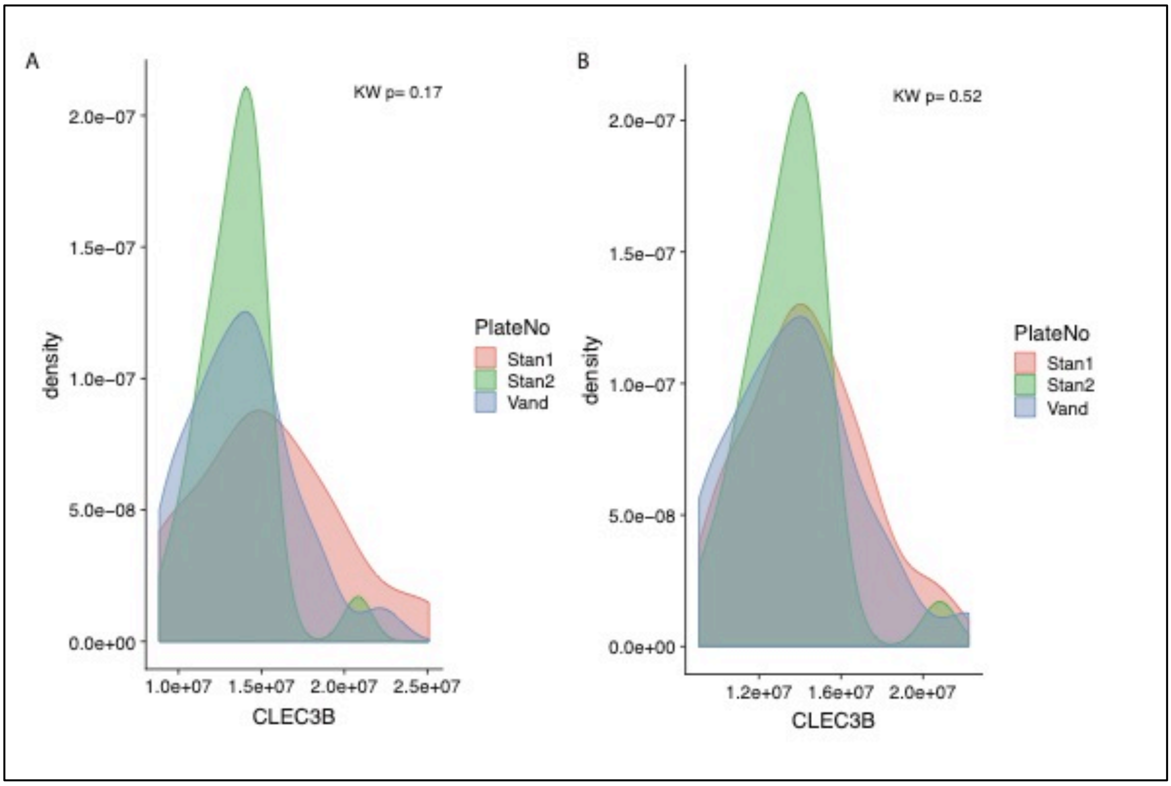


Figure 5.7: Density plot showing distribution of CLEC3B concentration measurements between plates. A: Before plate correction, B: After plate correction

Kruskal-Wallis p-value shown

Similarly, in analysis of GDF15, both Stanford plates read significantly higher than the Vanderbilt results when the Myriad RBM dataset values were taken into account (Figure 5.8a).

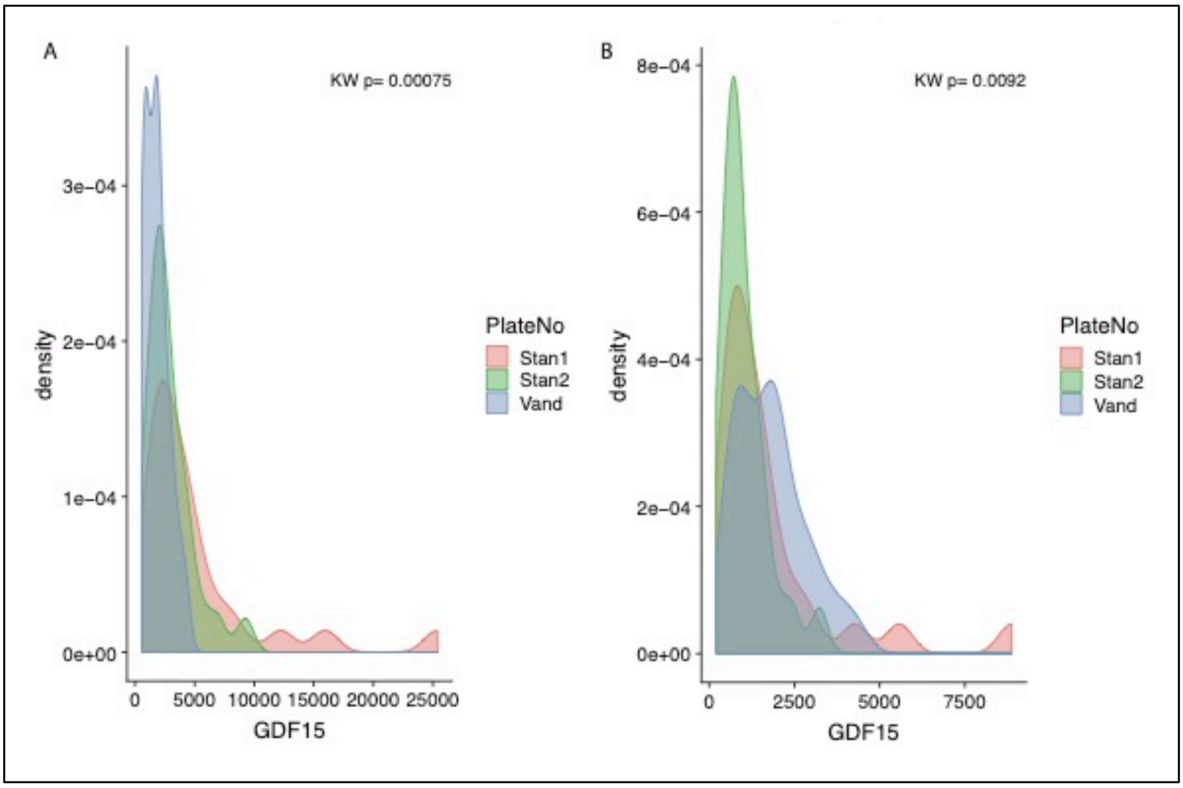


Figure 5.8: Density plot showing GDF15 concentration distributions on each plate. A: Before plate correction, B: After plate correction.

Kruskal Wallis p-value shown.

For this correction, the averaged results for each patient on the Stanford plates were corrected against each corresponding Vanderbilt value. The results were then averaged to generate a single correction factor which was applied to both Stanford plates.

This correction represents an improvement in plate variability, but does not completely eliminate the difference which may relate to biological differences between patient classes on each plate. Protein concentration plate comparison plots for all proteins after the above two corrections are applied are given in Appendix 3 – External validation QC, Protein concentration plate distribution plots.

5.4.4 NT-proBNP failed assays

Validation assays for NTproBNP were performed initially using Luminex technology, and subsequently by ELISA assay. Both assays failed quality control with very poor assay sensitivity for low NTproBNP concentrations, and the majority of sample results reading lower than background signal in our Luminex assays (Figure 5.9), and a high proportion samples outside the standard curve range in our ELISA assays (Figure 5.10).

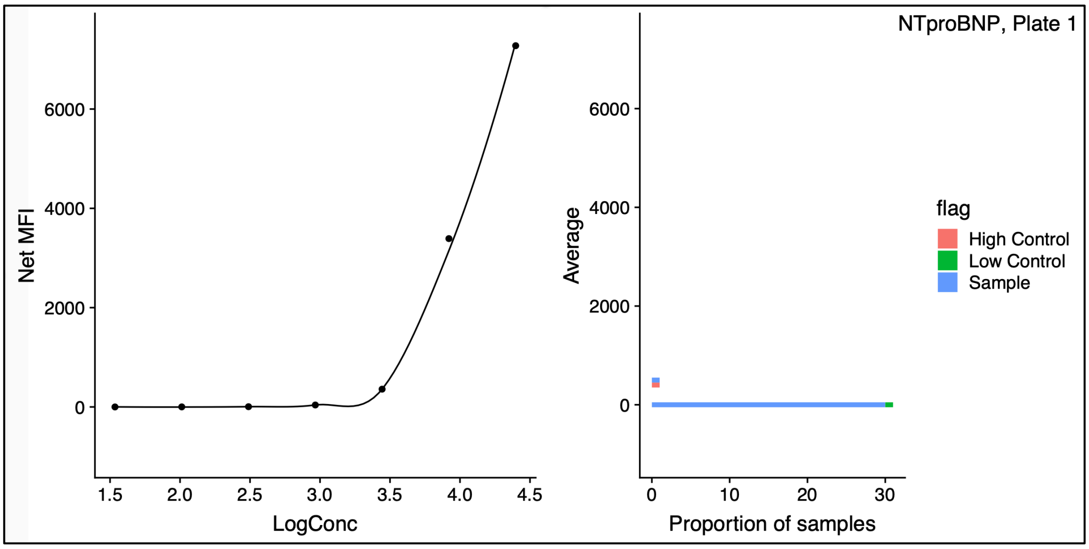


Figure 5.9: NTproBNP Luminex assay results for Stanford samples, plate 1.

Figure shows the standard curve on the left, and Net MFI readings for all samples on the right.

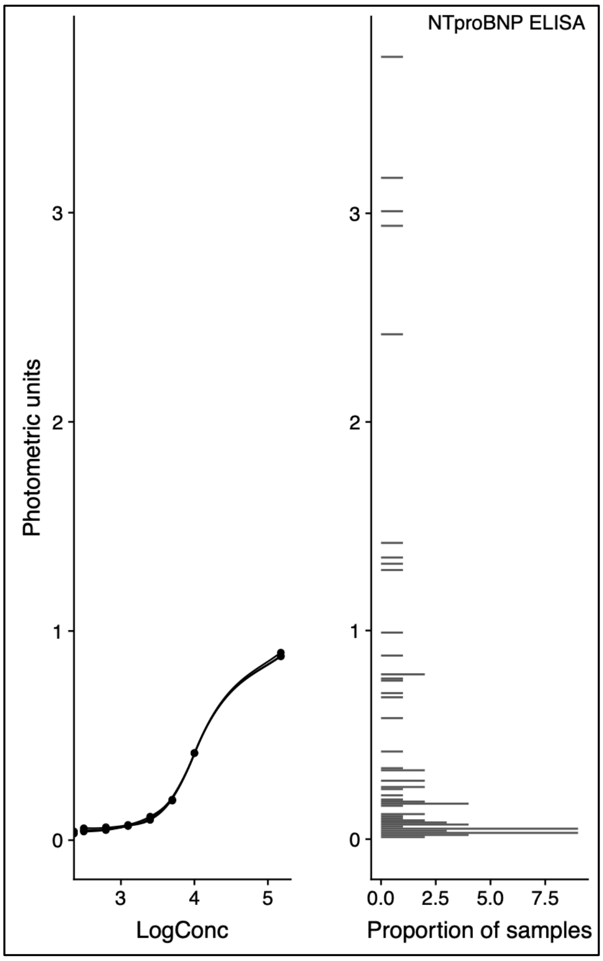


Figure 5.10: NTproBNP ELISA assays for all Stanford samples.

Standard curves for both plates on the left, with photometric detection units for all samples on the right.

When the NTproBNP ELISA results were compared to the clinical NTproBNP results available from Stanford university, taken and analysed at the time of patient visit, there was no correlation between measured concentrations.

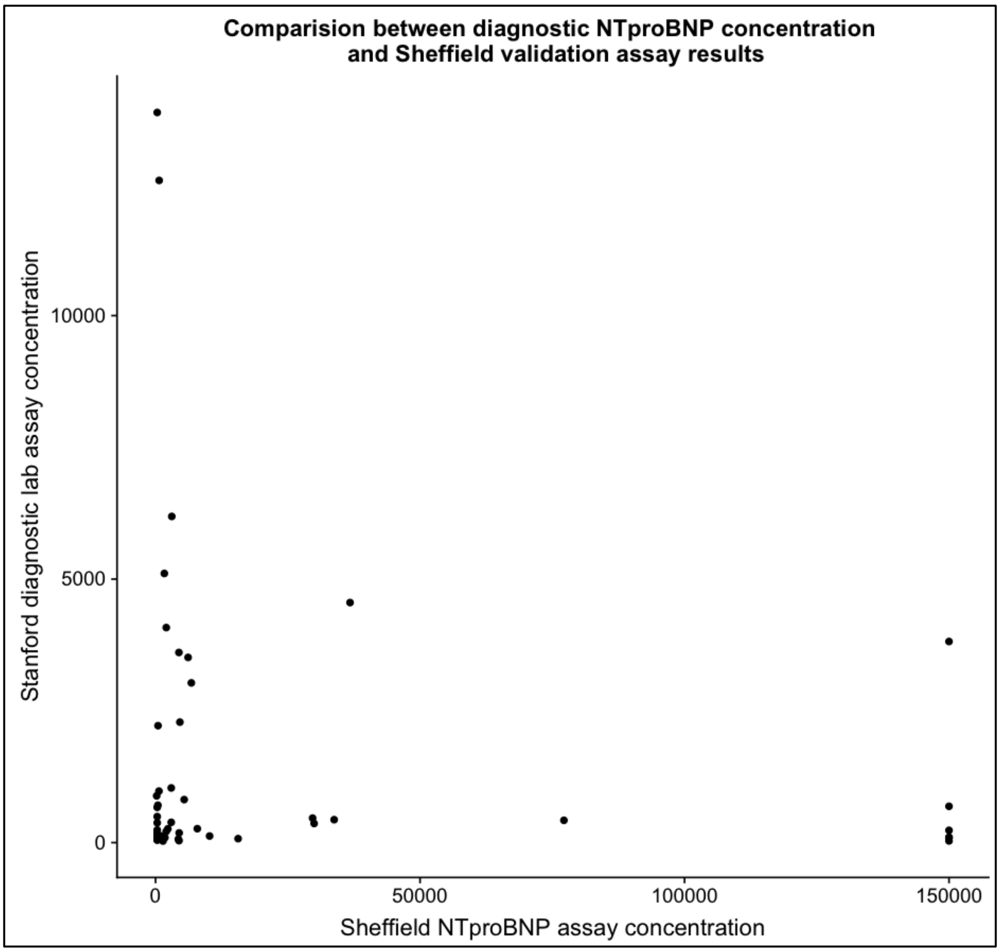


Figure 5.11: Comparison between available clinical NTproBNP assay results (from Stanford University research laboratory) and ELISA validation results.

On this basis, NTproBNP was removed from further consideration in the classifying model.

5.4.5 Assessing agreement between derivation and validation assays

Assessing the agreement between measurement methods is a large analysis in its own right and our study was never designed for this purpose. Nevertheless we have data from two different assays designed to measure the same protein concentrations and therefore some analysis of the agreement by the two methods is important.

Our dataset allows for only limited investigation of the two methods, directly by assessing the differences between measurements of 5 paired samples analysed in by both methods, and indirectly by observing and comparing the distribution of protein concentration measurements between the derivation and validation cohorts.

Agreement is an important concept to differentiate from correlation when analysing two measurement methods. Correlation is concerned with finding a trend in similarity between two variables, whereas to quantify agreement between two measurement methods it is more appropriate to consider the differences in paired measurements between the two methods. The most commonly used technique for doing so is the Bland-Altman plot, in which the difference between paired measurements is plotted against the mean measurement, with indicators for mean difference and 2 standard deviations from that mean shown. 2 measurement methods are considered to be in agreement if >95% of values fall within 2 standard deviations.(Giavarina, 2015)

A Bland-Altman analysis of our data, directly comparing the two measurement methods can be done using only the 5 patient samples which were included in both the derivation and validation assays. For this analysis, the paired measurements from the derivation dataset were analysed against the mean average paired measurement across all plates and replicates in the validation dataset. Given the small numbers, no robust statistical conclusions can be drawn, however the plots allow for inspection of the relationship between measurements (Figure 5.12). Each protein is represented by a scatterplot showing the direct relationship between derivation and validation measurements, and a Bland-Altman plot showing the analysis of differences. It is clear that there is some variability in between the measurement methods, with some outlying measurements significantly affecting correlation statistics given the small sample numbers.

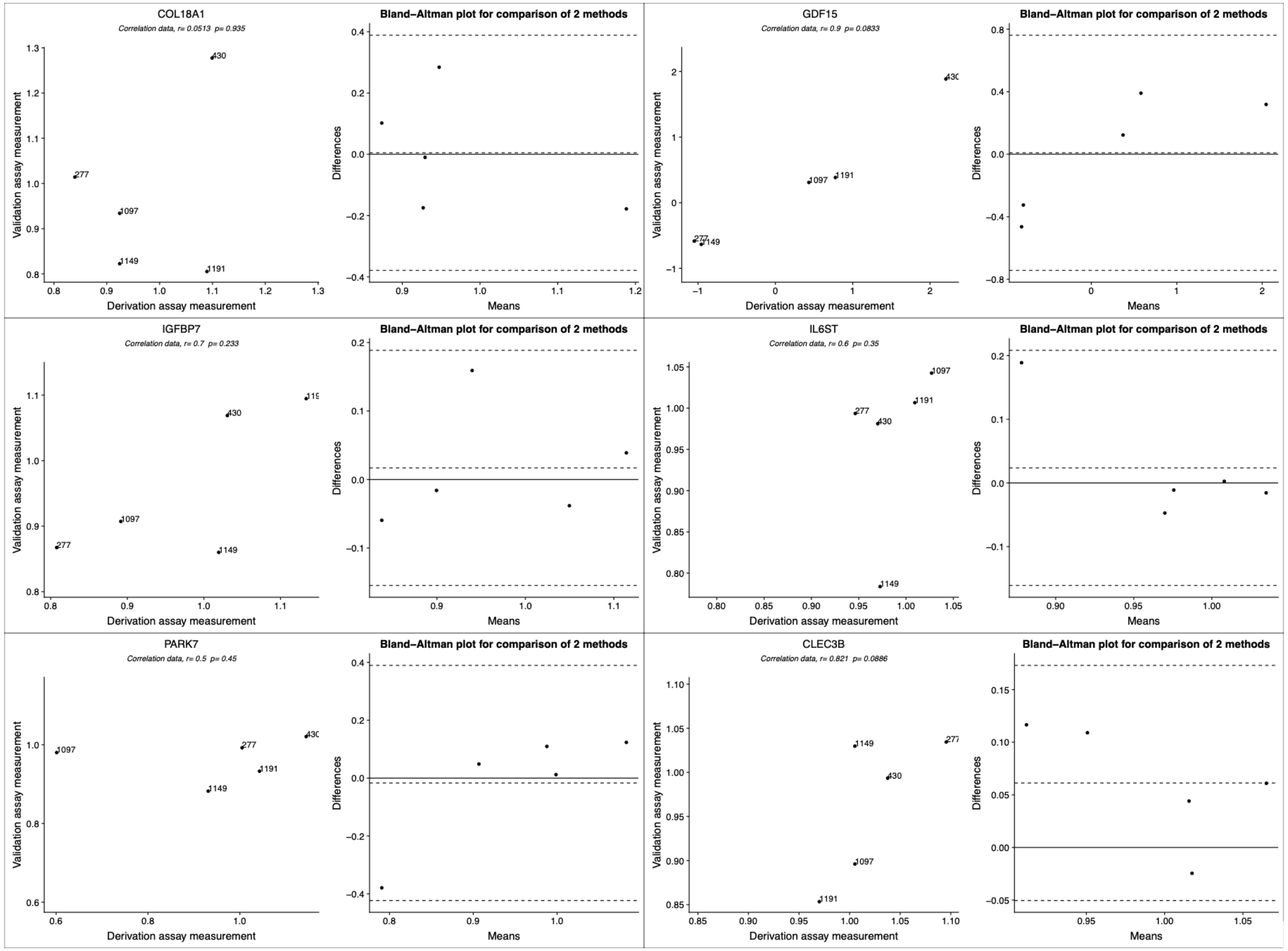


Figure 5.12: Bland-Altman analysis for agreement between derivation and validation assays

For each protein a scatterplot of the logConcentration for each method is shown, along with a Bland-Altman plot for analysis of differences. Within the Bland-Altman plot the middle dotted horizontal line represents the mean difference and the upper and lower dotted horizontal lines represent 2 standard deviations from that mean.

Given the small number of paired samples available for analysis, we also examined whether the protein general unpaired concentration distribution in our validation analyses matched the protein concentration distribution within the Myriad RBM derivation dataset (Figure 5.13). The disease group composition is similar in both the derivation and validation cohort and therefore if the selected proteins are related to the disease group we would expect some similarity in the general protein concentration measurements between the two cohorts.

From the paired sample analysis we can see that there is some suggestion of bias between the two methods for certain proteins, this is also evident from this unpaired analysis of a larger number of samples, particularly for CLEC3B, IGFBP7, IL6ST and PARK7. As in the paired sample analysis, the general measurements between cohorts for GDF15 are very similar.

COL18A1 showed little evidence of correlation between the derivation and validation cohort (Figure 5.12) and was significantly affected by a large proportion of datapoints falling outside the limit of detection in our validation assays (Paragraph 9.3.3). It is therefore likely an unreliable measurement in the validation cohort and does not appear in our final classifying model.

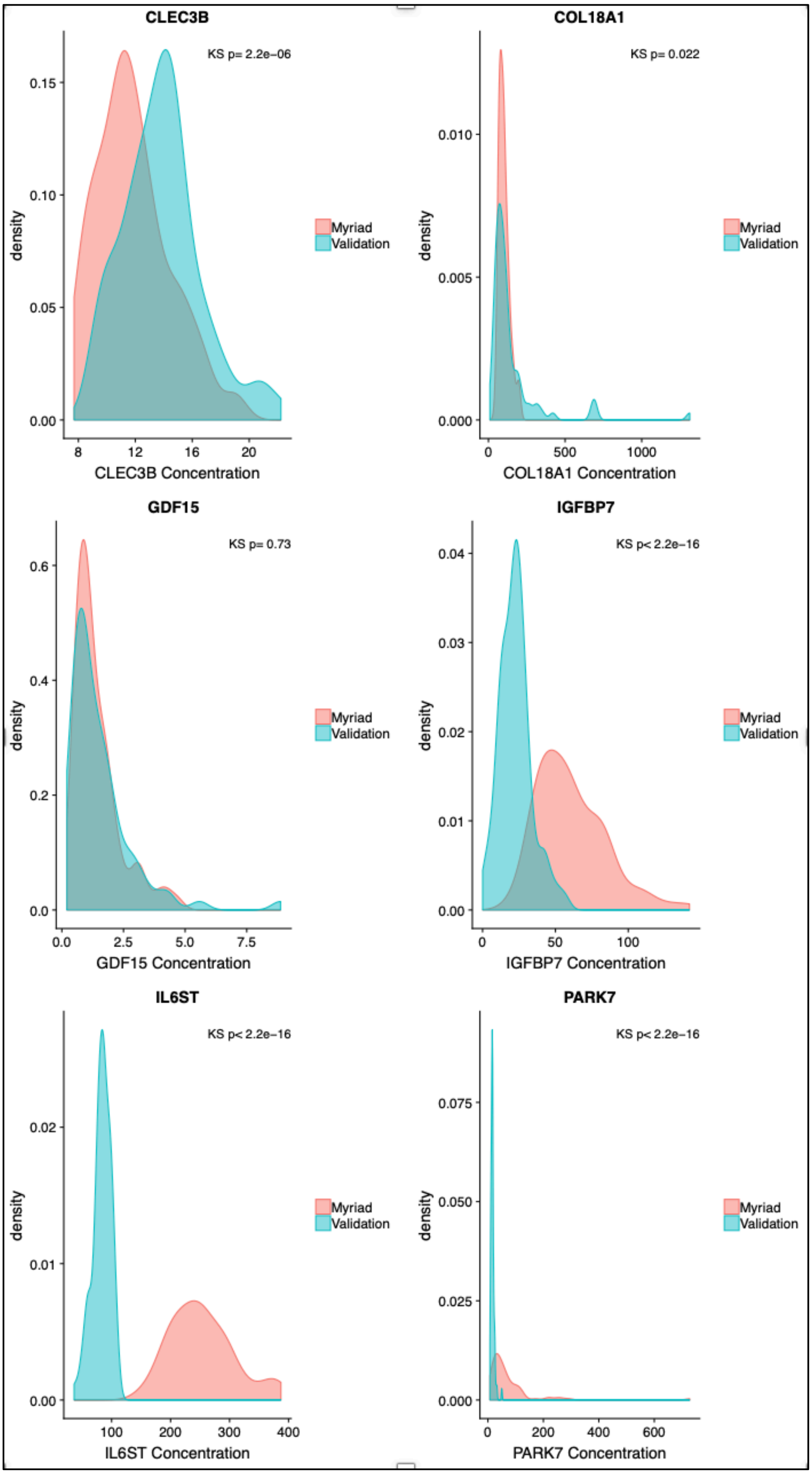


Figure 5.13: Comparison of protein concentration distributions between Validation dataset and Myriad dataset

Comparison of each of the remaining 6 proteins after NTproBNP dropped. Comparison statistic given is Kolmogorov-Smirnov p-value.

Formal statistical analysis of these protein distributions with Kolmogorov-Smirnoff test demonstrates statistically significant differences between cohorts in the measurement distributions for CLEC3B, IGFBP7, IL6ST and PARK7. GDF-15 concentrations are well matched between the analyses.

Simple median correction was applied to the validation dataset resulting in an improved data distribution match between the analyses (Figure 5.14). These adjusted protein data were taken as the final validation dataset for statistical analysis.

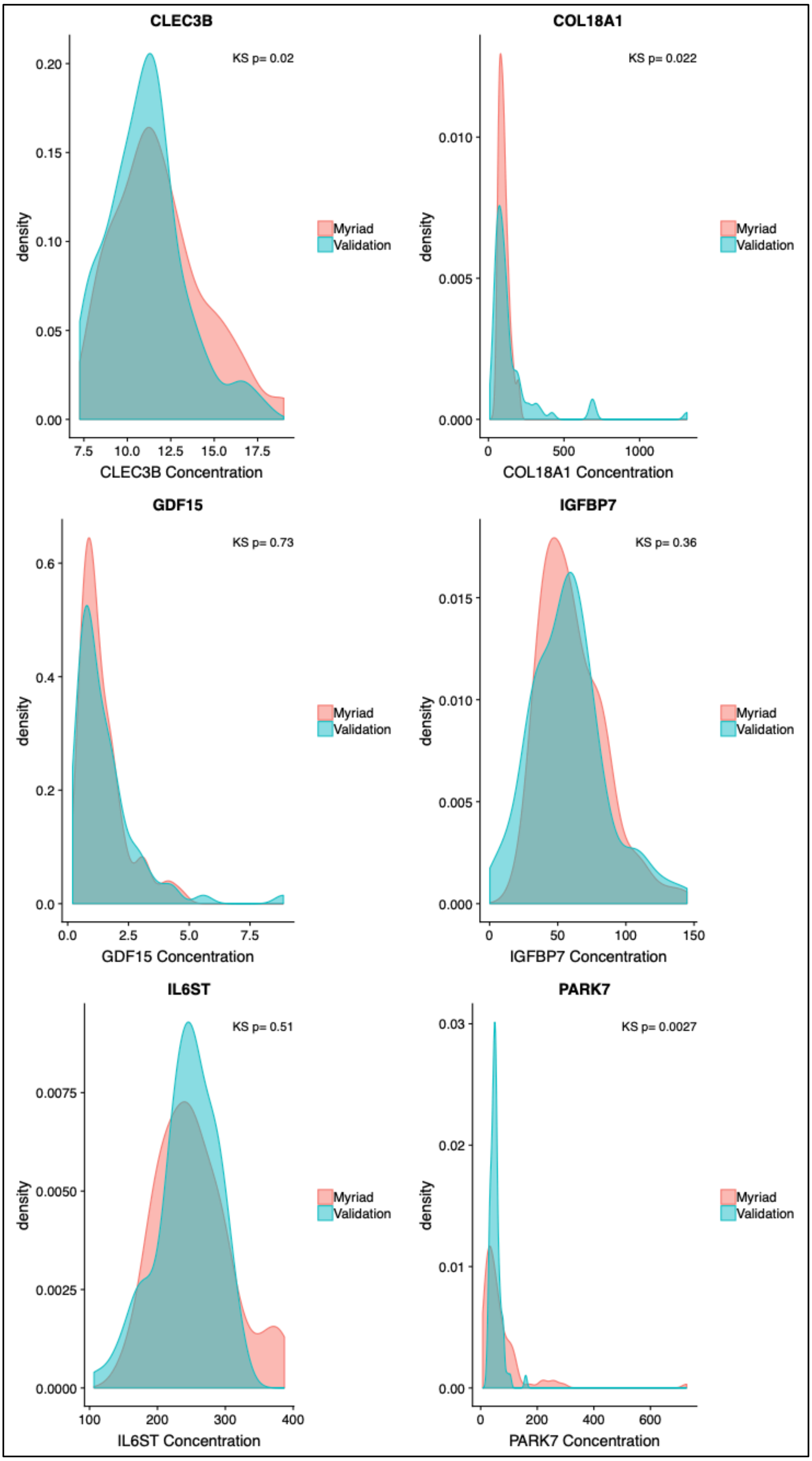


Figure 5.14: Comparison of protein concentration distribution between validation and Myriad datasets after median corrections.

5.4.6 Statistical analysis of model using external validation cohort

Comparison of protein distributions by disease classification in each of the validation and derivation cohorts demonstrates the issue of overfitting in the proteins dropped from the 7 protein panel to derive the final 3 protein panel (Figure 5.15).

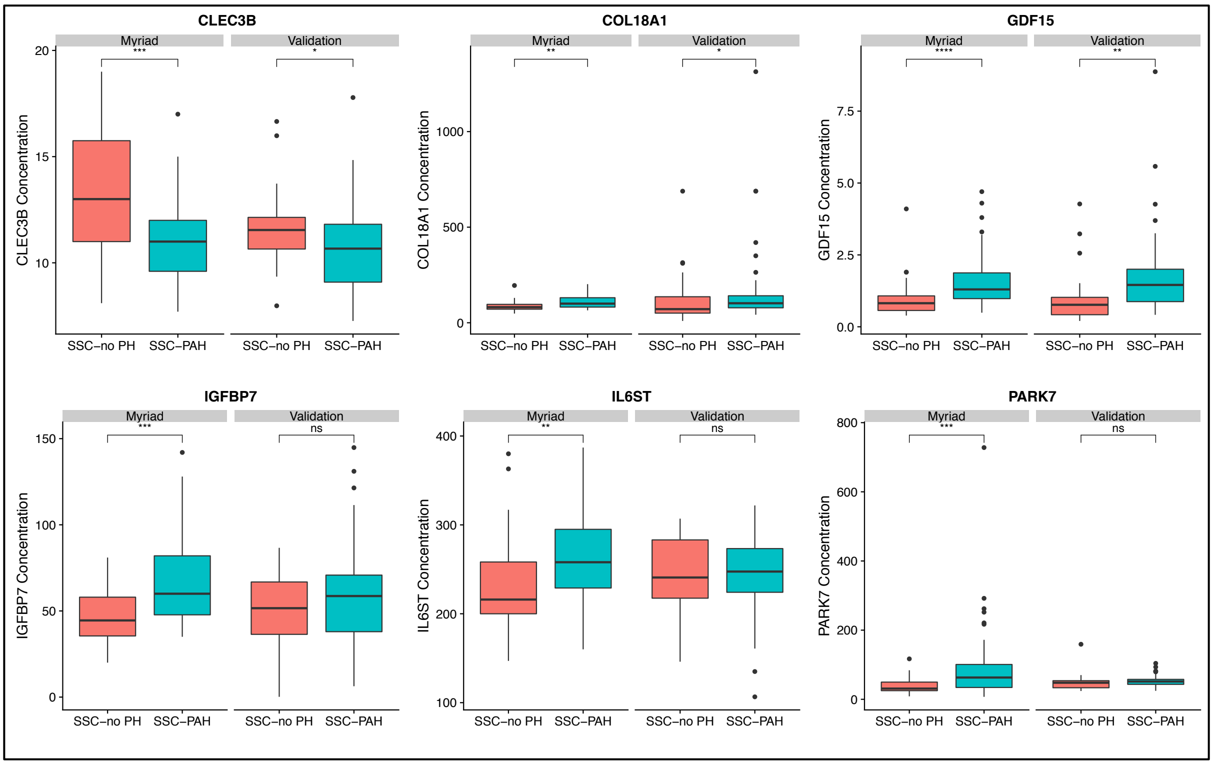


Figure 5.15: Boxplots of protein concentration in derivation and validation cohorts

These data show that although there are significantly altered protein concentrations in the derivation cohort for IGFBP7, IL6ST and PARK7, these significant differences are not demonstrated in the validation cohort. It is therefore likely that retaining all of these variables in the classification model would represent a model overfit to the derivation cohort only. Significant changes are shown in both cohorts for CLEC3B, COL18A1 and GDF15.

For entry of the validation dataset into the final classifying model data were prepared in the same way as for derivation modelling with log transformation and scaling of the dataset. The external validation cohort was modelled blind to the true clinical classification as the samples were received and analysed before the matching clinical phenotype information were available. Data were entered into the final logistic regression model for classification (Table 4.6). The resulting scores are shown in Figure 5.16a.

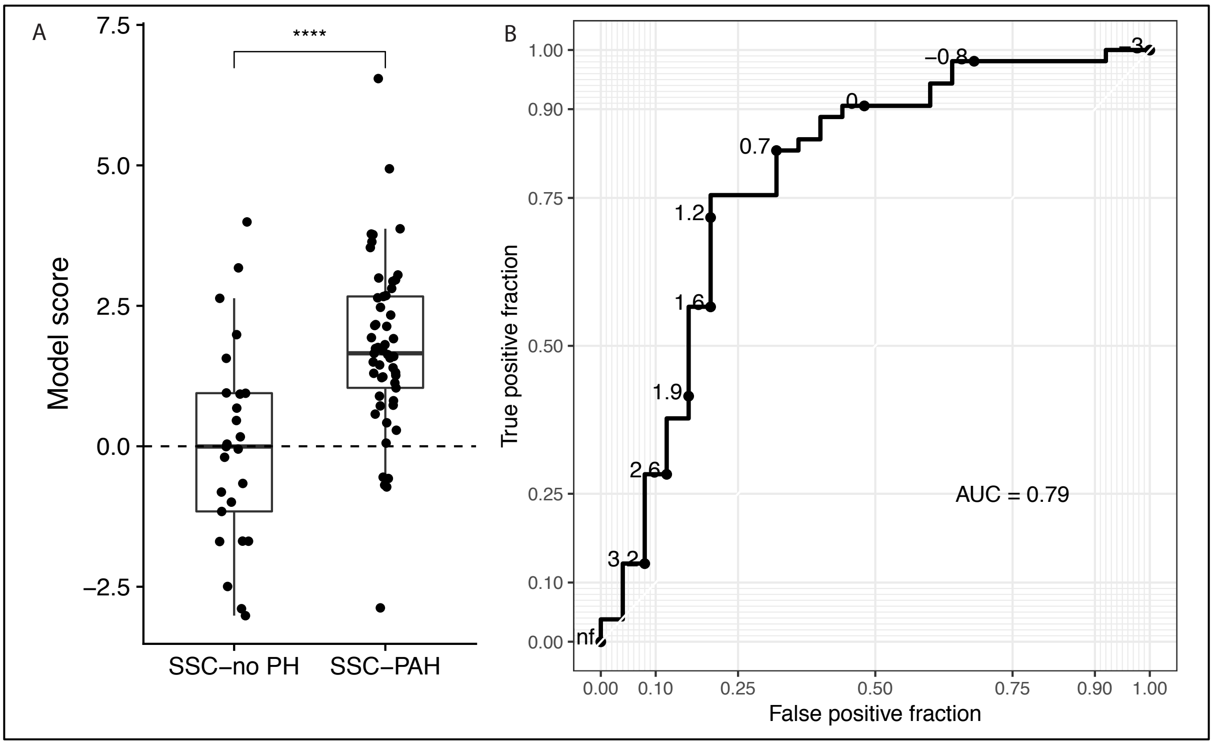


Figure 5.16: A: Boxplot showing model scores for external validation cohort, B: ROC showing classification utility in validation cohort.

Our model classifies PAH in SSc within the external validation cohort with an AU-ROC for of 0.79 (Figure 5.16b) which is reasonable given that the derivation cohort are all sampled at disease baseline when treatment naïve, and the validation cohort are a prevalent group of patients, 32 (60%) of whom are established on PAH targeted treatment (Figure 5.17).

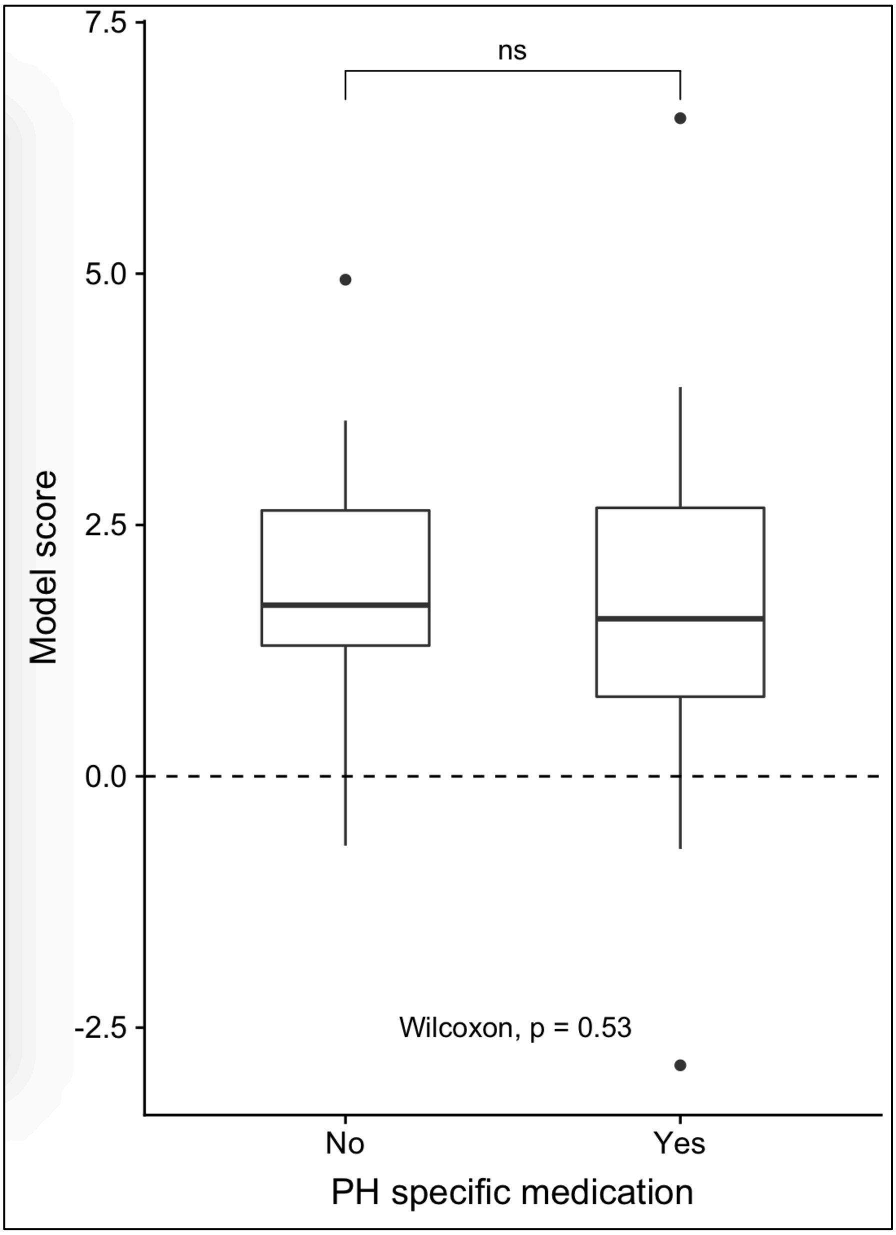


Figure 5.17: Subanalysis of Figure 5.16 showing number of SSc-PAH patients from the validation cohort who are already on PAH targeted treatments

5.5 Discussion

In this chapter I set out to test the fit of the classifying model to assess whether this model would generalise and prove an accurate classifier when applied to validation samples.

The robustness of the classifying model was tested through three different methods, each time proving its utility for the classification of PAH in SSc, and also generating some evidence for this panel in the detection of PAH in patients with idiopathic disease. The first two analyses were based on the dataset received from the assays performed by Myriad RBM, allowing for tight control over any intra-assay variation.

K-fold cross validation is a technique often used when there is no external validation cohort for more rigorous testing. Our K-fold cross validation with a large number of repeats allowed for adequate randomisation in the way the sample set were split into folds for testing, to maximise the accuracy of the final average result. Based on the derivation cohort which demonstrated a diagnostic accuracy of 85% when the classifying model is applied across the entire derivation cohort, k-fold cross validation confirmed this result with an average accuracy of 83% when the model is repeatedly optimised and tested in these derivation and validation subsets.

Although not intended for use in the IPAH subclass of patients, the initial Myriad cohort of patient samples included a cohort with both IPAH and healthy volunteers which were analysed under the same high-quality controlled conditions and therefore returned directly comparable concentrations for the same protein analytes as those for the SSc derivation work. Applying the unaltered classification model to the IPAH/HV cohort provided an alternative test for the validity of these proteins in the pathophysiology of PAH, and demonstrates that these are likely related to the PAH process rather than any other comorbidity associated with the SSc disease process. Validation in this cohort proved beneficial with an AU-ROC for classification identical to that found for the classification of PAH in SSc – 0.87.

The most rigorous test of model validity is testing against an external validation cohort. For this purpose we tested our model against patient samples from Stanford and Vanderbilt University pulmonary vascular disease services. It was not possible to obtain a suitable cohort of patients to match exactly to our baseline treatment naive derivation cohort. The external samples were obtained from patients with prevalent PAH, and therefore represent patients at an established stage of disease, which differ from the potentially more subtle pathophysiological changes of early PAH, most notable amongst this is the rise in NTproBNP caused by increasing right ventricular wall stress in progressing pulmonary vascular disease, which would be absent in the most early stages of a developing pulmonary vasculopathy. Applying our model to this cohort returned an AU-ROC for classification of 0.79 which is promising despite the clinical differences in this validation cohort.

The external validation was affected by several limiting factors which are likely to have had some bearing on the outcome, both technical variability within the assays performed, and variability between the patient cohorts used for derivation and validation. The validation assay technique run in-house differed from the very tightly controlled, automated commercial process used by Myriad RBM for the derivation assays. The antibody pairs are likely to be different, with the validation assays performed using commercially available assay kits, and the derivation assays using antibody pairs developed in house by Myriad RBM. It is therefore likely that the antibody pairs used targeted different epitopes on the target proteins which may have affected the results seen. Further to this, the validation assays were performed manually with some evidence of technical variability affecting the final concentration results.

Patient selection may have differed during the selection and recruitment of patients between the derivation and validation cohorts. In Sheffield we recruit patients at baseline referral immediately after diagnostic investigation to confirm and accurately phenotype the disease class, but before commencement of treatment. This can still be a significant time after onset of symptoms as we know there is often a delay to recognition of possible underlying pulmonary vascular disease and referral for definitive investigation. Patients provided by the external validation centres (Stanford and Vanderbilt) appear to differ in the stage of disease at recruitment as most patients were established on pulmonary vasodilator therapy by the time of sampling. Several patients in the external cohort had elevated NTproBNP concentrations at the time of sample, and combined with the evidence of frequent established treatments for these patients, suggests that these patients are recruited at a later stage of the disease process with well established PH. These patients are different to our ideal target patient cohort who are a group of patients with early stage pre-symptomatic disease who would benefit from early identification and treatment. The serum sampling technique, processing and storage also differed between all three PH centres involved.

When I examined the results of the external validation in detail (Figure 5.16), it became apparent that some of the patients clinically classified in the SSc-no PH cohort could have been falsely classified into this group. Although they do not meet the strict criteria for diagnosis of PH, there were some abnormalities that may alter the outcome of our validation. 5 patients in this group had a PVR ≥ 3 Wu, and 3 had a laboratory measured NTproBNP > 450 pg/ml. No detail is available as to whether these patients progressed on to develop PAH, however these results suggest a component of cardiovascular disease affecting this control group.

The final important consideration regarding potential sources of error inherent in the classification model goes to the fundamental purpose of this research and its target population. Both the derivation and validation cohorts of patients in this study were recruited from PH referral centres, therefore a preselected group of patients who had been referred for investigation of possible PH likely based on suggestive symptoms or basic investigations. This group of patients is therefore enriched for the presence of PH which will affect the quality of the SSc control cohort, with a higher likelihood of physiological (non-PH) abnormality in this group. It is possible that the classifying potential of our model is underestimated when tested in this cohort, and would be improved by testing against a true validation cohort such as an unselected group of SSc patients in the rheumatology clinic.

6 In vitro mechanisms

6.1 Introduction

The discovery analysis was performed using the Myriad RBM DiscoveryMAP platform of analytes which are broad in nature and not specifically targeted at cardiovascular disease. A secondary aim of our project was to identify proteins which have not previously been described in relation to PAH to allow us to identify potential novel proteins and pathways involved in this complex pathophysiology and as potential new targets for treatment.

Our derivation dataset contains 27 proteins for which there is a statistically significant protein concentration difference between SSc-PAH and SSc-no PH (based on false detection rate adjusted p-values). Of these, our final classification model retains of only three proteins: GDF-15; PARK7 and CLEC3B.

GDF-15 is reported by a reasonable body of published evidence which describe a clearly established role for altered expression of this protein in PAH of various subtypes. It is a member of the TGFß superfamily of cytokines, involved in regulating cell growth and differentiation. GDF-15 expression is known to be upregulated in lung tissue of patients with PAH, localising to endothelial cells and areas of active vascular remodelling such as plexiform lesions.(Nickel et al., 2011) Furthermore, circulating concentrations of GDF15 have been shown to be more significantly elevated in SSc-PAH than in IPAH despite less severe haemodynamic changes, with concentrations correlated to multiple indices of PH. Importantly, GDF-15 was found to predict mortality in SSc-PAH with greater accuracy than NTproBNP.(Meadows et al., 2011) Despite an established relationship between GDF-15 and PAH, evidence regarding the presence of GDNF family receptor alpha-like (GFRAL), the receptor for GDF-15, in lung tissues is not currently available.

PARK7 is a protein which is ubiquitously expressed in human tissues. Genetic study of PARK7 has found that loss of function mutations are associated with familial Parkinson's disease, and gain in function mutations with unregulated cell survival in cancer biology.(Vasseur et al., 2009) Multiple intracellular functions have been associated with PARK7 including resistance to oxidative stress in vascular cells (Lee et al., 2006, Taira et al., 2004, Vasseur et al., 2009) and more recently an association has been made between PARK7 expression and modulation of endothelial nitric oxide synthase activity.(Won et al., 2014) PARK7 also exists as a secreted protein, associated with promoting angiogenesis by stimulating angiogenesis through the FGF-1 receptor.(Kim et al., 2012) To our knowledge, PARK7 has not previously been studied in relation to human PAH, however these mechanisms present in other pathologies would support the need for investigation of a similar role in pulmonary vascular disease.

Tetranectin (CLEC3B) represents the least well described protein from our classifying model. It is a known component of proteolytic and fibrinolytic processes, known to bind plasminogen resulting in enhanced activation of plasminogen to plasmin.(Chen et al., 2015) Tetranectin has been described in relation to the enhanced matrix turnover and infiltration in multiple malignant processes.(deVries et al., 1996, Obrist et al., 2004) In cardiovascular disease, Tetranectin has been found decreased in the serum of patients with atherosclerotic disease, with the decrease proportional to the extent of atherosclerosis. Examination of atherosclerotic tissue samples found significantly increased Tetranectin concentrations leading Chen et al. to hypothesise that decreased circulating concentrations are related to an enhanced uptake of the protein into atherogenic lesions.(Chen et al., 2015) Alteration Tetranectin biology has not previously been reported in pulmonary arterial hypertension.

GDF-15, PARK7 and CLEC3B are the constituent proteins of our classifying model and as such these are the proteins that we chose to take forward to study in cell biology.

6.2 Aim

As GDF-15 is already well established in SSc-PAH we sought to identify the presence or absence of its receptor (GFRAL) in human and animal tissues.

PARK7 and CLEC3B are previously less well described and as such we initially sought to identify whether these proteins can be identified in human and animal lung tissues, and subsequently to take these forward into cell culture experiments to look for any typical biological changes which might relate to a pathophysiological mechanism of action known key to vascular remodelling. In order to assess whether animal models of PAH could be used for further studies of these proteins we also sought to identify the presence or absence of these proteins in rat lung tissues.

6.3 Results

For each immunohistochemistry experiment, 3 slides were stained with the example showing the best available section to include both lung parenchyma and pulmonary vasculature.

6.3.1 Immunohistochemistry

6.3.1.1 Negative control

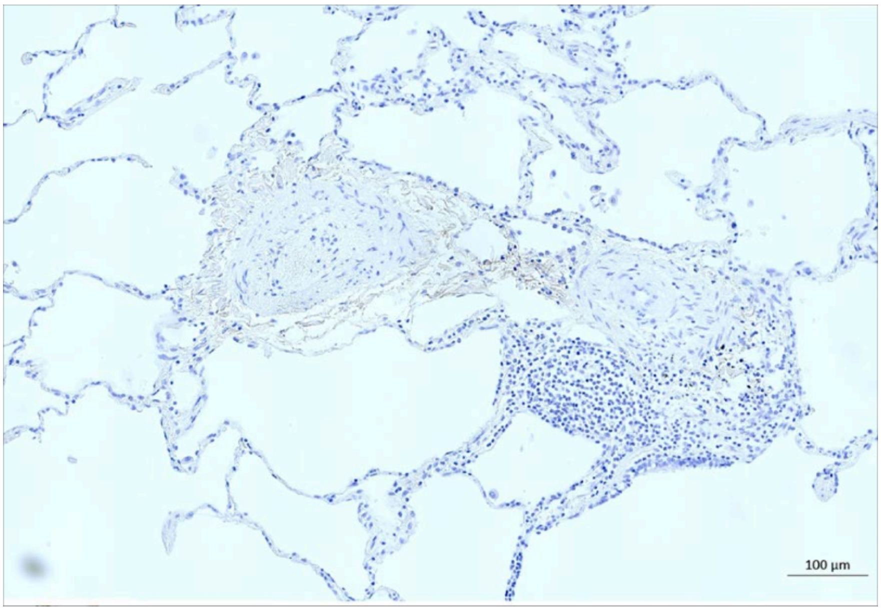


Figure 6.1: Immunohistochemistry: Negative control

Lung tissue sections, processed as per IHC protocol, but without primary antibody. Human PAH lung section at 65x magnification

This negative control section demonstrates remodelled small vessels from a human patient with IPAH. There is neither specific nor background DAB staining evident on this section, produced using the same protocol as for all following IHC sections, but in the absence of a primary antibody.

6.3.1.2 PARK7

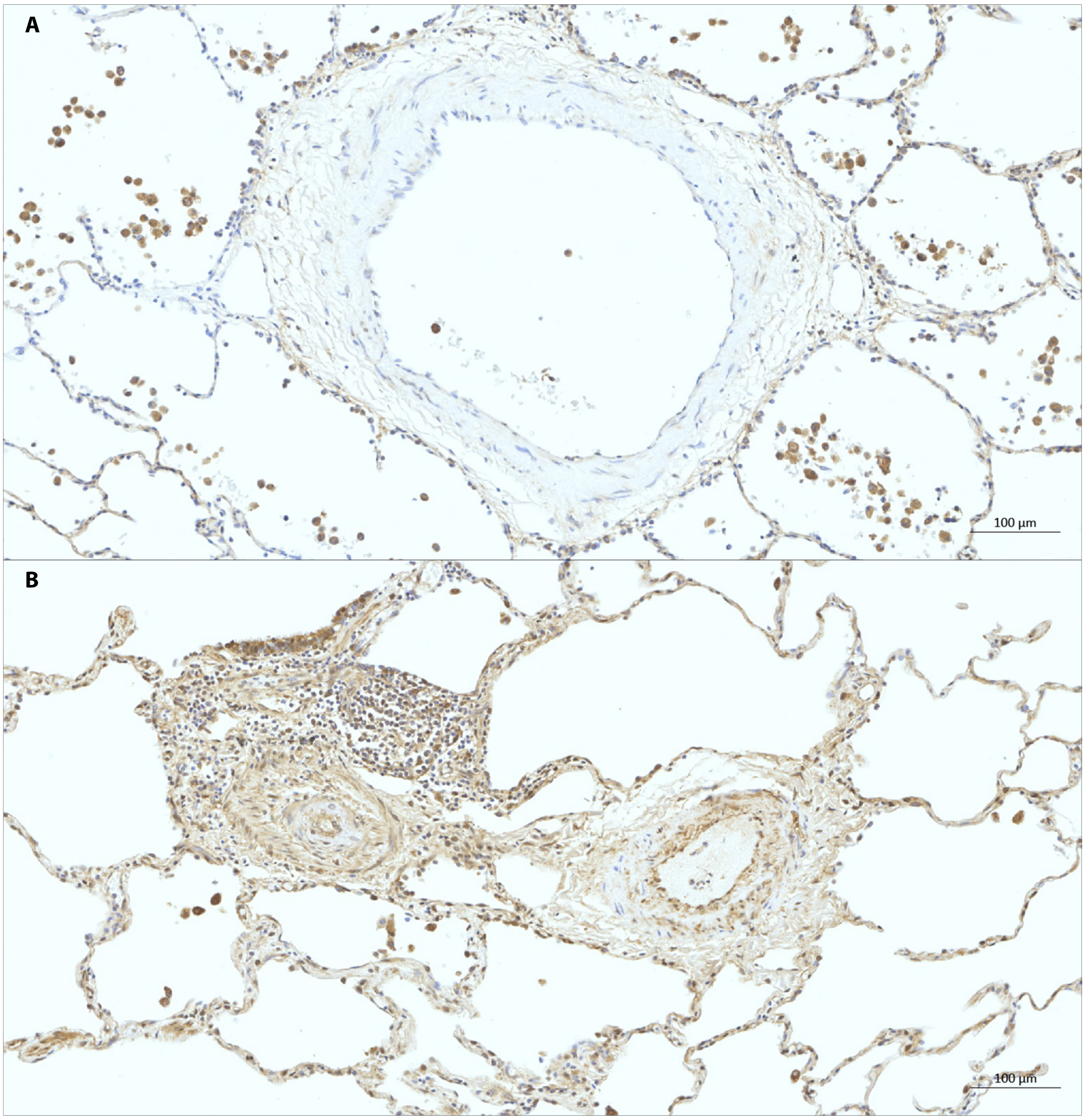


Figure 6.2: Immunohistochemistry: PARK7

Human lung tissue sections stained for presence of PARK7, imaged at x65 magnification. A: Healthy lung tissue; B: PAH lung tissue

Staining for PARK7 is evident in the alveolar cells, and inflammatory cells present in the healthy human lung, with an absence of staining in the vascular media or intimal layers (Figure 6.2a). In contrast, there appears significantly greater staining for PARK7 in remodelled vessels of human IPAH with evident staining in the remodelled vessel wall and endothelium (Figure 6.2b).

6.3.1.3 CLEC3B

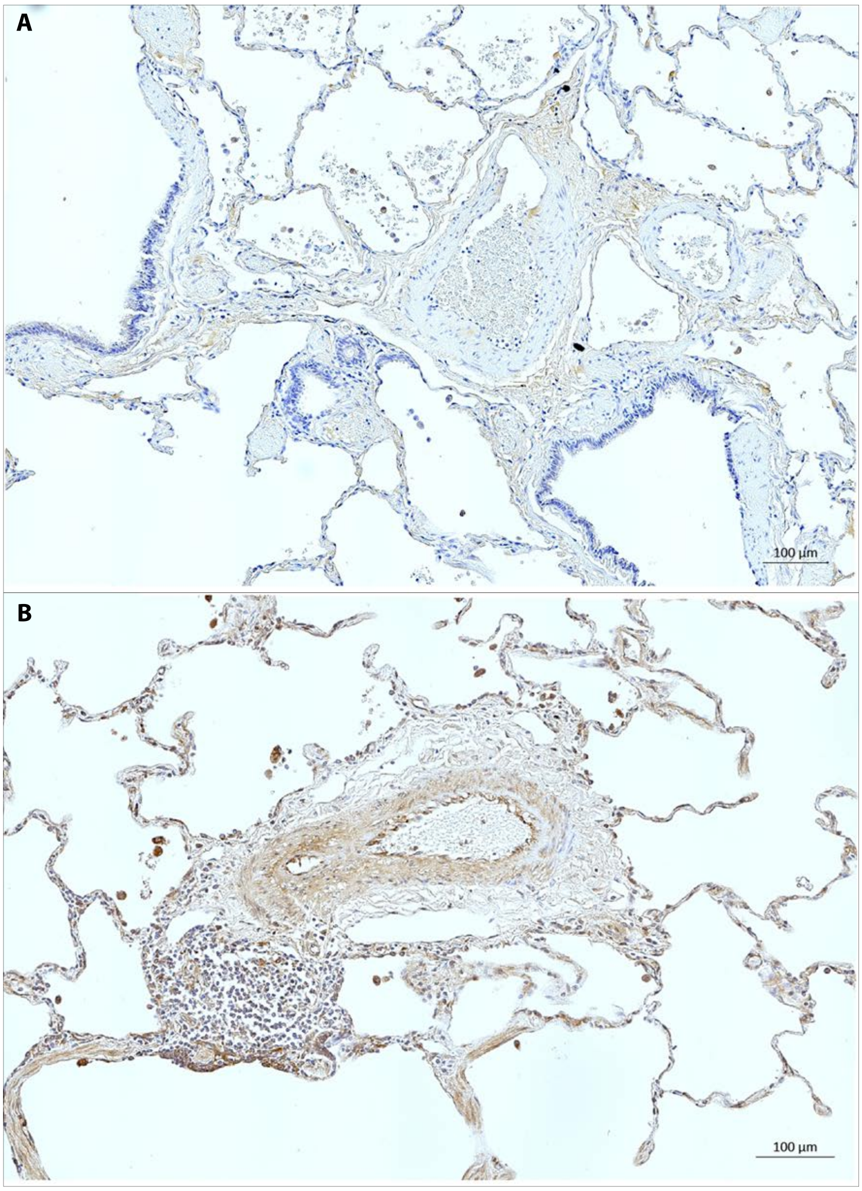


Figure 6.3: Immunohistochemistry: CLEC3B

Human lung tissue sections stained for presence of CLEC3B, imaged at x65 magnification. A: Healthy lung tissue; B: PAH lung tissue

In healthy tissues (Figure 6.3a) only very limited staining is seen in the vascular adventitial layers, however this contrasts markedly with PAH lung tissues (Figure 6.3b) which show staining for CLEC3B in the remodelled vascular media and endothelial layer, as well as some staining in the alveolar tissues.

6.3.1.4 GFRAL

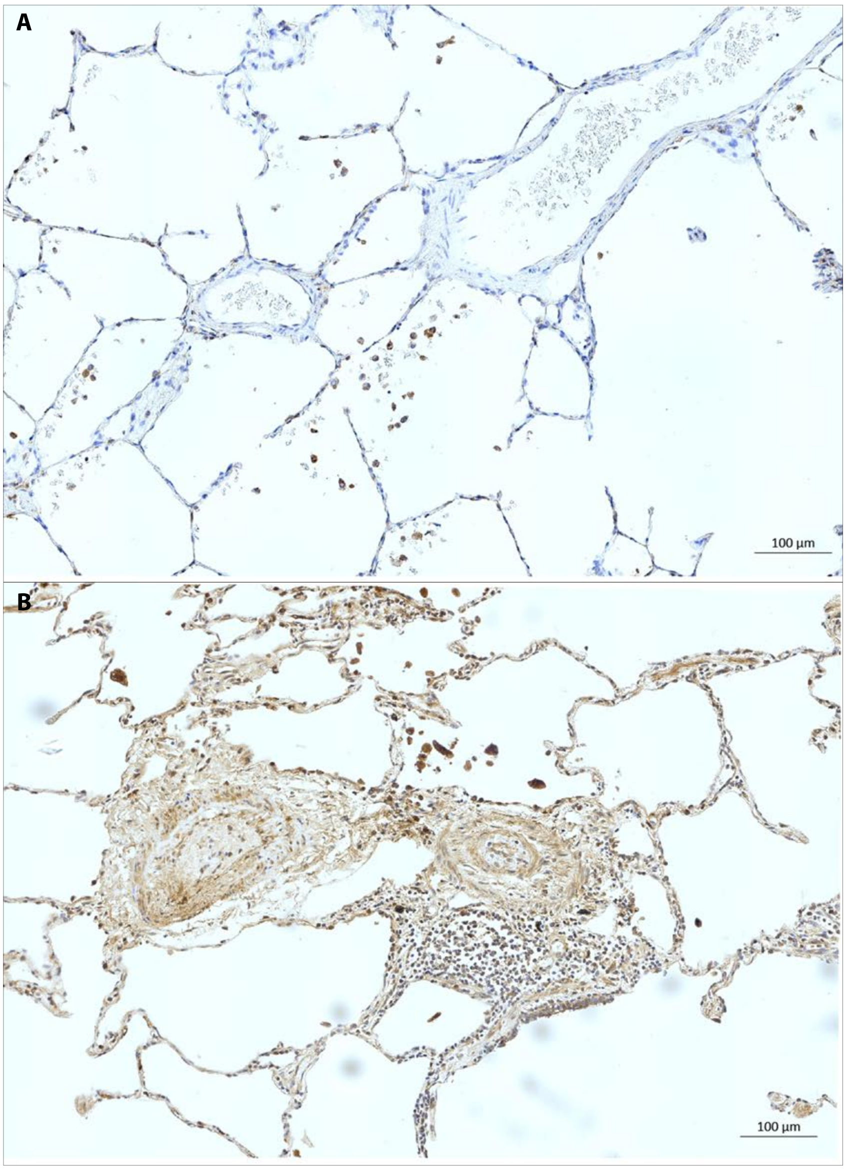


Figure 6.4: Immunohistochemistry: GFRAL

Human lung tissue sections stained for presence of GFRAL, imaged at x65 magnification. A: Healthy lung tissue; B: PAH lung tissue

Again, there is a clear difference for GFRAL staining between human healthy (Figure 6.4a) and PAH diseased (Figure 6.4b) lung tissues. There is no evidence of staining in healthy tissues, but again evidence of staining in the remodelled vessel wall.

6.3.1.5 Animal Models

In order to determine whether the target proteins are expressed in animal models, I also stained 3 sections of lung tissue from Sugen hypoxic rat models of PAH (Figure 6.5). This was to determine whether rat models would be suitable for further mechanistic/ knockdown experiments to further examine the role of our target proteins in the pathogenesis of PAH.

These sections confirm the presence of our target proteins and in a similar distribution seen in diseased human tissues.

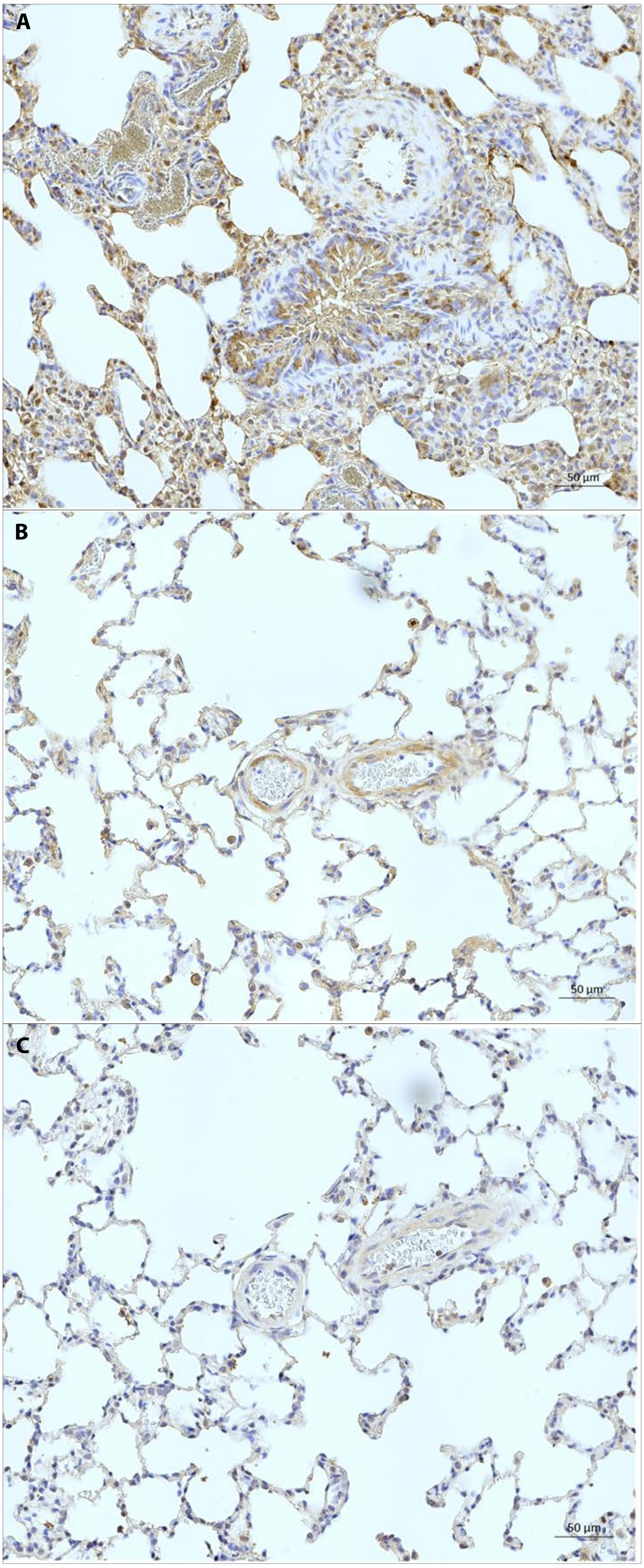


Figure 6.5: Immunohistochemistry in animal models.

Lung tissues from Sugen hypoxic rat models of PAH. A: stained for PARK7; B: stained for CLEC3B; C: stained for GFRAL.

6.3.2 Cell culture

6.3.2.1 Phenotyping of HPAEC

The majority of cell culture work was done using commercial HPAEC bought in and used from passage 3. A few experiments were done using human pulmonary artery endothelial cells acquired from colleagues at Imperial University and stored in our tissue bank. To verify the cell type under investigation we used IHC to phenotype the cells as described in section 2.6.2.1.

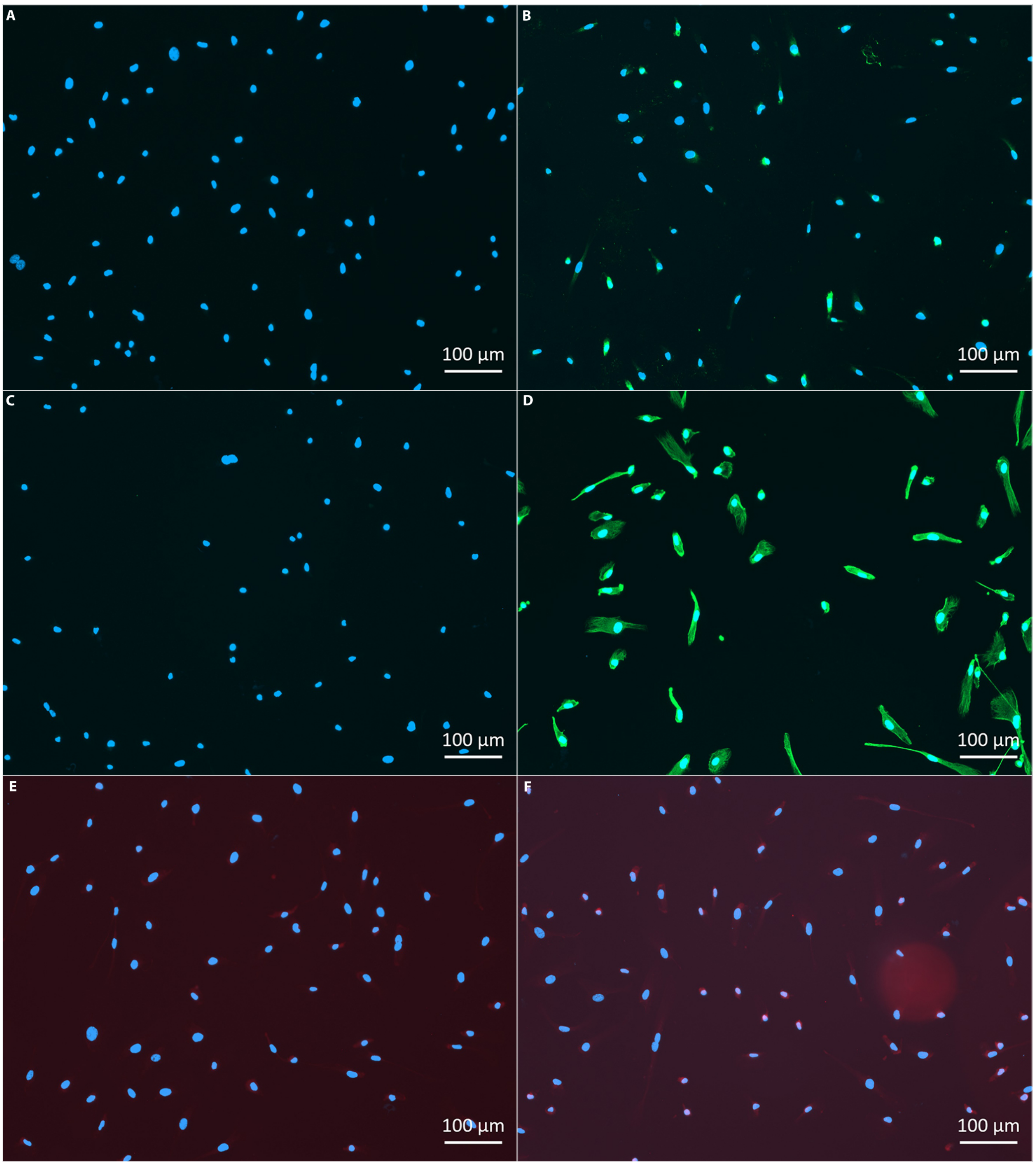


Figure 6.6: Immunofluorescent IHC for cell phenotype

Images at x20 magnification. A: vWF stain negative control; B: vWF stain; C: Vimentin stain negative control; D: Vimentin stain; E: SMA stain negative control; F: SMA stain. A-D Alexa fluor 488 (green), E-F Alexa fluor 555 (red). All samples counterstained with DAPI.

Fluorescent IHC (Figure 6.6) shows no evidence of background staining for negative controls (slides A, C and E), positive vWF (slide B) and Vimentin (slide D) staining, but negative SMA staining (slide F) which is consistent with an endothelial cell phenotype.

6.3.2.2 PARK7

To study the effect of PARK7 on cell types primarily involved in the pathophysiology of PAH we first conducted experiments by attempting to stimulate cells with the protein to investigate for any evidence of altered cell proliferation in both HPAEC and HPASMC, and for evidence of altered cell migration or angiogenesis in HPAEC only.

Three proliferation assays (n=3) were performed on HPAEC, using two different batches of commercially bought cells (Lonza CC-2350). The first two assays used the same batch of cells at passages 3 and 6, and the third assay using a second batch of cells at passage 5. These assays investigate the effect of direct simulation of HPAECs with recombinant human PARK7 protein (Abcam, Cat: ab124312).

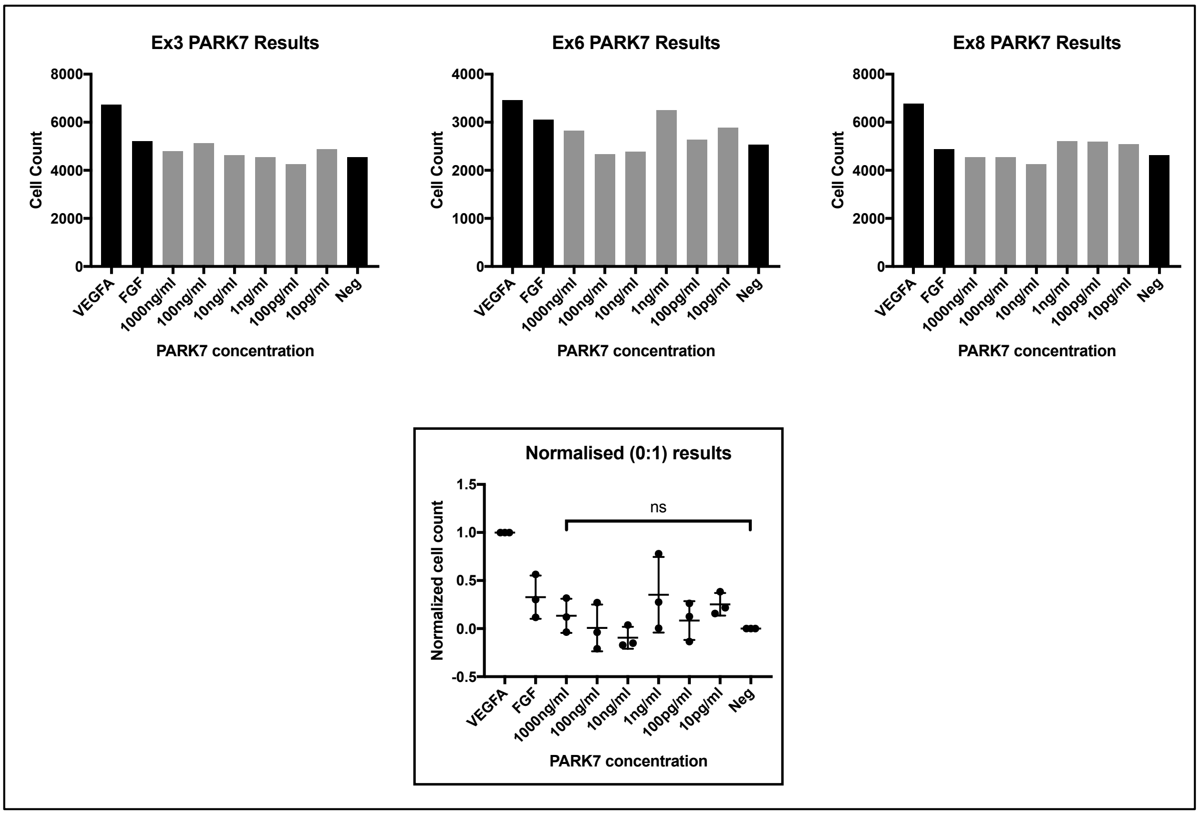


Figure 6.7: Proliferation assays for PARK7 effect on HPAEC.

Top panels showing average results from replicates in individual proliferation assays, with the lower panel showing combined results, normalized between VEGFA positive control and Negative control. ANOVA statistic shown.

Response from these HPAEC to VEGFA positive control stimulation was present but limited. From these data, there is no evidence of a direct effect of PARK7 stimulation on proliferation of HPAEC (Figure 6.7).

PARK7 stimulation, using the same recombinant human PARK7 protein, was also tested to evaluate any proliferation effect on commercial HPASMC (Lonza CC-2581). 2 assays (n=2) were performed on a single batch of cells at passages 4 and 5.

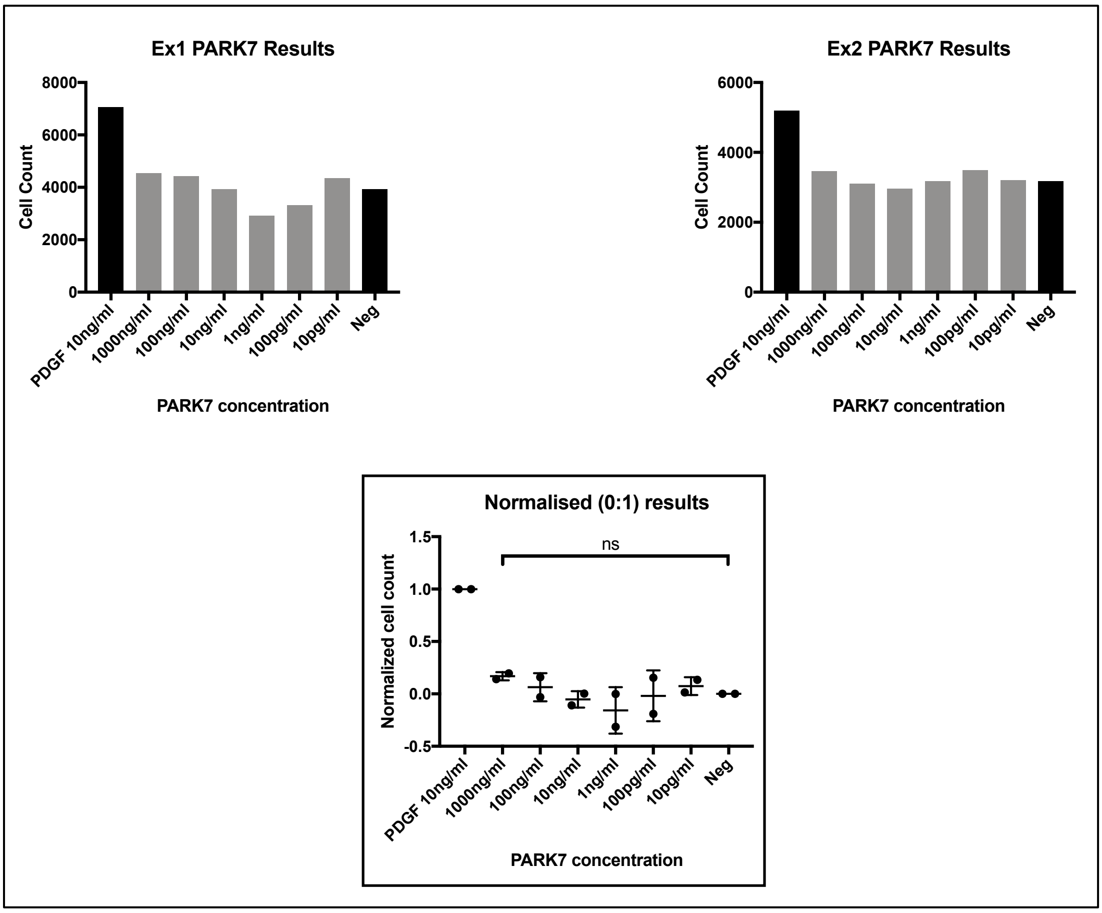


Figure 6.8: Proliferation assay for PARK7 effect on HPASMC.

Top panels show average results from replicates for individual proliferation assays, the lower panels showing the combined results normalized between PDGF positive control and negative control. ANOVA statistic shown.

This experiment shows an appropriate proliferation response from HPASMCs to the PDGF positive control, however shows no evidence of proliferation in response to direct stimulation with PARK7 (Figure 6.8).

3 angiogenesis assays (n=3) were performed using Imperial University HPAECs at passages 3, 4 and 5. Although tube networks developed appropriately (representative images shown in Figure 6.9), the there was no difference demonstrated for total tube length between positive and negative control (Figure 6.10), therefore a failed experiment, and from this data we can draw no meaningful conclusions regarding any effect of direct PARK7 stimulation on angiogenesis in HPAEC.

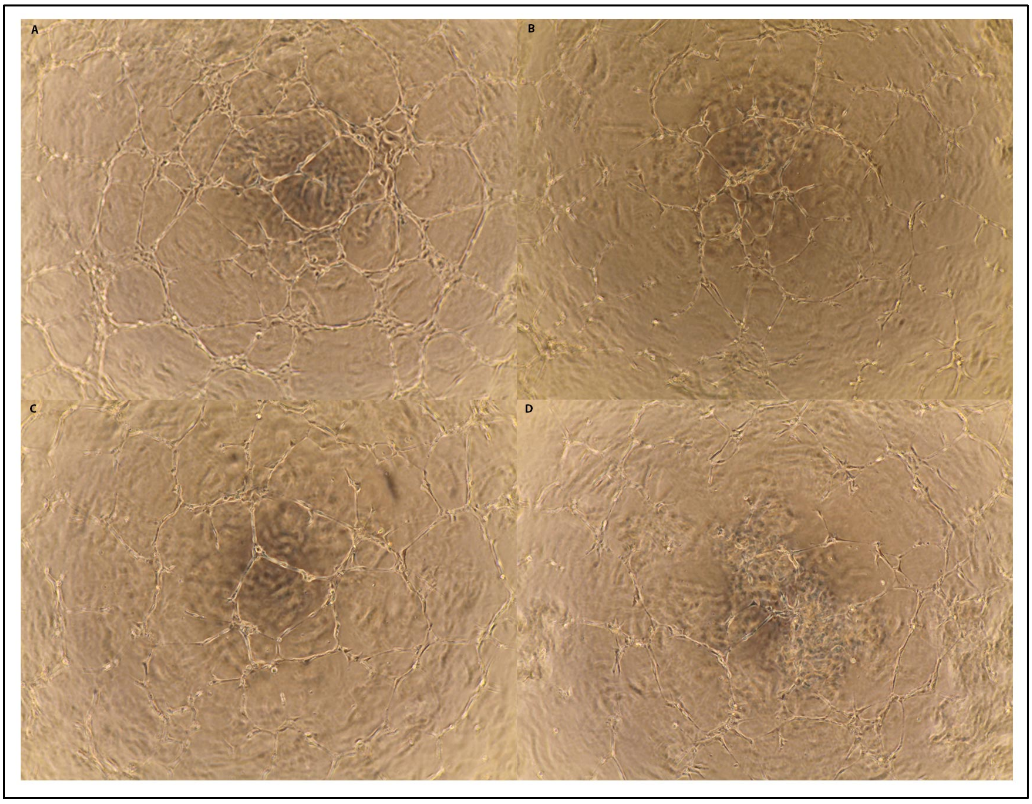


Figure 6.9: Images from angiogenesis assay Ex16, HPAEC stimulated with PARK7.

A: Positive control VEGFA; B: PARK7 1000ng/ml; C: PARK7 1ng/ml; D: Negative control.

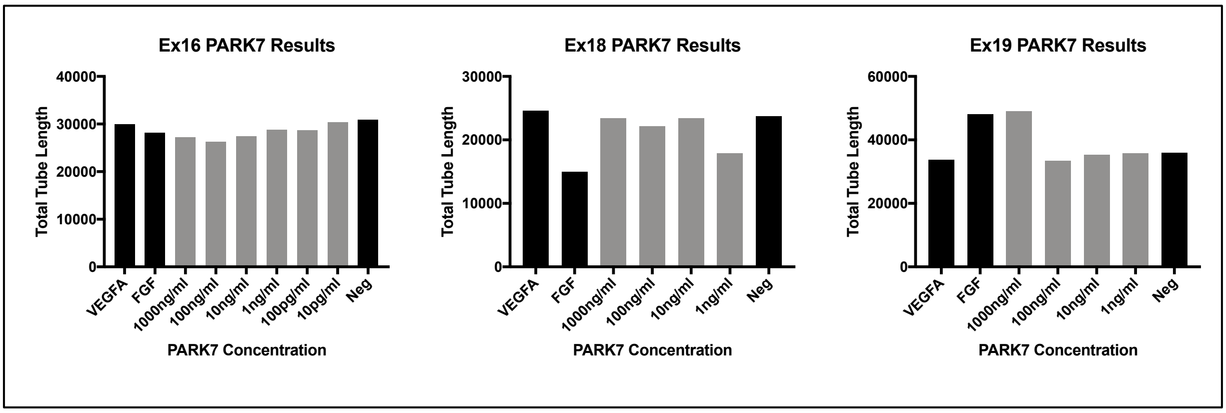


Figure 6.10: Angiogenesis assay for HPAEC stimulated with PARK7.

Showing average results for replicates from individual angiogenesis experiments. Metric shown is total tube length.

3 migration assays (n=3) were performed to investigate the effect of direct stimulation of HPAEC with PARK7 for altering endothelial cell migration. 2 different batches of commercial cells were used in this experiment, 1 batch of cells at passage 6, and another batch of cells at passage 4 and 5.

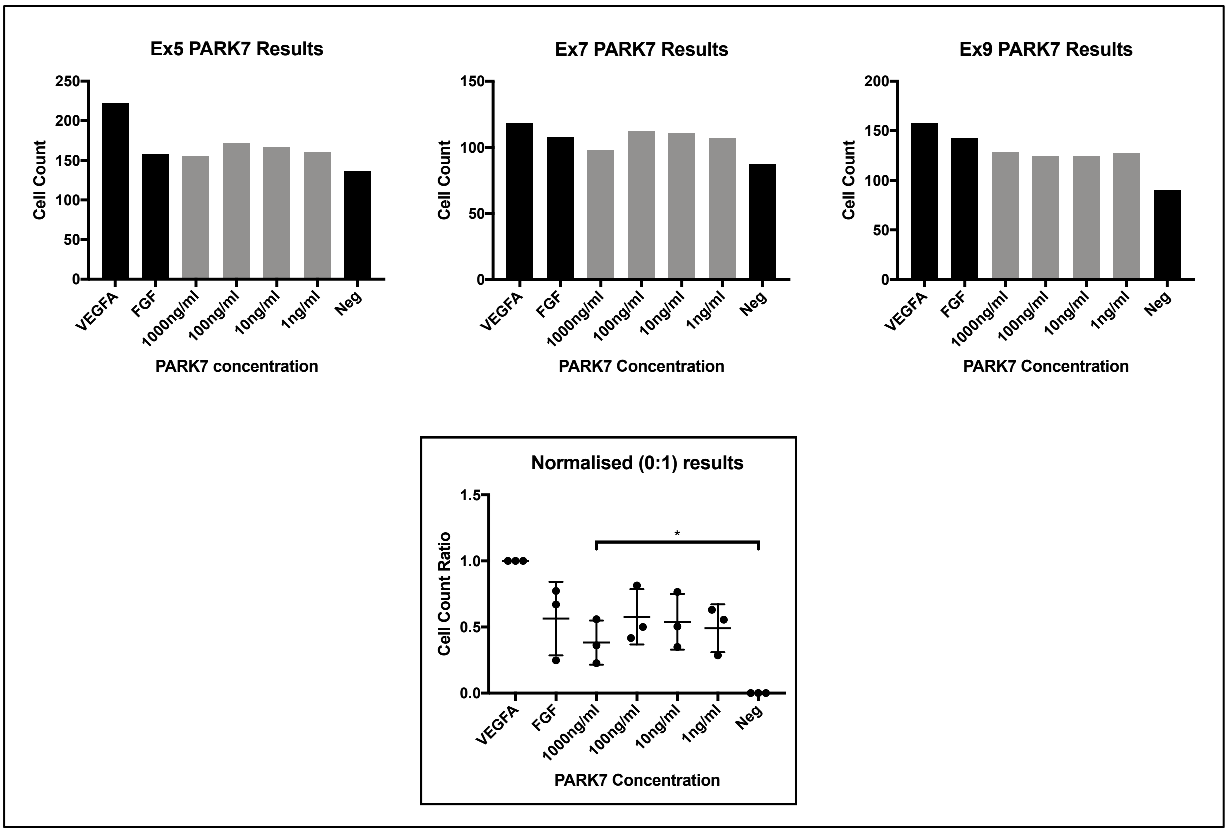


Figure 6.11: Migration assay for HPAEC stimulated with PARK7.

Top panels show average results of replicates for individual experiments, with lower panel showing the combined results normalized between VEGFA positive control and negative control.

Migration response to positive control stimulation with VEGFA was very limited, but despite this there was a small significant difference between all concentrations of PARK7 stimulation against the negative control suggesting some effect of direct stimulation by PARK7 for increasing HPAEC migration (Figure 6.11).

6.3.2.3 CLEC3B

The effect of CLEC3B on pulmonary vascular cells was investigated in parallel to the experiments described for the investigation of the effect of PARK7 in section 6.3.2.2. The number of experiments performed, source of cells used and passage numbers are therefore the same as in the corresponding assays previously described.

The effect of directly stimulating pulmonary vascular cells with recombinant human CLEC3B (R&D systems, Cat: 5170-CL-050) was investigated in proliferation assays (n=3) (Figure 6.12).

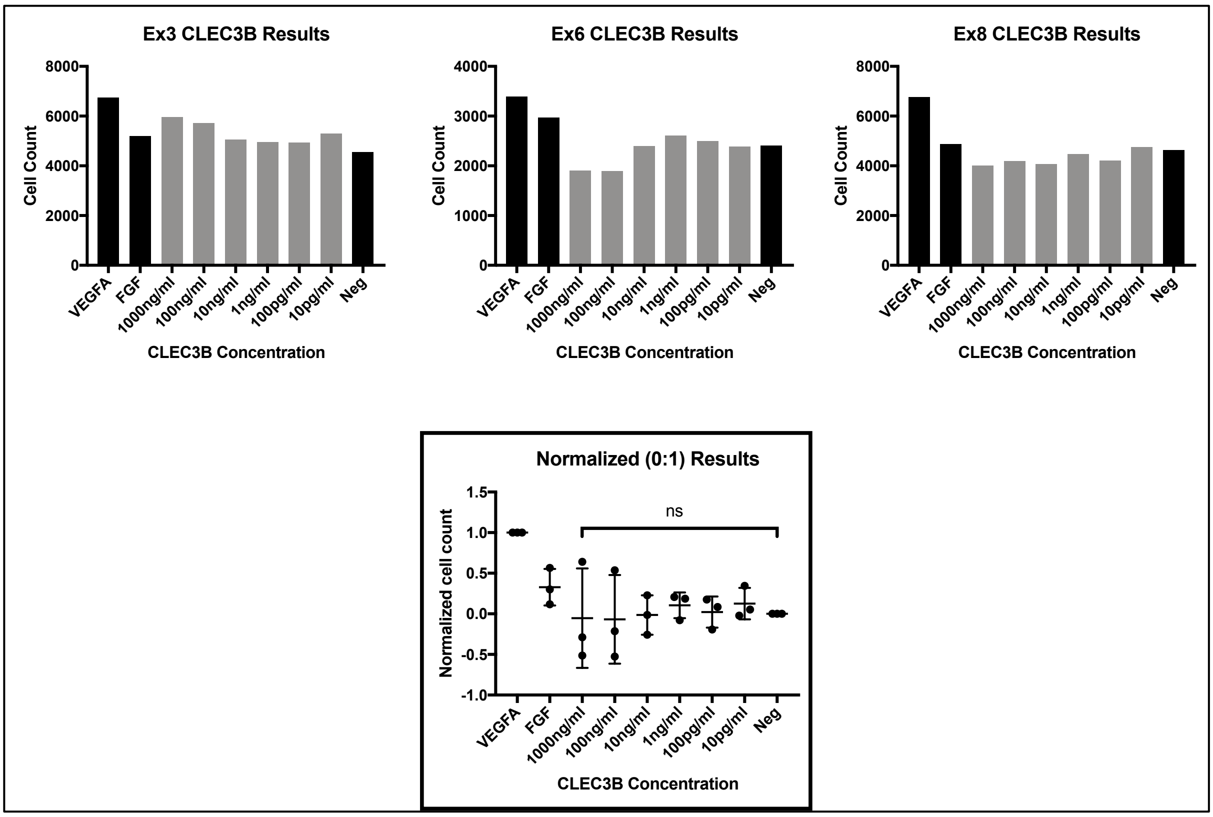


Figure 6.12: Proliferation assay for effect of CLEC3B on HPAEC

Top panels showing average results from replicates in individual proliferation assays, with the lower panel showing combined results, normalized between VEGFA positive control and Negative control. ANOVA statistic shown.

The results are variable between each assay, but neither individual assays, nor grouped analysis show any evidence that proliferation of HPAEC is altered by direct simulation with recombinant human CLEC3B.

Similarly, proliferation assays in HPASMCs did not show any evidence of altered cell proliferation when similarly stimulated with recombinant human CLEC protein (Figure 6.13).

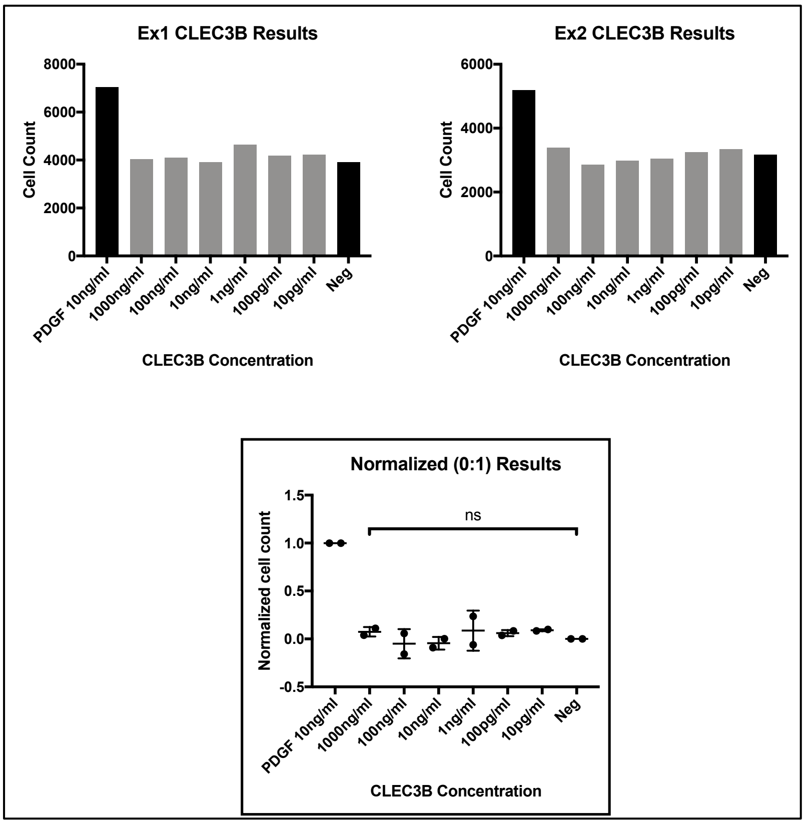


Figure 6.13: Proliferation assay for effect of CLEC3B on HPASMC

Top panels show average results from replicates for individual proliferation assays, the lower panels showing the combined results normalized between PDGF positive control and negative control. ANOVA statistic shown.

Angiogenesis assays were performed to investigate whether recombinant human CLEC3B protein could stimulate angiogenesis in HPAEC. 3 experiments were performed in HPAECs on a growth factor reduced Matrigel matrix, however there was little difference in total tube length measured between the VEGFA positive control, and the negative control and therefore the assays were deemed to have failed (Figure 6.14).

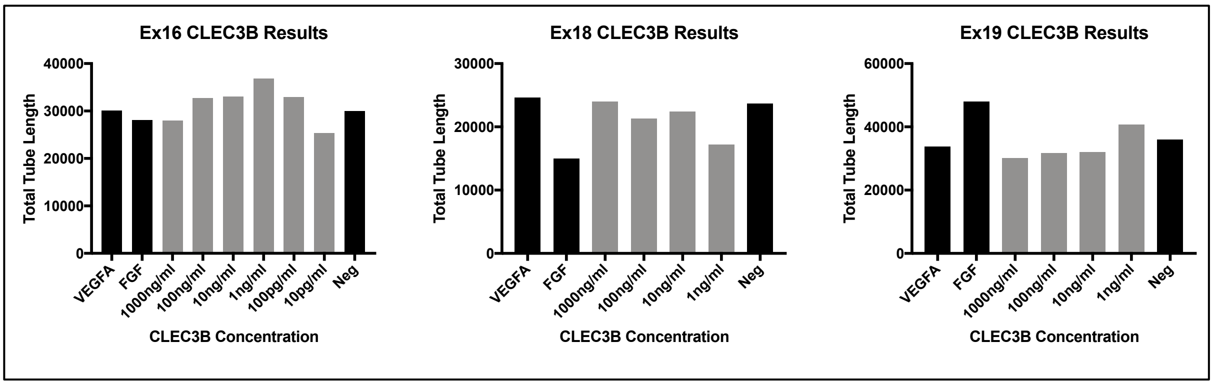


Figure 6.14: Angiogenesis assay for HPAEC stimulated with CLEC3B.

Showing average results for replicates from individual angiogenesis experiments. Metric shown is total tube length.

HPAEC migration was assessed in response to stimulation with CLEC3B protein (n=3 assays) but no difference was demonstrated between CLEC3B stimulated cells and their corresponding negative control (Figure 6.15).

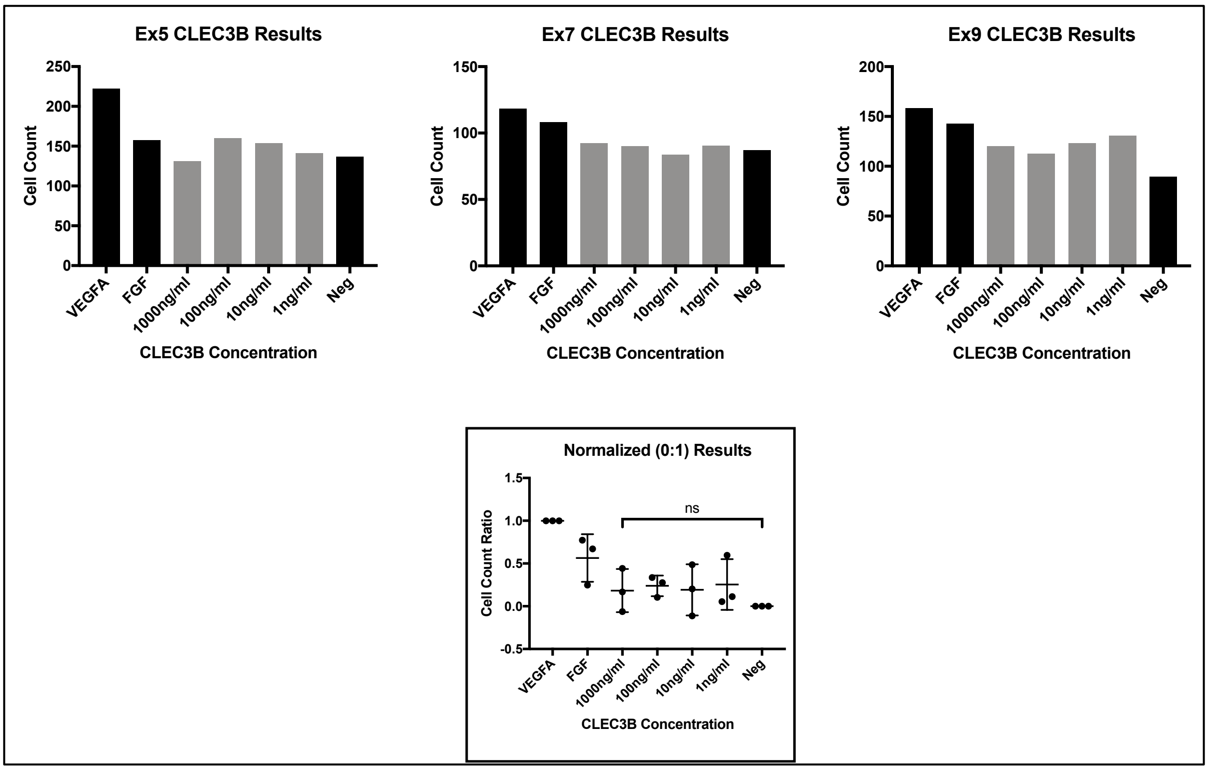


Figure 6.15:Migration assay for HPAEC stimulated with CLEC3B.

Top panels show average results of replicates for individual experiments, with lower panel showing the combined results normalized between VEGFA positive control and negative control.

6.3.2.4 PARK7 and CLEC3B knockdown

Some small significant effects on migration were noted on direct stimulation of HPAEC with PARK7, but other experiments with direct stimulation of pulmonary vascular cells with PARK7 and CLEC3B showed negative results. I therefore wanted to investigate the effect of knocking down these proteins by transfecting cells with appropriate siRNA before repeating the migration assays to look for any change in phenotype.

HPAEC were transfected with siRNA targeting PARK7 and CLEC3B, as well as non-targeting siRNA as control according to the method given in section 2.6.2.5:Transfection. This experiment was repeated three times using HPAEC at passage 6 and 7 from one batch of cells, and passage 5 from a second batch of cells. Non-targeting siRNA was only available for transfection in the final experiment. Migration assays were then performed as previously described. Protein knockdown was quantified using Western blot (as per 2.6.2.6:Protein quantification and Western blot) on protein isolated from cells used in the final experiment.

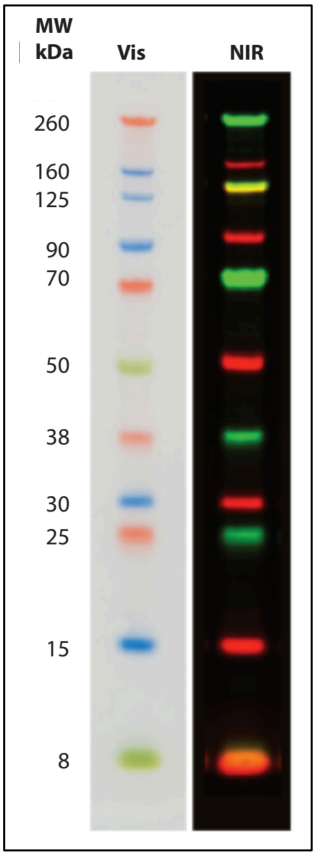


Figure 6.16: Chameleon duo pre-stained ladder

Two colour near-infrared image shown on the right, with visible image on the left.(Li-Cor, 2015)

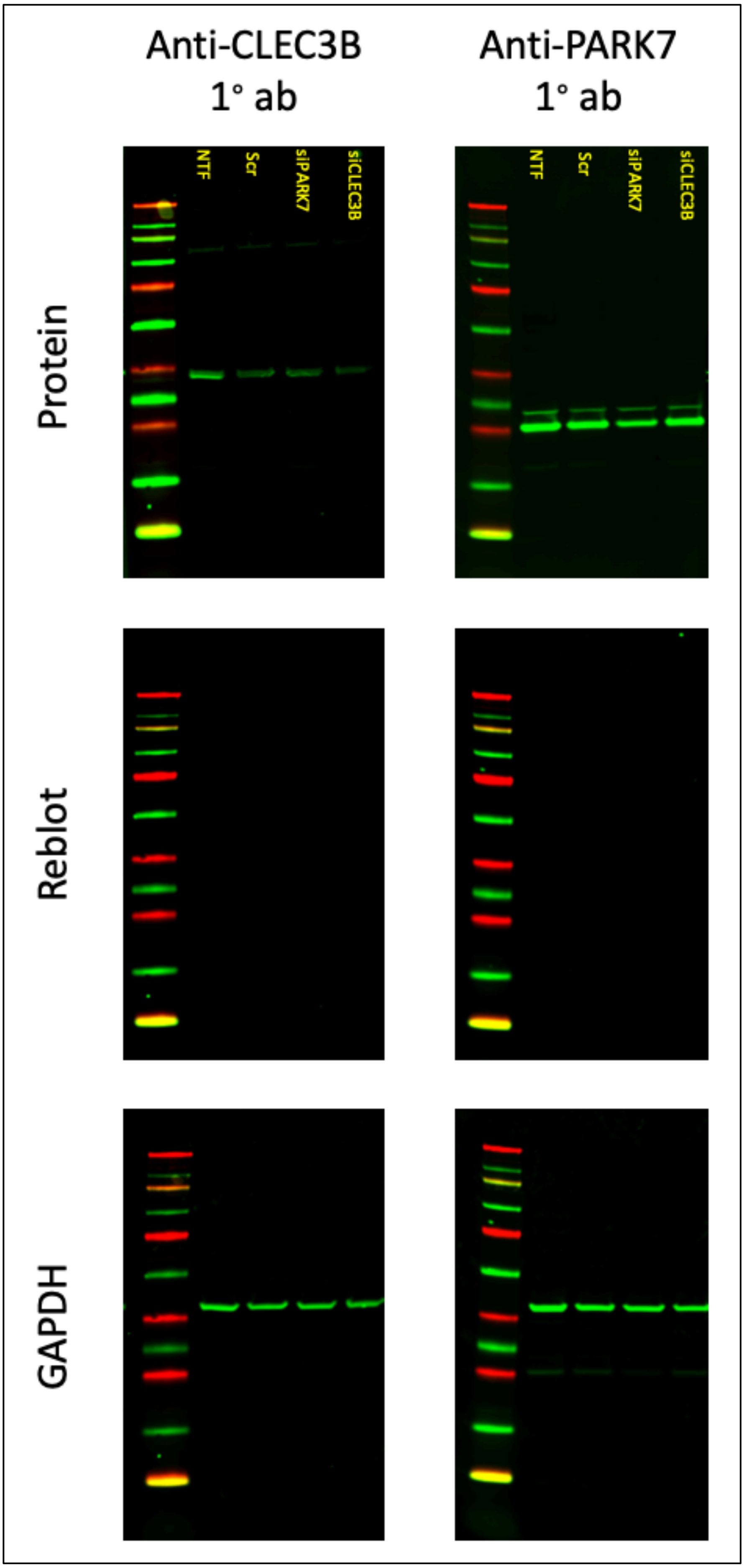


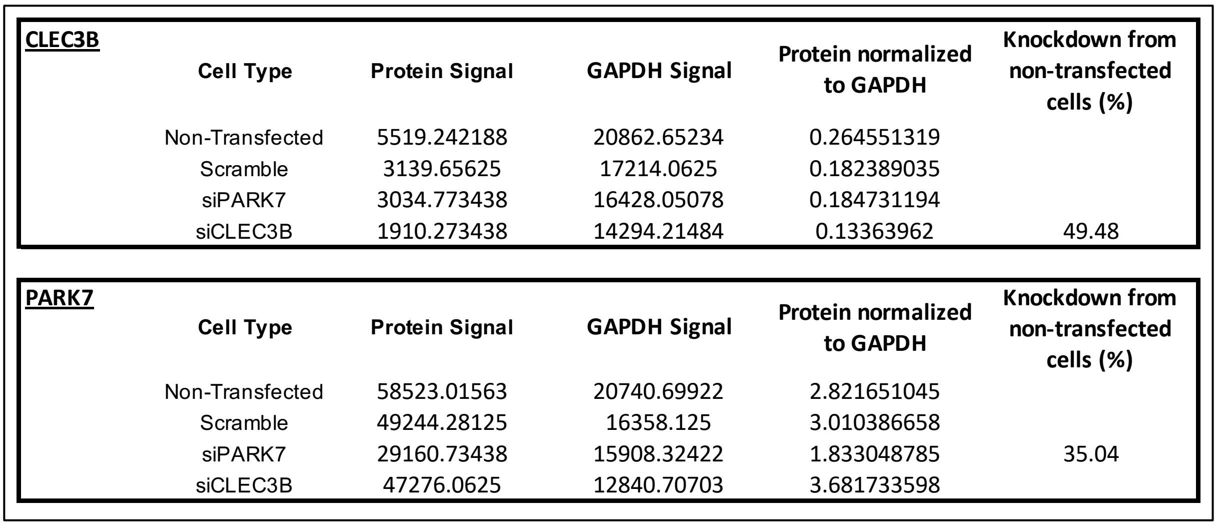
Figure 6.17: Western blot for PARK7 and CLEC3B knockdown.

Top panels show Western blot results with specific protein directed primary antibodies as shown above each column.

Middle panels demonstrate complete stripping of antibody from membrane before analysing GAPDH.

Lower panels show results of primary antibody against GAPDH, used to normalize protein signal for each cell type analysed.

Cell types are shown above each band indicating which cell types protein has been isolated from: NTF: Not transfected (control cells); Scr: Scramble transfection (Non-targeting siRNA); siPARK7: cells transfected with siRNA targeting PARK7; siCLEC3B: cells transfected with siRNA targeting CLEC3B.



CLEC3B	**Cell Type**	**Protein Signal**	**GAPDH Signal**	**Protein normalized to GAPDH**	**Knockdown from non-transfected cells (%)**
	Non-Transfected	5519.242188	20862.65234	0.264551319	
	Scramble	3139.65625	17214.0625	0.182389035	
	siPARK7	3034.773438	16428.05078	0.184731194	
	siCLEC3B	1910.273438	14294.21484	0.13363962	49.48

PARK7	**Cell Type**	**Protein Signal**	**GAPDH Signal**	**Protein normalized to GAPDH**	**Knockdown from non-transfected cells (%)**
	Non-Transfected	58523.01563	20740.69922	2.821651045	
	Scramble	49244.28125	16358.125	3.010386658	
	siPARK7	29160.73438	15908.32422	1.833048785	35.04
	siCLEC3B	47276.0625	12840.70703	3.681733598	

Table 6.1: Results of Western blot analysis

Showing the cell type protein was isolated from, the measured protein signal, associated GAPDH signal and normalized protein signal. Knockdown was calculated only for the transfection under test from the relevant membrane. Top panel shows results from the membrane analysed with anti-CLEC3B primary antibody and lower panel showing results from the membrane analysed with anti-PARK7 antibody.

The Western blot show a good result with protein detection at approximately 35 kDa for CLEC3B and 25 kDa for PARK7 (Figure 6.17). Analysis of the Western blot performed on cells taken at 48 hours after transfection showed knockdown of 49% for CLEC3B and 35% for PARK7 (Table 6.1).

6.3.2.4.1 Migration assays in PARK7 and CLEC3B knockdown HPAEC

Migration assays (n=3) were performed on cells at 48 hours after transfection (Figure 6.18). The migrated cell counts demonstrate a negative effect on endothelial cell migration due to the transfection process alone, with no further additional change noted with specific knockdown of CLEC3B. There is however evidence of a further reduction in endothelial cell migration in the cells specifically targeted to knockdown PARK7.

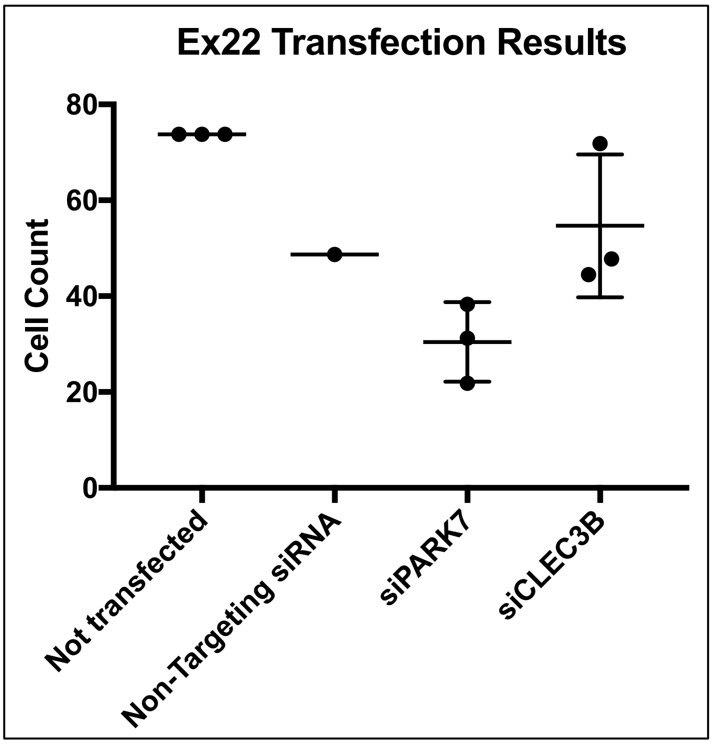


Figure 6.18: PARK7 and CLEC3B knockdown migration assay

Combined results of migration assays using cells transfected with siRNA targeting PARK7 and CLEC3B, non-targeting siRNA and non-transfected cells. All experiments normalized to the non-transfected cell count.

6.4 Discussion

In this chapter I set out to identify whether our proteins of interest are expressed in lung tissues of patients with pulmonary arterial hypertension and, more specifically, whether these localize to remodelled pulmonary vessels which would suggest a closer relationship between the protein and the underlying pathophysiology of PAH, rather than expression as a response to disease such as is seen with NTproBNP expression responding to elevated right ventricular pressures. Through cell culture experiments, I then looked for a possible role in PAH pathogenesis for PARK7 and CLEC3B, as a role for these proteins is less well described than that of GDF-15 which has previously been well studied.

Immunohistochemistry was performed on sections of lung tissue taken from explanted lungs from a patient with idiopathic pulmonary arterial hypertension as the closest available model to that of systemic sclerosis related pulmonary arterial hypertension. The same remodelled vessels were used in each disease example.

IHC demonstrates the widespread expression of PARK7 protein in lung parenchyma and alveolar cells in both healthy and diseased tissues. In contrast to healthy tissues, in PAH there is significantly greater expression of PARK7 in pulmonary vascular tissues, with particular localisation to the vascular endothelium. This change is consistent with the previously described increase in serum concentrations of PARK7 in patients with SSc-PAH compared to control, and suggests that PARK7 may have a direct role in the vasculopathy underlying PAH. In the study of bone fracture repair and wound healing, Kim & Shin et al. associated PARK7 expression with induction of angiogenesis in endothelial cells through stimulation of the fibroblast growth factor 1 (FGF-1) receptor.(Kim et al., 2012) My in vitro experiments could not replicate angiogenesis finding in human pulmonary artery endothelial cells as these experiments failed, however a small but significant increase in endothelial cell migration was seen when HPAEC's were stimulated with PARK7 protein. When PARK7 is selectively knocked down, endothelial cell migration is negatively affected. These results support the previously described role for PARK7 seen in wound healing and tissue regeneration, and these findings in pulmonary artery endothelial cells support the hypothesis that this protein plays an important role in the underlying pathophysiology of PAH.

Staining for CLEC3B in PAH tissues shows significantly increased expression of this protein in the pulmonary vascular tissue, localising particularly to the smooth muscle and endothelial layers in contrast to control tissues where there is no staining seen. This apparent increase in CLEC3B at the tissue level is in contrast to the results of our serum studies which reported reduced concentrations of CLEC3B in patients with PAH. Similar contrasting results were noted by Chen & Han et al. who found a decreased serum concentration of CLEC3B in patients with coronary artery disease, but a higher concentration in histological sections of diseased coronary arteries as compared to controls.(Chen et al., 2015) It is therefore possible that CLEC3B plays an important role in multiple vascular pathologies, being sequestered from the circulation by diseased and remodelling vessels. No studies have reported an association between CLEC3B and PAH. Stimulation of pulmonary vascular endothelial and smooth muscle cells with CLEC3B in vitro cell culture did not demonstrate any altered cell biology. Similarly, knocking down CLEC3B did not affect HPAEC migration. The known function of CLEC3B protein is to be involved in plasminogen activation and tissue remodelling, particularly reported to be involved in the invasion of malignant cancer biology, which I would suggest supports a reasonable hypothesis for a similar role in the vascular remodelling underlying PAH.(deVries et al., 1996, Hogdall et al., 2002) This proteolytic activity is likely to be a tissue based effector response to upstream PAH pathophysiology, and less likely to act as a pathway mediator of endothelial or smooth muscle cell activation.

GFRAL, a receptor for the signalling of GDF-15, has not previously been reported outside the central nervous system. Although GDF-15 has been widely reported in PAH, the signalling pathway for its action has not been identified. IHC on diseased vessels in PAH demonstrates staining for the presence of GFRAL, which to my knowledge is the first report of this receptor in tissues of patients with PAH. The presence of this receptor in lung tissues would provide a link for the action of GDF-15 to the pathological phenotype.

IHC staining in tissues taken from rat models of PAH demonstrates similar staining for PARK7 and CLEC3B to that of human tissues, suggesting that a rat model of disease could be used to test for the effect of these proteins, or the knockout of these proteins in in-vivo experimentation. GFRAL did not show evidence of staining in the rat model.

7 General Discussion

Most previous research, which currently guides treatment decisions in SSc-PAH, has included this subgroup of patients in the larger group under the heading “pulmonary arterial hypertension” due to the apparent pathophysiological similarities understood at the time. It is rapidly becoming clear that SSc-PAH is a distinct pathophysiological condition, separated from idiopathic pulmonary arterial hypertension by differences in disease progression and mortality; treatment response; and histological findings.

Within systemic sclerosis, prevalence of pulmonary hypertension is relatively high, with a currently established screening programme based on expert consensus through international guidelines and the published data from the DETECT study. Small studies have suggested that diagnosing and treating patients with SSc-PAH earlier in their condition confers a survival advantage. The DETECT algorithm has significant limitations, in that it relies on the availability of respiratory function tests, the availability of and intra-observer variability of echocardiography, and a returns high false positive rate of 65% leading to an excess of invasive right heart catheter studies being undertaken.

We have demonstrated a potential predictive model based solely on the results of a peripheral blood test, a procedure which is generally considered simple and acceptable to patients. This model positions itself as a tool for use in general rheumatology to identify patients with pulmonary arterial hypertension, without the multi-modality testing which is currently required. It is likely that echocardiographic estimation of PASP and right heart function will remain as a major investigation for the determination of the need for definitive invasive testing.

The combined measurement of Tetranectin, Protein DJ-1 and Growth differentiation factor 15 produced a predictive tool which could classify patients with SSc into those with and without PAH with a diagnostic accuracy of 85%. The biology of Tetranectin and Protein DJ-1 have not previously been reported in pulmonary arterial hypertension. Immunohistochemistry suggests increases in these protein concentrations in lung tissues from patients with pulmonary arterial hypertension. Furthermore, these proteins appear to be expressed in animal models of PAH which could be used for further investigation of their pathobiology. Cell biology experiments lacked a strong signal to explain a clear mechanism of action of these proteins however generally supported theoretical mechanisms extrapolated from other pathological conditions.

To further this work, it would be useful to investigate the expression and relevance of these proteins in an unselected group of patients with systemic sclerosis, to gain some data on their potential as prognostic biomarkers, and as a true measure of their ability to detect disease at an early stage. This would require a protracted period of patient follow up from an unselected baseline rheumatology cohort to determine those that would go on to develop pulmonary hypertension. Within this cohort, it would also be useful to understand how change in these biomarkers reflects change in pulmonary vascular status over time, and to investigate any role for them in predicting response to treatment. This could be done by analysis of longitudinal samples during patient visits at the pulmonary vascular disease unit, and around the time of change in medication. Further work is also required to examine the biological mechanisms of these proteins in their relationship to pulmonary vascular disease.

7.1 Limitations of this study

There were several important limitations to this work that are important to acknowledge as they may have an influence on the conclusions drawn.

Study size

Due to large expected natural intra- and inter-person variability, finding a characteristic pattern which accurately distinguishes one condition between two groups is very difficult with a small cohort of patients. Most predictive tools are derived and based on the study of large cohorts of patients, and therefore these tools can be developed using relatively simple statistical methods. Due to the rarity of PAH as a condition, large cohorts of patients, particularly with specific disease subtypes such as our cohort with systemic sclerosis related disease, are difficult to obtain without large scale and possibly international collaboration. Collaborative approaches to these questions raise alternative challenges such as differences in patient selection, screening programmes, severity of disease, and treatment status amongst other differences affecting patients recruited to these registries. We acknowledge the small number of patient samples in our study, but were careful in the selection of appropriate patients with careful review of notes and diagnostic investigations before including patients in the analysis. It was ambitious from the outset to approach the question of predictive classification modelling in a cohort with small sample numbers. The small number available required a complicated approach to the statistical analysis, using statistical methods and machine learning algorithms specifically designed to handle data with this problem, however despite this we must recognise that without a broad patient sample, the risk of overfitting a statistical model to our derivation cohort remained very high.

Derivation array

Although the Myriad RBM platform is calibrated and controlled to clinical laboratory standards for absolute quantification of protein concentration, which provides much tighter quality control than other available proteomic arrays, the breadth of the proteome tested is much smaller. The DiscoveryMAP array is not specific to cardiovascular diseases, lending itself to testing a wide variety of research questions, but as a consequence is more limited in terms of returning data focussed on the proteomic spectrum more specifically associated with cardiovascular physiology. In contrast, other platforms which provide relative quantification can measure vastly greater number of analytes in order to include this level of proteomic detail. It is reasonable to suggest that, on this basis, there may be other protein targets which we have not measured which may provide better classification of disease in our cohort of patients.

Multiplex assays in general have some limitation in terms of specificity when it comes to measuring multiple targets in the same assay. Single target assays such as ELIZA remain the gold standard as they can be specifically tailored for the detection of a single target. As the number of analytes increases in a multiplex assay, so does the risk of reducing specificity. Several patient factors have previously been shown to interfere with the specificity of results measured in multiplex systems, so absolute protein results need to be interpreted with caution and ideally subsequently confirmed with single target assays.(Christiansson et al., 2014)

Patient Cohort

The aim of this study was to examine whether a protein or panel of proteins could be derived which would accurately predict the presence of PAH in a cohort of patients with SSc. Underlying this aim is the knowledge that patients with SSc have a high lifetime prevalence of PAH, and that early detection and initiation of targeted treatments can significantly alter potential outcomes for patients. It is important for the applicability of a predictive model to be derived from the study of patients from a representative population. The patients included in our study were the best available to us, but by no means perfect. They were treatment naïve patients, recruited at first presentation to the pulmonary hypertension unit after referral from the parent rheumatology team. They were therefore patients for whom some suspicion of pulmonary hypertension had already been raised, perhaps due to the development of typical symptoms. The ideal cohort from which to recruit patients would be the rheumatology clinic, from unselected patients with systemic sclerosis, and then follow these to see which would go on to develop pulmonary hypertension. The time course required with this study design made this approach impractical to take forward in this study. The outcome is that true application of this tool is for the prediction of PAH in SSc patients at the point of referral for investigation of potential PAH. The model requires rigorous validation in an unselected SSc cohort from the rheumatology clinic before it can be reliably applied in this setting.

SSc-no PH Control Group

I have already discussed the limitation caused by the recruitment of patients after referral for PH investigation, however this is likely to more significantly affect the SSc-no PH control cohort as these patients were also recruited from the pool of patients referred for investigation of possible PH. These patients were proven not to have pulmonary hypertension at cardiac catheterisation, and to our knowledge none went on to develop PH during subsequent follow up. These patients cannot however be considered “normal” patients with SSc but without PH as there must have been some element of their symptomatology or monitoring which raised enough concern about the possibility of PH to prompt their referral.

In depth analysis of the phenotype of patients included in the SSc-no PH control arm demonstrates significant abnormalities in the baseline cardiorespiratory physiology (Table 4.1) which may differ from other patients with SSc but without PH. Although less severe than for patients in the SSc-PAH cohort, the TLCO is reduced from normal in the control cohort. The pulmonary haemodynamics are also abnormal (mPAP 21 mmHg, normal considered to be <14 mmHg), although do not meet the threshold for diagnosis of pulmonary hypertension. The ideal study would have recruited control patients from an unselected population of SSc patients in a rheumatology clinic, who would then need to have undergone invasive measurement of pulmonary haemodynamics. Subjecting normal patients to invasive medical testing in this way is ethically questionable making recruitment in this way impractical.

External validation cohort

After deriving our classification model, and internal validation in our cohort of patients, we went on to validate in an external group of patients recruited from international collaborators. It was difficult to identify a centre which recruited patients in a similar way to us, at treatment naïve baseline. Both centres providing validation samples also provided matching phenotype data which demonstrated that the majority of patients provided were prevalent PAH patients already well established on treatment and therefore potentially at a significantly different disease timepoint than our target diagnostic stage patient. There were also some subtle differences between centres in patient selection, sample storage and the processing of biological samples from the study patients. All of these factors may potentially alter the proteomic profile present for analysis and alter the outcome of validation.

7.2 Reflections on this work

Given the time to conduct this research again and in the ideal situation I would have spent less time on the derivation cohort phenotyping exercise early on in the project. This was not lost time however, as a tightly phenotyped cohort was fundamental to understanding the sources of error in some of the earlier works. It was during this early period that I learned my skills in research programming which are the foundation of this type of research, having never used these platforms previously.

The work would be improved by closer links with rheumatology as a speciality, and the ability to develop a cohort of unselected patients at least to act as a control cohort, to more clearly demonstrate the protein concentration changes against those patients who have developed pulmonary arterial hypertension. It would also benefit greatly from a larger patient cohort as has been demonstrated throughout.

This work has produced a predictive protein panel capable of predicting pulmonary arterial hypertension in patients with systemic sclerosis, but the pathophysiological mechanisms of these proteins remain unclear. This is something that I would have liked to have developed further through cell biology and potentially in vivo models.

7.3 Closing comments

There have been a number of assay platforms recently developed which can return precise measurement data on a very large number of analytes. The bio-informatic methodology used can be easily translated into other classifying problems which are affected by small sample numbers and a very large number of independent variables.

8 References

Abdi, H. and L. J. Williams (2010). Principle component analysis. *WIREs Computational Statistics* 2(4) 433-459.

Abraham, D. and O. Distler (2007). How does endothelial cell injury start? The role of endothelin in systemic sclerosis. *Arthritis Res Ther* 9 Suppl 2 S2.

Abraham, D. J., T. Krieg, J. Distler, et al. (2009). Overview of pathogenesis of systemic sclerosis. *Rheumatology (Oxford)* 48 Suppl 3 iii3-7.

Adela, R. and S. K. Banerjee (2015). GDF-15 as a Target and Biomarker for Diabetes and Cardiovascular Diseases: A Translational Prospective. *J Diabetes Res* 2015 490842.

Agresti, A. (2007). *An introduction to categorical data analysis*: Wiley-Interscience.

Altorok, N., Y. Wang and B. Kahaleh (2014). Endothelial dysfunction in systemic sclerosis. *Curr Opin Rheumatol* 26(6) 615-620.

Andreassen, A. K., R. Wergeland, S. Simonsen, et al. (2006). N-terminal pro-B-type natriuretic peptide as an indicator of disease severity in a heterogeneous group of patients with chronic precapillary pulmonary hypertension. *Am J Cardiol* 98(4) 525-529.

Angelini, D. J., Q. Su, K. Yamaji-Kegan, et al. (2009). Resistin-like molecule-beta in scleroderma-associated pulmonary hypertension. *Am J Respir Cell Mol Biol* 41(5) 553-561.

Avouac, J., P. Airo, C. Meune, et al. (2010). Prevalence of pulmonary hypertension in systemic sclerosis in European Caucasians and metaanalysis of 5 studies. *J Rheumatol* 37(11) 2290-2298.

Avouac, J., C. Meune, C. Chenevier-Gobeaux, et al. (2015). Cardiac biomarkers in systemic sclerosis: contribution of high-sensitivity cardiac troponin in addition to N-terminal pro-brain natriuretic peptide. *Arthritis Care Res (Hoboken)* 67(7) 1022-1030.

Badesch, D. B., G. E. Raskob, C. G. Elliott, et al. (2010). Pulmonary arterial hypertension: baseline characteristics from the REVEAL Registry. *Chest* 137(2) 376-387.

Biomarkers Definitions Working, G. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 69(3) 89-95.

Boucly, A., J. Weatherald, L. Savale, et al. (2017). Risk assessment, prognosis and guideline implementation in pulmonary arterial hypertension. *Eur Respir J* 50(2).

Budhiraja, R., R. M. Tuder and P. M. Hassoun (2004). Endothelial dysfunction in pulmonary hypertension. *Circulation* 109(2) 159-165.

Cacoub, P., M. Karmochkine, R. Dorent, et al. (1996). Plasma levels of thrombomodulin in pulmonary hypertension. *American Journal of Medicine* 101(2) 160-164.

Carpentier, G. *Angiogenesis Analyzer.* [*online*]. Available at: https://imagej.nih.gov/ij/macros/toolsets/Angiogenesis Analyzer.txt [Accessed 13th March 2017.

Chen, Y. J., H. Han, X. X. Yan, et al. (2015). Tetranectin as a Potential Biomarker for Stable Coronary Artery Disease. *Scientific Reports* 5.

Christiansson, L., S. Mutjoki, B. Simonsson, et al. (2014). The use of multiplex platforms for absolute and relative protein quantification of clinical material. *EuPA Open Proteomics* 3 37-47.

Chung, L., R. M. Fairchild, D. E. Furst, et al. (2017). Utility of B-type natriuretic peptides in the assessment of patients with systemic sclerosis-associated pulmonary hypertension in the PHAROS registry. *Clin Exp Rheumatol* 35 Suppl 106(4) 106-113.

Coghlan, J. G., C. P. Denton, E. Grunig, et al. (2014). Evidence-based detection of pulmonary arterial hypertension in systemic sclerosis: the DETECT study. *Ann Rheum Dis* 73(7) 1340-1349.

Coghlan, J. G., M. Wolf, O. Distler, et al. (2018). Incidence of pulmonary hypertension and determining factors in patients with systemic sclerosis. *Eur Respir J* 51(4).

Colle, I. O., R. Moreau, E. Godinho, et al. (2003). Diagnosis of portopulmonary hypertension in candidates for liver transplantation: a prospective study. *Hepatology* 37(2) 401-409.

Condliffe, R. and G. Kovacs (2018). Identifying early pulmonary arterial hypertension in patients with systemic sclerosis. *Eur Respir J* 51(4).

Conley, B. A., J. D. Smith, M. Guerrero-Esteo, et al. (2000). Endoglin, a TGF-beta receptor-associated protein, is expressed by smooth muscle cells in human atherosclerotic plaques. *Atherosclerosis* 153(2) 323-335.

Cool, C. D., D. Kennedy, N. F. Voelkel, et al. (1997). Pathogenesis and evolution of plexiform lesions in pulmonary hypertension associated with scleroderma and human immunodeficiency virus infection. *Hum Pathol* 28(4) 434-442.

Cool, C. D., N. F. Voelkel and T. Bull (2011). Viral infection and pulmonary hypertension: is there an association? *Expert Rev Respir Med* 5(2) 207-216.

Coral-Alvarado, P. X., M. F. Garces, J. E. Caminos, et al. (2010). Serum endoglin levels in patients suffering from systemic sclerosis and elevated systolic pulmonary arterial pressure. *Int J Rheumatol* 2010.

Dalonzo, G. E., R. J. Barst, S. M. Ayres, et al. (1991). Survival in Patients with Primary Pulmonary-Hypertension - Results from a National Prospective Registry. *Annals of Internal Medicine* 115(5) 343-349.

Deng, Z. M., J. H. Morse, S. L. Slager, et al. (2000). Familial primary pulmonary hypertension (gene PPH1) is caused by mutations in the bone morphogenetic protein receptor-II gene. *American Journal of Human Genetics* 67(3) 737-744.

deVries, T. J., P. E. J. deWit, I. Clemmensen, et al. (1996). Tetranectin and plasmin/plasminogen are similarly distributed at the invasive front of cutaneous melanoma lesions. *Journal of Pathology* 179(3) 260-265.

Eddahibi, S., M. Humbert, S. Sediame, et al. (2000). Imbalance between platelet vascular endothelial growth factor and platelet-derived growth factor in pulmonary hypertension. Effect of prostacyclin therapy. *Am J Respir Crit Care Med* 162(4 Pt 1) 1493-1499.

Edelman, J. D. (2007). Clinical presentation, differential diagnosis, and vasodilator testing of pulmonary hypertension. *Semin Cardiothorac Vasc Anesth* 11(2) 110-118.

Fijalkowska, A., M. Kurzyna, A. Torbicki, et al. (2006). Serum N-terminal brain natriuretic peptide as a prognostic parameter in patients with pulmonary hypertension. *Chest* 129(5) 1313-1321.

Fishman, A. P. (2004). Primary pulmonary arterial hypertension - A look back. *Journal of the American College of Cardiology* 43(12) 2s-4s.

Gabrielli, A., E. V. Avvedimento and T. Krieg (2009). Scleroderma. *N Engl J Med* 360(19) 1989-2003.

Galie, N., M. Humbert, J. L. Vachiery, et al. (2016). 2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European

Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT). *Eur Heart J* 37(1) 67-119.

Galie, N., V. V. McLaughlin, L. J. Rubin, et al. (2019). An overview of the 6th World Symposium on Pulmonary Hypertension. *Eur Respir J* 53(1).

Gatzoulis, M. A., R. Alonso-Gonzalez and M. Beghetti (2009). Pulmonary arterial hypertension in paediatric and adult patients with congenital heart disease. *Eur Respir Rev* 18(113) 154-161.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochem Med (Zagreb)* 25(2) 141-151.

Gold, L., D. Ayers, J. Bertino, et al. (2010). Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* 5(12) e15004.

Gore, B., M. Izikki, O. Mercier, et al. (2014). Key role of the endothelial TGF-beta/ALK1/endoglin signaling pathway in humans and rodents pulmonary hypertension. *PLoS One* 9(6) e100310.

Hachulla, E., V. Gressin, L. Guillevin, et al. (2005). Early detection of pulmonary arterial hypertension in systemic sclerosis: a French nationwide prospective multicenter study. *Arthritis Rheum* 52(12) 3792-3800.

Hadengue, A., M. K. Benhayoun, D. Lebrec, et al. (1991). Pulmonary hypertension complicating portal hypertension: prevalence and relation to splanchnic hemodynamics. *Gastroenterology* 100(2) 520-528.

Hao, Y., V. Thakkar, W. Stevens, et al. (2015). A comparison of the predictive accuracy of three screening models for pulmonary arterial hypertension in systemic sclerosis. *Arthritis Res Ther* 17 7.

Hatano, S. and T. Strasser (1975). Primary Pulmonary Hypertension. Report on a WHO Meeting. Geneva, World Health Organization.

Heath, D. and J. E. Edwards (1958). Pathology of Hypertensive Pulmonary Vascular Disease - Description of 6 Grades of Structural Changes in the Pulmonary Arteries with Special Reference to Congenital Cardiac Septal Defects. *Circulation* 18(4) 533-547.

Hickey, P. M., A. Lawrie and R. Condliffe (2018). Circulating Protein Biomarkers in Systemic Sclerosis Related Pulmonary Arterial Hypertension: A Review of Published Data. *Front Med (Lausanne)* 5 175.

Hoeper, M. M., T. Kramer, Z. Pan, et al. (2017). Mortality in pulmonary arterial hypertension: prediction by the 2015 European pulmonary hypertension guidelines risk stratification model. *Eur Respir J* 50(2).

Hogdall, C. K., I. J. Christensen, R. W. Stephens, et al. (2002). Serum tetranectin is an independent prognostic marker in colorectal cancer and weakly correlated with plasma suPAR, plasma PAI-1 and serum CEA. *APMIS* 110(9) 630-638.

Horstmeyer, A., C. Licht, G. Scherr, et al. (2005). Signalling and regulation of collagen I synthesis by ET-1 and TGF-beta1. *FEBS J* 272(24) 6297-6309.

Humbert, M. and H. A. Ghofrani (2016). The molecular targets of approved treatments for pulmonary arterial hypertension. *Thorax* 71(1) 73-83.

Humbert, M., O. Sitbon, A. Chaouat, et al. (2006). Pulmonary arterial hypertension in France: results from a national registry. *Am J Respir Crit Care Med* 173(9) 1023-1030.

Humbert, M., A. Yaici, P. de Groote, et al. (2011). Screening for pulmonary arterial hypertension in patients with systemic sclerosis: clinical characteristics at diagnosis and long-term survival. *Arthritis Rheum* 63(11) 3522-3530.

Hurdman, J., R. Condliffe, C. A. Elliot, et al. (2012). ASPIRE registry: assessing the Spectrum of Pulmonary hypertension Identified at a REferral centre. *Eur Respir J* 39(4) 945-955.

Kiely, D. G., C. A. Elliot, I. Sabroe, et al. (2013). Pulmonary hypertension: diagnosis and management. *BMJ* 346 f2028.

Kim, J. M., H. I. Shin, S. S. Cha, et al. (2012). DJ-1 promotes angiogenesis and osteogenesis by activating FGF receptor-1 signaling. *Nature Communications* 3.

Kylhammar, D., B. Kjellstrom, C. Hjalmarsson, et al. (2018). A comprehensive risk stratification at early follow-up determines prognosis in pulmonary arterial hypertension. *European Heart Journal* 39(47) 4175-4181.

Lee, C. K., H. J. Park, H. H. So, et al. (2006). Proteomic profiling and identification of cofilin responding to oxidative stress in vascular smooth muscle. *Proteomics* 6(24) 6455-6475.

Leuchte, H. H., M. Holzapfel, R. A. Baumgartner, et al. (2004). Clinical significance of brain natriuretic peptide in primary pulmonary hypertension. *Journal of the American College of Cardiology* 43(5) 764-770.

Li-Cor (2015). Chameleon Duo Pre-Stained Protein Ladder.

Ling, Y., M. K. Johnson, D. G. Kiely, et al. (2012). Changing demographics, epidemiology, and survival of incident pulmonary arterial hypertension: results from the pulmonary hypertension registry of the United Kingdom and Ireland. *Am J Respir Crit Care Med* 186(8) 790-796.

Machado, R. D., L. Southgate, C. A. Eichstaedt, et al. (2015). Pulmonary Arterial Hypertension: A Current Perspective on Established and Emerging Molecular Genetic Defects. *Hum Mutat* 36(12) 1113-1127.

Manichaikul, A., L. Sun, A. C. Borczuk, et al. (2017). Plasma Soluble Receptor for Advanced Glycation End Products in Idiopathic Pulmonary Fibrosis. *Ann Am Thorac Soc* 14(5) 628-635.

Mathai, S. C., M. Bueso, L. K. Hummers, et al. (2010). Disproportionate elevation of N-terminal pro-brain natriuretic peptide in scleroderma-related pulmonary hypertension. *European Respiratory Journal* 35(1) 95-104.

McMahan, Z., F. Schoenhoff, J. E. Van Eyk, et al. (2015). Biomarkers of pulmonary hypertension in patients with scleroderma: a case-control study. *Arthritis Res Ther* 17 201.

Meadows, C. A., M. G. Risbano, L. Zhang, et al. (2011). Increased expression of growth differentiation factor-15 in systemic sclerosis-associated pulmonary arterial hypertension. *Chest* 139(5) 994-1002.

Meloche, J., A. Courchesne, M. Barrier, et al. (2013). Critical role for the advanced glycation end-products receptor in pulmonary arterial hypertension etiology. *J Am Heart Assoc* 2(1) e005157.

Mercie, P., M. Seigneur, J. Constans, et al. (1997). [Assay of plasma thrombomodulin in systemic diseases]. *Rev Med Interne* 18(2) 126-131.

Mukerjee, D., D. St George, B. Coleiro, et al. (2003). Prevalence and outcome in systemic sclerosis associated pulmonary arterial hypertension: application of a registry approach. *Ann Rheum Dis* 62(11) 1088-1093.

Nagaya, N., T. Nishikimi, M. Uematsu, et al. (2000). Plasma brain natriuretic peptide as a prognostic indicator in patients with primary pulmonary hypertension. *Circulation* 102(8) 865-870.

Nickel, N., H. Golpon, M. Greer, et al. (2012). The prognostic impact of follow-up assessments in patients with idiopathic pulmonary arterial hypertension. *Eur Respir J* 39(3) 589-596.

Nickel, N., D. Jonigk, T. Kempf, et al. (2011). GDF-15 is abundantly expressed in plexiform lesions in patients with pulmonary arterial hypertension and affects proliferation and apoptosis of pulmonary endothelial cells. *Respir Res* 12 62.

Nickel, N., T. Kempf, H. Tapken, et al. (2008). Growth differentiation factor-15 in idiopathic pulmonary arterial hypertension. *Am J Respir Crit Care Med* 178(5) 534-541.

Nicolls, M. R., L. Taraseviciene-Stewart, P. R. Rai, et al. (2005). Autoimmunity and pulmonary hypertension: a perspective. *Eur Respir J* 26(6) 1110-1118.

Obrist, P., G. Spizzo, C. Ensinger, et al. (2004). Aberrant tetranectin expression in human breast carcinomas as a predictor of survival. *Journal of Clinical Pathology* 57(4) 417-421.

Odler, B., V. Foris, A. Gungl, et al. (2018). Biomarkers for Pulmonary Vascular Remodeling in Systemic Sclerosis: A Pathophysiological Approach. *Front Physiol* 9 587.

Olsson, A. K., A. Dimberg, J. Kreuger, et al. (2006). VEGF receptor signalling - in control of vascular function. *Nat Rev Mol Cell Biol* 7(5) 359-371.

Overbeek, M. J., M. C. Vonk, A. Boonstra, et al. (2009). Pulmonary arterial hypertension in limited cutaneous systemic sclerosis: a distinctive vasculopathy. *Eur Respir J* 34(2) 371-379.

Papaioannou, A. I., E. Zakynthinos, K. Kostikas, et al. (2009). Serum VEGF levels are related to the presence of pulmonary arterial hypertension in systemic sclerosis. *BMC Pulm Med* 9 18.

Parasuraman, S., S. Walker, B. L. Loudon, et al. (2016). Assessment of pulmonary artery pressure by echocardiography-A comprehensive review. *Int J Cardiol Heart Vasc* 12 45-51.

Park, J. E., H. H. Chen, J. Winer, et al. (1994). Placenta Growth-Factor - Potentiation of Vascular Endothelial Growth-Factor Bioactivity, in-Vitro and in-Vivo, and High-Affinity Binding to Flt-1 but Not to Flk-1/Kdr. *Journal of Biological Chemistry* 269(41) 25646-25654.

PH, N. A. o. (2014). National Audit of Pulmonary Hypertension 2014.

Price, L. C., S. J. Wort, F. Perros, et al. (2012). Inflammation in pulmonary arterial hypertension. *Chest* 141(1) 210-221.

Ramjug, S., N. Hussain, J. Hurdman, et al. (2017). Idiopathic and Systemic Sclerosis-Associated Pulmonary Arterial Hypertension: A Comparison of Demographic, Hemodynamic, and MRI Characteristics and Outcomes. *Chest* 152(1) 92-102.

Rhee, R. L., N. B. Gabler, S. Sangani, et al. (2015). Comparison of Treatment Response in Idiopathic and Connective Tissue Disease-associated Pulmonary Arterial Hypertension. *Am J Respir Crit Care Med* 192(9) 1111-1117.

Rhodes, C. J., J. Wharton, P. Ghataorhe, et al. (2017). Plasma proteome analysis in patients with pulmonary arterial hypertension: an observational cohort study. *Lancet Respir Med* 5(9) 717-726.

Schindelin, J., I. Arganda-Carreras, E. Frise, et al. (2012). Fiji: an open-source platform for biological-image analysis. *Nat Methods* 9(7) 676-682.

Shibuya, M. (2011). Vascular Endothelial Growth Factor (VEGF) and Its Receptor (VEGFR) Signaling in Angiogenesis: A Crucial Target for Anti- and Pro-Angiogenic Therapies. *Genes Cancer* 2(12) 1097-1105.

Simonneau, G., D. Montani, D. S. Celermajer, et al. (2019). Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *Eur Respir J* 53(1).

Sitbon, O., C. Lascoux-Combe, J. F. Delfraissy, et al. (2008). Prevalence of HIV-related pulmonary arterial hypertension in the current antiretroviral therapy era. *Am J Respir Crit Care Med* 177(1) 108-113.

Sohail, A., H. B. Korejo, A. S. Shaikh, et al. (2019). Correlation between Echocardiography and Cardiac Catheterization for the Assessment of Pulmonary Hypertension in Pediatric Patients. *Cureus* 11(8) e5511.

Steen, V. D. (2005). Autoantibodies in systemic sclerosis. *Semin Arthritis Rheum* 35(1) 35-42.

Stratton, R. J., L. Pompon, J. G. Coghlan, et al. (2000). Soluble thrombomodulin concentration is raised in scleroderma associated pulmonary hypertension. *Ann Rheum Dis* 59(2) 132-134.

Strimbu, K. and J. A. Tavel (2010). What are biomarkers? *Curr Opin HIV AIDS* 5(6) 463-466.

Taira, T., Y. Saito, T. Niki, et al. (2004). DJ-1 has a role in antioxidative stress to prevent cell death. *Embo Reports* 5(2) 213-218.

Tyndall, A. J., B. Bannert, M. Vonk, et al. (2010). Causes and risk factors for death in systemic sclerosis: a study from the EULAR Scleroderma Trials and Research (EUSTAR) database. *Ann Rheum Dis* 69(10) 1809-1815.

Vandecasteele, E., B. Drieghe, K. Melsens, et al. (2017). Screening for pulmonary arterial hypertension in an unselected prospective systemic sclerosis cohort. *Eur Respir J* 49(5).

Vasseur, S., S. Afzal, J. Tardivel-Lacombe, et al. (2009). DJ-1/PARK7 is an important mediator of hypoxia-induced cellular responses. *Proc Natl Acad Sci U S A* 106(4) 1111-1116.

Voelkel, N. F. and J. Gomez-Arroyo (2014). The role of vascular endothelial growth factor in pulmonary arterial hypertension. The angiogenesis paradox. *Am J Respir Cell Mol Biol* 51(4) 474-484.

Wagenvoo.Ca and Wagenvoo.N (1970). Primary Pulmonary Hypertension - a Pathologic Study of Lung Vessels in 156 Clinically Diagnosed Cases. *Circulation* 42(6) 1163-&.

Walker, U. A., A. Tyndall, L. Czirjak, et al. (2007). Clinical risk assessment of organ manifestations in systemic sclerosis: a report from the EULAR Scleroderma Trials And Research group database. *Ann Rheum Dis* 66(6) 754-763.

Warwick, G., P. S. Thomas and D. H. Yates (2008). Biomarkers in pulmonary hypertension. *European Respiratory Journal* 32(2) 503-512.

Welsh, B. T. and J. Mapes. *An overview of assay quality systems at Myriad RBM, Inc.* [*online*]. Available at: https://myriadrbm.com/quality-control-white-paper/.

Williams, M. H., C. E. Handler, R. Akram, et al. (2006). Role of N-terminal brain natriuretic peptide (N-TproBNP) in scleroderma-associated pulmonary arterial hypertension. *Eur Heart J* 27(12) 1485-1494.

Won, K. J., S. H. Jung, S. H. Jung, et al. (2014). DJ-1/park7 modulates vasorelaxation and blood pressure via epigenetic modification of endothelial nitric oxide synthase. *Cardiovascular Research* 101(3) 473-481.

9 Appendices

9.1 Appendix 1 – Protein decode

Analytes	Symbol
A disintegrin and metalloproteinase with thrombospondin motifs 8	ADAMTS8
6Ckine	CCL21
Adiponectin	ADIPOQ
Adrenomedullin	ADM
Aggrecan core protein	ACAN
Aldose Reductase	AKR1B1
Alpha-1-acid glycoprotein 1	ORM1
Alpha-1-Antitrypsin	SERPINA1
Alpha-1-Microglobulin	AMBP
Alpha-2-Macroglobulin	A2M
Alpha-Fetoprotein	AFP
Amphiregulin	AREG
Angiogenin	ANG
Angiopoietin-1	ANGPT1
Angiopoietin-2	ANGPT1
Angiopoietin-related protein 4	ANGPTL4
Angiotensin-Converting Enzyme	ACE
Antileukoproteinase	SLPI
Antithrombin-III	SERPINC1
Apolipoprotein	LPA
Apolipoprotein A-I	APOA1
Apolipoprotein A-II	APOA2
Apolipoprotein B	APOB
Apolipoprotein C-I	APOC1
Apolipoprotein C-III	APOC3
Apolipoprotein D	APOD
Apolipoprotein E	APOE
Apolipoprotein H	APOH
AXL Receptor Tyrosine Kinase	AXL
B cell-activating factor	TNFSF13B
B Lymphocyte Chemoattractant	CXCL13
Beta Amyloid 1-40	APP40
Beta Amyloid 1-42	APP42
Beta-2-Microglobulin	B2M
Beta-microseminoprotein	MSMB
Betacellulin	BTC
Brain-Derived Neurotrophic Factor	BDNF
C-Peptide	CPEP

C-Reactive Protein	CRP
Cadherin-1	CDH1
Calbindin	CALB1
Cancer Antigen 125	MUC16
Cancer Antigen 15-3	MUC1
Cancer Antigen 19-9	NAA15
Carbonic anhydrase 9	CA9
Carcinoembryonic Antigen	CEACAM5
Carcinoembryonic antigen-related cell adhesion molecule 1	CEACAM1
Carcinoembryonic antigen-related cell adhesion molecule 6	CEACAM6
Cathepsin B	CTSB
Cathepsin D	CTSD
C-C motif chemokine 15	CCL15
CD5 Antigen-like	CD5L
CD27 antigen	CD27
CD40 Ligand	CD40LG
CD163	CD163
Cellular Fibronectin	FN1
Ceruloplasmin	CP
Chemokine CC-4	CCR4
Chromogranin-A	CHGA
Ciliary Neurotrophic Factor	CNTF
Clusterin	CLU
Collagen IV	COL4A3
Complement C3	C3
Complement Factor H	CFH
Complement Factor H _ Related Protein 1	CFHR1
Cystatin-A	CSTA
Cystatin-B	CSTB
Cystatin-C	CST3
Decorin	DCN
Dickkopf-related protein 1	DKK1
Dipeptidyl peptidase IV	DPP4
Dopamine beta-hydroxylase	DBH
E-Selectin	SELE
EN-RAGE	S100A12
Endoglin	ENG
Endostatin	COL18A1
Eotaxin-1	CCL11

Eotaxin-2	CCL24
Eotaxin-3	CCL26
Epidermal Growth Factor	EGF
Epidermal Growth Factor Receptor	EGFR
Epiregulin	EREG
Epithelial cell adhesion molecule	EPCAM
Epithelial-Derived Neutrophil-Activating Protein 78	CXCL5
Erythropoietin	EPO
Factor VII	F7
Fas Ligand	FASLG
FASLG Receptor	FAS
Fatty Acid-Binding Protein, adipocyte	FABP4
Fatty Acid-Binding Protein, heart	FABP3
Fatty Acid-Binding Protein, liver	FABP1
Ferritin	FRTN
Fetuin-A	AHSG
Fibrinogen	FGA
Fibroblast Growth Factor 21	FGF21
Fibroblast growth factor 23	FGF23
Fibulin-1C	FBLN1
Ficolin-3	FCN3
Folate receptor gamma	FOLR3
Follicle-Stimulating Hormone	FSHB
Galectin-3	LGALS3
Gastric inhibitory polypeptide	GIP
Gelsolin	GSN
Glucagon-like Peptide 1, total	GCG
Glucose-6-phosphate Isomerase	GPI
Glutathione S-Transferase Mu 1	GSTM1
Glycogen phosphorylase isoenzyme BB	PYGB
Granulocyte Colony-Stimulating Factor	CSF3
Granulocyte-Macrophage Colony-Stimulating Factor	CSF2
Growth/differentiation factor 15	GDF15
Growth/differentiation factor 11	GDF11
Growth Hormone	GH1
Growth-Regulated alpha protein	CXCL1
Haptoglobin	HP
HE4	WFDC2
Heat-Shock protein 70	HSP71
Hemopexin	HPX
Heparin-Binding EGF-Like Growth Factor	HBEGF
Hepatocyte Growth Factor	HGF
Hepatocyte Growth Factor receptor	MET

Hepsin	HPN
Human Chorionic Gonadotropin beta	CGB5
Human Epidermal Growth Factor Receptor 2	ERBB2
Immunoglobulin A	IgA
Immunoglobulin E	IgE
Immunoglobulin M	IGHM
Insulin	INS
Insulin-like Growth Factor-Binding Protein 1	IGFBP1
Insulin-like Growth Factor-Binding Protein 2	IGFBP2
Insulin-like Growth Factor-Binding Protein 7	IGFBP7
Intercellular Adhesion Molecule 1	ICAM1
Interferon alpha	IFNA1
Interferon gamma	IFNG
Interferon gamma Induced Protein 10	CXCL10
Interferon-inducible T-cell alpha chemoattractant	CXCL11
Interleukin-1 alpha	IL1A
Interleukin-1 beta	IL1B
Interleukin-1 receptor antagonist	IL1RN
Interleukin-1 receptor type 1	IL1R1
Interleukin-1 receptor type 2	IL1R2
Interleukin-2	IL2
Interleukin-2 receptor alpha	IL2RA
Interleukin-3	IL3
Interleukin-4	IL4
Interleukin-5	IL5
Interleukin-6	IL6
Interleukin-6 receptor	IL6R
Interleukin-6 receptor subunit beta	IL6ST
Interleukin-7	IL7
Interleukin-8	CXCL8
Interleukin-10	IL10
Interleukin-12 Subunit p40	IL12B
Interleukin-12 Subunit p70	IL12A
Interleukin-13	IL13
Interleukin-15	IL15
Interleukin-16	IL16
Interleukin-17	IL17A
Interleukin-18	IL18
Interleukin-18-binding protein	IL18BP
Interleukin-22	IL22
Interleukin-23	IL23A
Interleukin-31	IL31
Kallikrein-5	KLK5

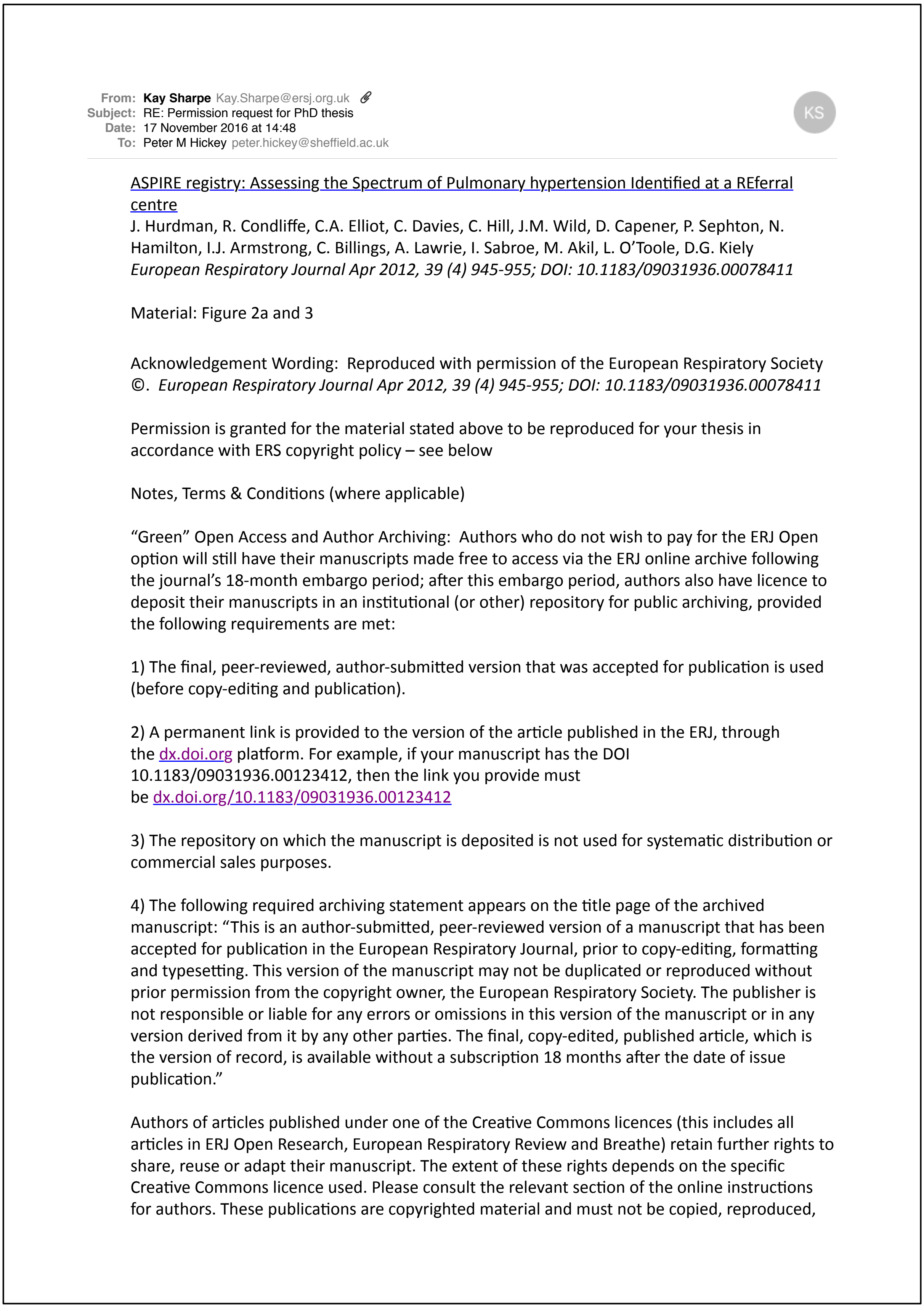
Kallikrein-7	KLK7
Kidney Injury Molecule-1	HAVCR1
Lactoferrin	LTF
Lactoylglutathione lyase	GLO1
Latency-Associated Peptide of Transforming Growth Factor beta 1	TGFB1
Lectin-Like Oxidized LDL Receptor 1	OLR1
Leptin	LEP
Leptin Receptor	LEPR
Leucine-rich alpha-2-glycoprotein	LRG1
Lipocalin-1	LCN1
Luteinizing Hormone	LHB
Macrophage Colony-Stimulating Factor 1	CSF1
Macrophage-Derived Chemokine	CCL22
Macrophage Inflammatory Protein-1 alpha	CCL3
Macrophage Inflammatory Protein-1 beta	CCL4
Macrophage Inflammatory Protein-3 alpha	CCL20
Macrophage inflammatory protein 3 beta	CCL19
Macrophage Migration Inhibitory Factor	MIF
Macrophage-Stimulating Protein	MST1
Maspin	SERPINB5
Mast/stem cell growth factor receptor	KIT
Matrix Metalloproteinase-1	MMP1
Matrix Metalloproteinase-2	MMP2
Matrix Metalloproteinase-3	MMP3
Matrix Metalloproteinase-7	MMP7
Matrix Metalloproteinase-9	MMP9
Matrix Metalloproteinase-9, total	MMP9t
Matrix Metalloproteinase-10	MMP10
Mesothelin	MSLN
MHC class I chain-related protein A	MICA
Monocyte Chemotactic Protein 1	CCL2
Monocyte Chemotactic Protein 2	CCL8
Monocyte Chemotactic Protein 3	CCL7
Monocyte Chemotactic Protein 4	CCL13
Monokine Induced by Gamma Interferon	CXCL9
Myeloid Progenitor Inhibitory Factor 1	CCL23
Myeloperoxidase	MPO
Myoglobin	MB
N-terminal prohormone of brain natriuretic peptide	NTproBNP
Nerve Growth Factor beta	NGF

Neurofilament heavy polypeptide	NEFH
Neuron-Specific Enolase	ENO2
Neuronal Cell Adhesion Molecule	NRCAM
Neuropilin-1	NRP1
Neutrophil Activating Peptide 2	PPBP
Neutrophil Gelatinase-Associated Lipocalin	LCN2
Omentin	ITLN1
Osteocalcin	BGLAP
Osteopontin	SPP1
Osteoprotegerin	TNFRSF11B
P-Selectin	SELP
Pancreatic Polypeptide	PPY
Pancreatic secretory trypsin inhibitor	SPINK1
Paraoxonase-1	PON1
Pepsinogen I	PGA
Periostin	POSTN
Peroxiredoxin-4	PRDX4
Phosphoserine Aminotransferase	PSAT1
Pigment Epithelium Derived Factor	SERPINF1
Placenta Growth Factor	PGF
Plasminogen Activator Inhibitor 1	SERPINE1
Platelet endothelial cell adhesion molecule	PECAM1
Platelet-Derived Growth Factor BB	PDGFRB
Progranulin	GRN
Prolactin	PRL
Prostasin	PRSS8
Prostate-Specific Antigen, Free	PSAf
Prostate Specific Antigen, total	PSAt
Protein DJ-1	PARK7
Protein S100-A6	S100A6
Pulmonary and Activation-Regulated Chemokine	CCL18
Pulmonary surfactant-associated protein D	SFTPD
Receptor for advanced glycosylation end products	AGER
Receptor tyrosine-protein kinase erbB-3	ERBB3
Resistin	RETN
Retinol-binding protein 4	RBP4
S100 calcium-binding protein B	S100B
Sclerostin	SOST
Secreted protein acidic and rich in cysteine	SPARC
Serum Amyloid A Protein	SAA1
Sex Hormone-Binding Globulin	SHBG
Sortilin	SORT1

Squamous Cell Carcinoma Antigen-1	SERPINB3
ST2	ST2
Stem Cell Factor	KITLG
Stromal cell-derived factor-1	CXCL12
Superoxide Dismutase 1, soluble	SOD1
T-Cell-Specific Protein RANTES	CCL5
T Lymphocyte-Secreted Protein I-309	CCL1
Tamm-Horsfall Urinary Glycoprotein	UMOD
Tenascin-C	TNC
Tenascin-X	TNXB
Tetranectin	CLEC3B
Thrombin-Activatable Fibrinolysis	CPB2
Thrombomodulin	THBD
Thrombospondin-1	THBS1
Thymus and activation-regulated chemokine	CCL17
Thymus-Expressed Chemokine	CCL25
Thyroglobulin	TG
Thyroid-Stimulating Hormone	TSHB
Thyroxine-Binding Globulin	SERPINA7
Tissue Inhibitor of Metalloproteinases 1	TIMP1
Tissue Inhibitor of Metalloproteinases 2	TIMP2
Tissue Inhibitor of Metalloproteinases 3	TIMP3
Tissue type Plasminogen activator	PLAT
TNF-Related Apoptosis-Inducing Ligand Receptor 3	TNFRSF10C
Transferrin receptor protein 1	TFRC
Transforming Growth Factor beta-3	TGFB3
Transthyretin	TTR
Trefoil Factor 3	TFF3
Tumor Necrosis Factor alpha	TNF
Tumor Necrosis Factor beta	LTA
Tumor necrosis factor ligand superfamily member 12	TNFSF12
Tumor necrosis factor ligand superfamily member 13	TNFSF13
Tumor Necrosis Factor Receptor I	TNFRSF1A
Tumor necrosis factor receptor 2	TNFRSF1B
Tyrosine kinase with Ig and EGF homology domains 2	TEK
Urokinase-type Plasminogen Activator	PLAU
Urokinase-type plasminogen activator receptor	PLAUR
Uteroglobin	SCGB1A1

Vascular Cell Adhesion Molecule-1	VCAM1
Vascular Endothelial Growth Factor	VEGFA
Vascular endothelial growth factor D	FIGF
Vascular Endothelial Growth Factor Receptor 1	FLT1
Vascular Endothelial Growth Factor Receptor 2	KDR
Visceral adipose tissue _ derived serpin A12	SERPINA12
Visfatin	NAMPT
Vitamin D-Binding Protein	GC
Vitamin K-Dependent Protein S	PROS1
Vitronectin	VTN
von Willebrand Factor	VWF
Bone morphogenetic protein 9	GDF2
YKL-40	CHI3L1

9.2 Appendix 2 – Permissions



From: **Kay Sharpe** Kay.Sharpe@ersj.org.uk
Subject: RE: Permission request for PhD thesis
Date: 17 November 2016 at 14:48
To: Peter M Hickey peter.hickey@sheffield.ac.uk

KS

ASPIRE registry: Assessing the Spectrum of Pulmonary hypertension Identified at a REferral centre
J. Hurdman, R. Condliffe, C.A. Elliot, C. Davies, C. Hill, J.M. Wild, D. Capener, P. Sephton, N. Hamilton, I.J. Armstrong, C. Billings, A. Lawrie, I. Sabroe, M. Akil, L. O'Toole, D.G. Kiely
European Respiratory Journal Apr 2012, 39 (4) 945-955; DOI: 10.1183/09031936.00078411

Material: Figure 2a and 3

Acknowledgement Wording: Reproduced with permission of the European Respiratory Society ©. *European Respiratory Journal Apr 2012, 39 (4) 945-955; DOI: 10.1183/09031936.00078411*

Permission is granted for the material stated above to be reproduced for your thesis in accordance with ERS copyright policy – see below

Notes, Terms & Conditions (where applicable)

"Green" Open Access and Author Archiving: Authors who do not wish to pay for the ERJ Open option will still have their manuscripts made free to access via the ERJ online archive following the journal's 18-month embargo period; after this embargo period, authors also have licence to deposit their manuscripts in an institutional (or other) repository for public archiving, provided the following requirements are met:

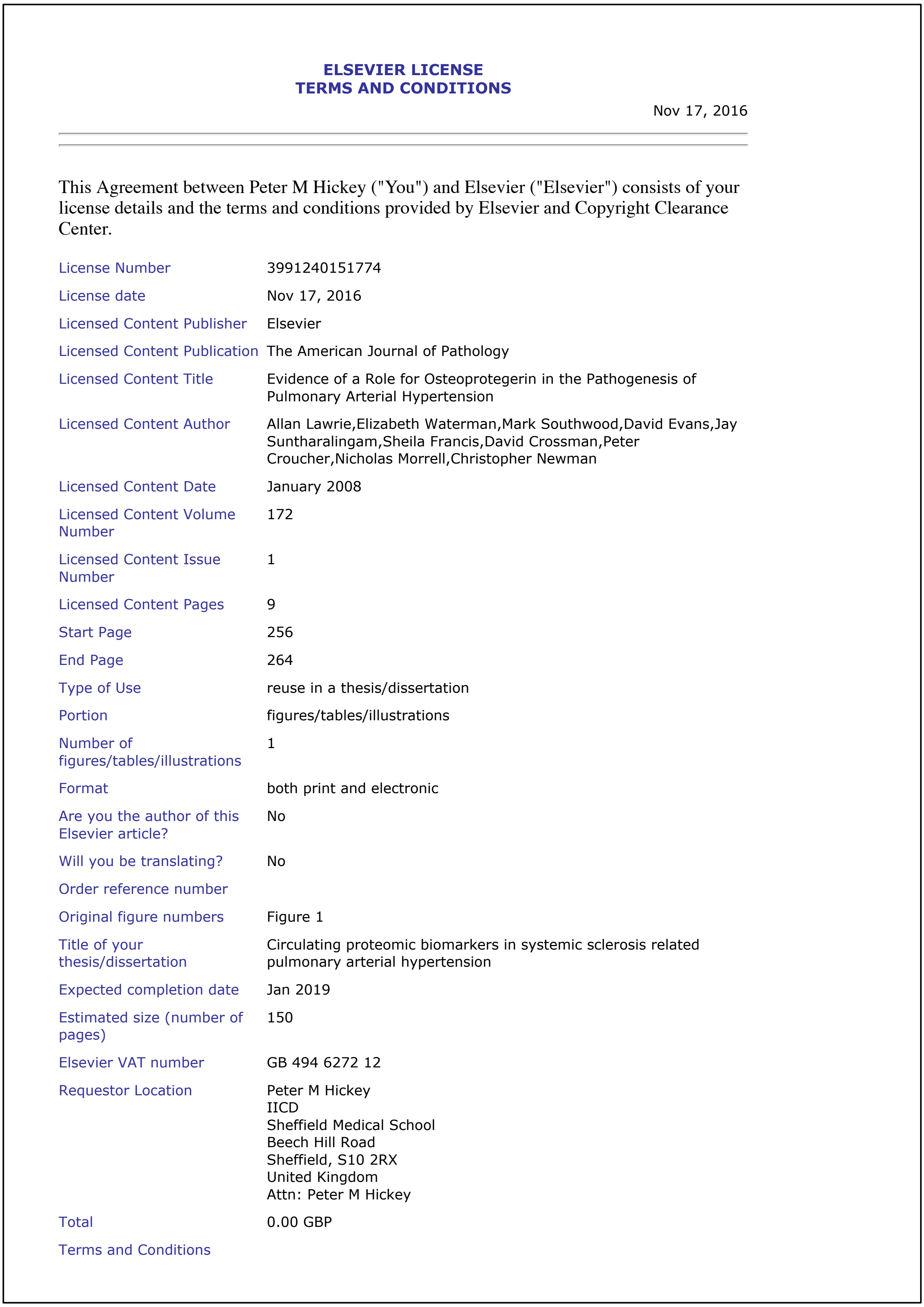
1) The final, peer-reviewed, author-submitted version that was accepted for publication is used (before copy-editing and publication).

2) A permanent link is provided to the version of the article published in the ERJ, through the dx.doi.org platform. For example, if your manuscript has the DOI 10.1183/09031936.00123412, then the link you provide must be dx.doi.org/10.1183/09031936.00123412

3) The repository on which the manuscript is deposited is not used for systematic distribution or commercial sales purposes.

4) The following required archiving statement appears on the title page of the archived manuscript: "This is an author-submitted, peer-reviewed version of a manuscript that has been accepted for publication in the European Respiratory Journal, prior to copy-editing, formatting and typesetting. This version of the manuscript may not be duplicated or reproduced without prior permission from the copyright owner, the European Respiratory Society. The publisher is not responsible or liable for any errors or omissions in this version of the manuscript or in any version derived from it by any other parties. The final, copy-edited, published article, which is the version of record, is available without a subscription 18 months after the date of issue publication."

Authors of articles published under one of the Creative Commons licences (this includes all articles in ERJ Open Research, European Respiratory Review and Breathe) retain further rights to share, reuse or adapt their manuscript. The extent of these rights depends on the specific Creative Commons licence used. Please consult the relevant section of the online instructions for authors. These publications are copyrighted material and must not be copied, reproduced,

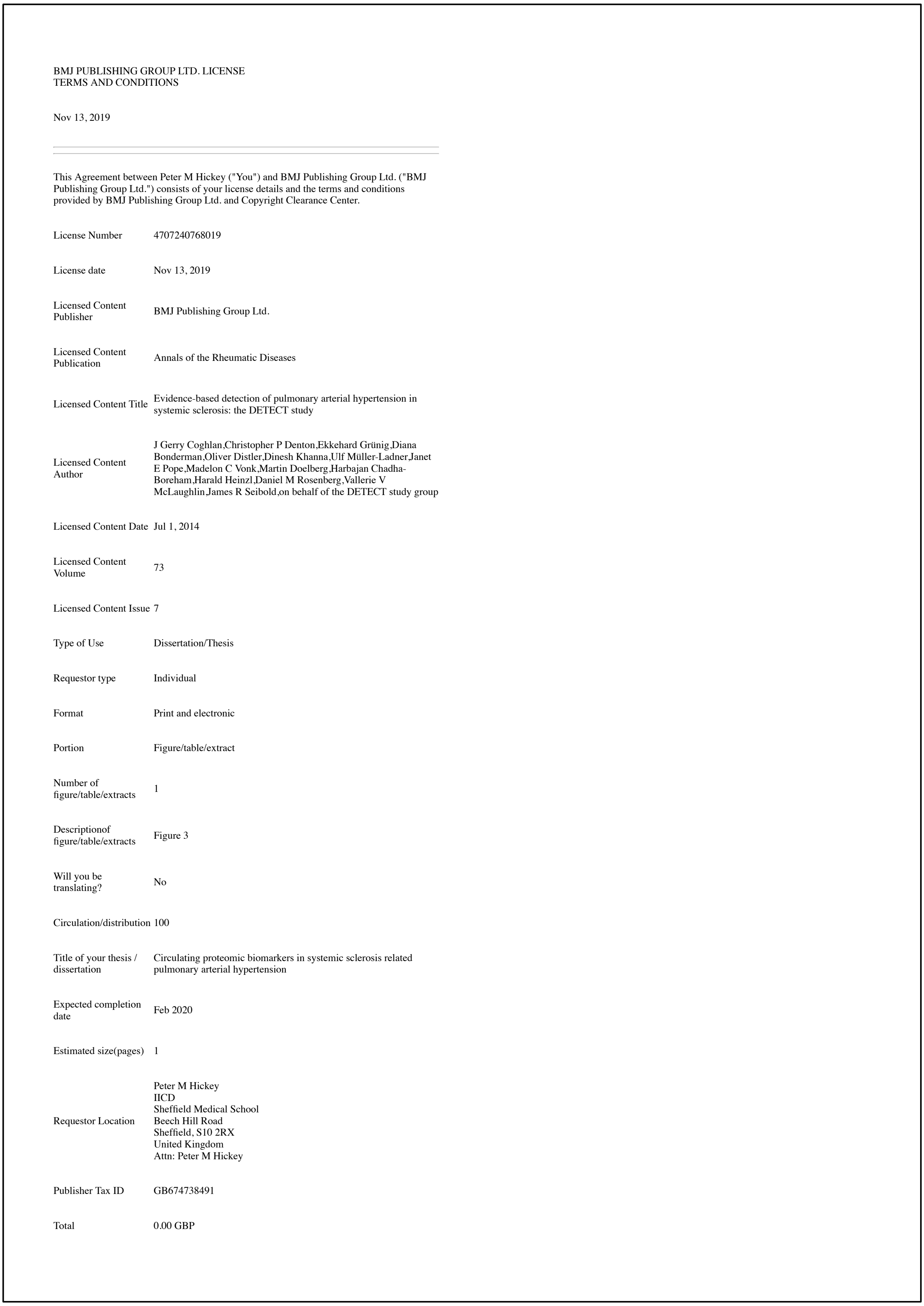


ELSEVIER LICENSE
TERMS AND CONDITIONS

Nov 17, 2016

This Agreement between Peter M Hickey ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	3991240151774
License date	Nov 17, 2016
Licensed Content Publisher	Elsevier
Licensed Content Publication	The American Journal of Pathology
Licensed Content Title	Evidence of a Role for Osteoprotegerin in the Pathogenesis of Pulmonary Arterial Hypertension
Licensed Content Author	Allan Lawrie,Elizabeth Waterman,Mark Southwood,David Evans,Jay Suntharalingam,Sheila Francis,David Crossman,Peter Croucher,Nicholas Morrell,Christopher Newman
Licensed Content Date	January 2008
Licensed Content Volume Number	172
Licensed Content Issue Number	1
Licensed Content Pages	9
Start Page	256
End Page	264
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Original figure numbers	Figure 1
Title of your thesis/dissertation	Circulating proteomic biomarkers in systemic sclerosis related pulmonary arterial hypertension
Expected completion date	Jan 2019
Estimated size (number of pages)	150
Elsevier VAT number	GB 494 6272 12
Requestor Location	Peter M Hickey IICD Sheffield Medical School Beech Hill Road Sheffield, S10 2RX United Kingdom Attn: Peter M Hickey
Total	0.00 GBP
Terms and Conditions	

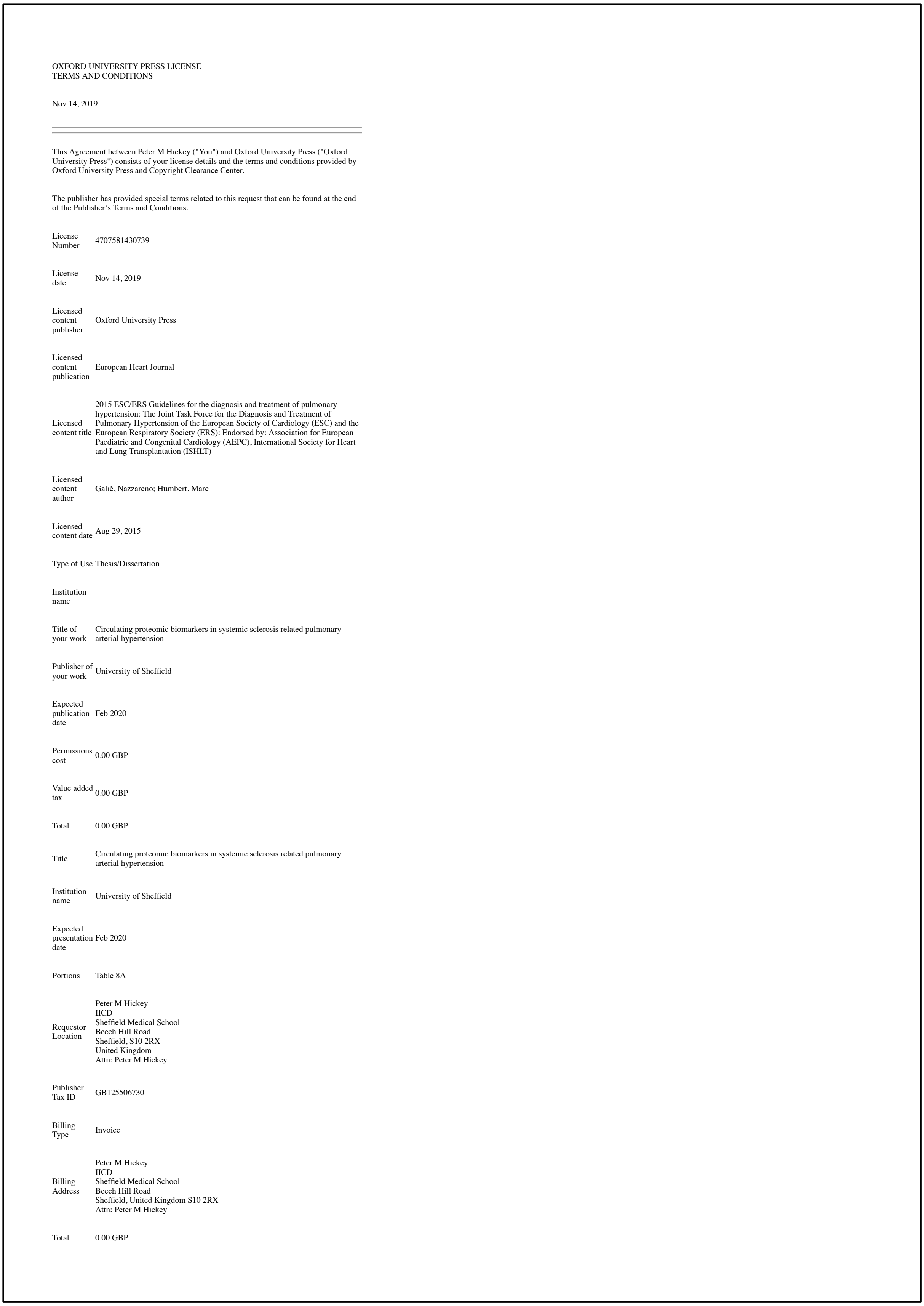


BMJ PUBLISHING GROUP LTD. LICENSE
TERMS AND CONDITIONS

Nov 13, 2019

This Agreement between Peter M Hickey ("You") and BMJ Publishing Group Ltd. ("BMJ Publishing Group Ltd.") consists of your license details and the terms and conditions provided by BMJ Publishing Group Ltd. and Copyright Clearance Center.

License Number	4707240768019
License date	Nov 13, 2019
Licensed Content Publisher	BMJ Publishing Group Ltd.
Licensed Content Publication	Annals of the Rheumatic Diseases
Licensed Content Title	Evidence-based detection of pulmonary arterial hypertension in systemic sclerosis: the DETECT study
Licensed Content Author	J Gerry Coghlan,Christopher P Denton,Ekkehard Grünig,Diana Bonderman,Oliver Distler,Dinesh Khanna,Ulf Müller-Ladner,Janet E Pope,Madelon C Vonk,Martin Doelberg,Harbajan Chadha-Boreham,Harald Heinzl,Daniel M Rosenberg,Vallerie V McLaughlin,James R Seibold,on behalf of the DETECT study group
Licensed Content Date	Jul 1, 2014
Licensed Content Volume	73
Licensed Content Issue	7
Type of Use	Dissertation/Thesis
Requestor type	Individual
Format	Print and electronic
Portion	Figure/table/extract
Number of figure/table/extracts	1
Descriptionof figure/table/extracts	Figure 3
Will you be translating?	No
Circulation/distribution	100
Title of your thesis / dissertation	Circulating proteomic biomarkers in systemic sclerosis related pulmonary arterial hypertension
Expected completion date	Feb 2020
Estimated size(pages)	1
Requestor Location	Peter M Hickey IICD Sheffield Medical School Beech Hill Road Sheffield, S10 2RX United Kingdom Attn: Peter M Hickey
Publisher Tax ID	GB674738491
Total	0.00 GBP



OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Nov 14, 2019

This Agreement between Peter M Hickey ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

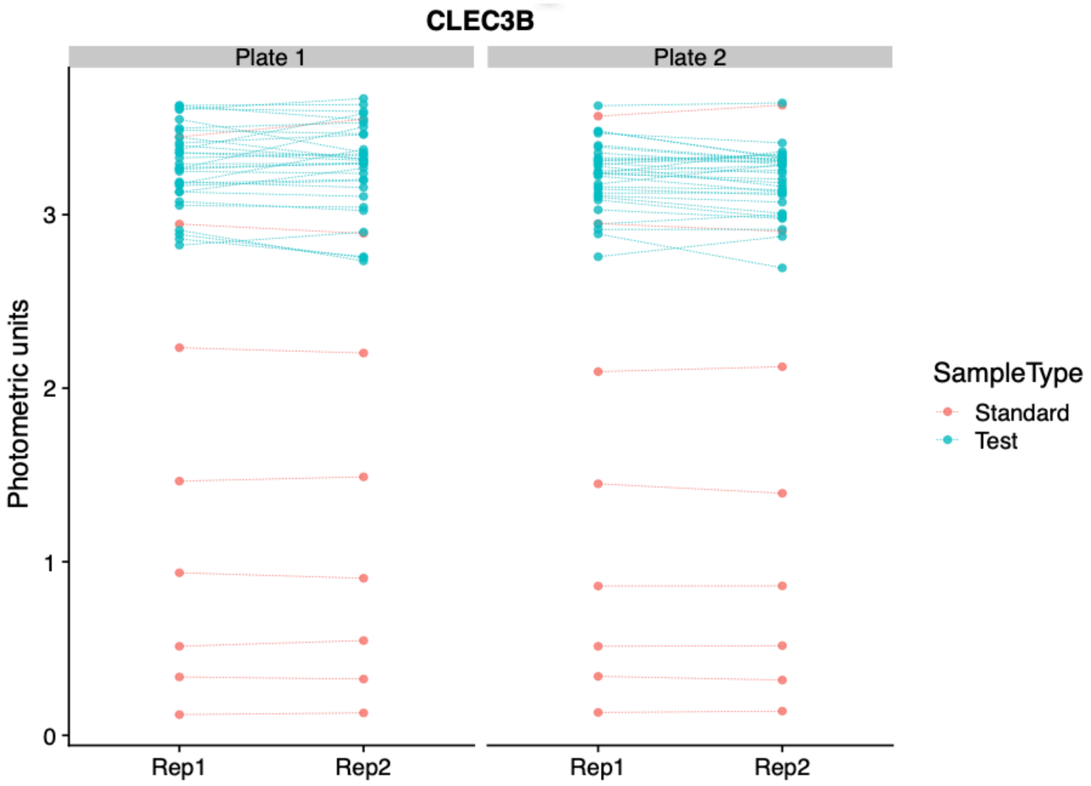
The publisher has provided special terms related to this request that can be found at the end of the Publisher's Terms and Conditions.

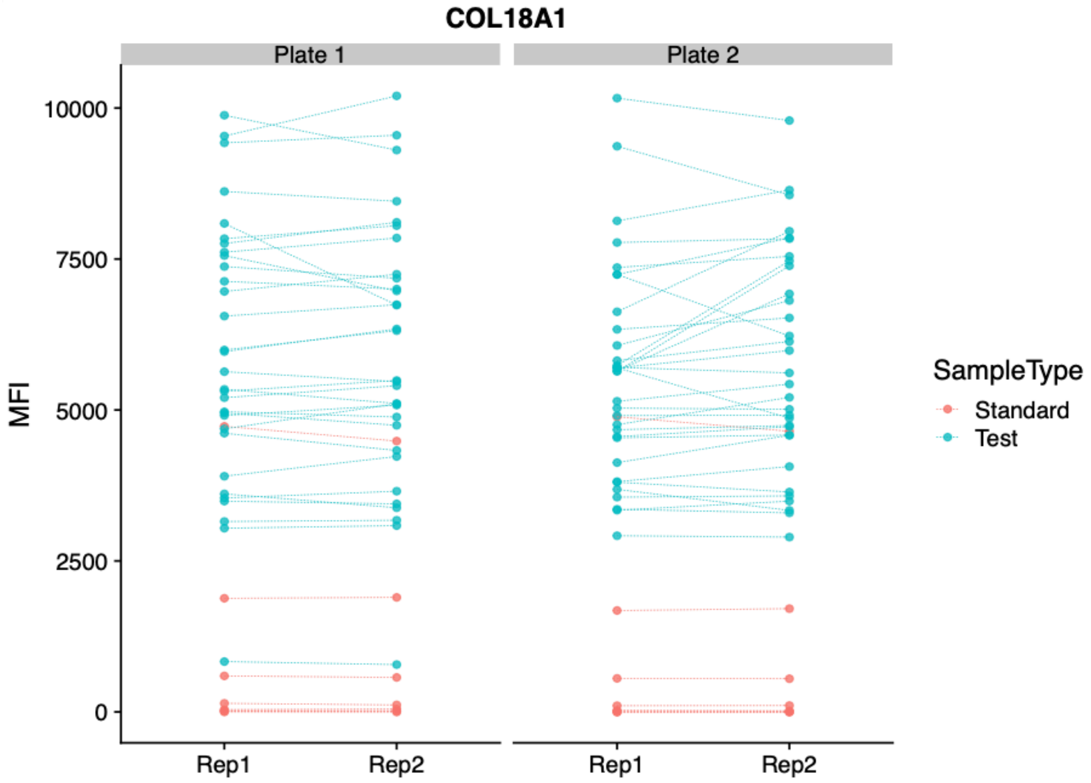
License Number	4707581430739
License date	Nov 14, 2019
Licensed content publisher	Oxford University Press
Licensed content publication	European Heart Journal
Licensed content title	2015 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension: The Joint Task Force for the Diagnosis and Treatment of Pulmonary Hypertension of the European Society of Cardiology (ESC) and the European Respiratory Society (ERS): Endorsed by: Association for European Paediatric and Congenital Cardiology (AEPC), International Society for Heart and Lung Transplantation (ISHLT)
Licensed content author	Galiè, Nazzareno; Humbert, Marc
Licensed content date	Aug 29, 2015
Type of Use	Thesis/Dissertation
Institution name	
Title of your work	Circulating proteomic biomarkers in systemic sclerosis related pulmonary arterial hypertension
Publisher of your work	University of Sheffield
Expected publication date	Feb 2020
Permissions cost	0.00 GBP
Value added tax	0.00 GBP
Total	0.00 GBP
Title	Circulating proteomic biomarkers in systemic sclerosis related pulmonary arterial hypertension
Institution name	University of Sheffield
Expected presentation date	Feb 2020
Portions	Table 8A
Requestor Location	Peter M Hickey IICD Sheffield Medical School Beech Hill Road Sheffield, S10 2RX United Kingdom Attn: Peter M Hickey
Publisher Tax ID	GB125506730
Billing Type	Invoice
Billing Address	Peter M Hickey IICD Sheffield Medical School Beech Hill Road Sheffield, United Kingdom S10 2RX Attn: Peter M Hickey
Total	0.00 GBP

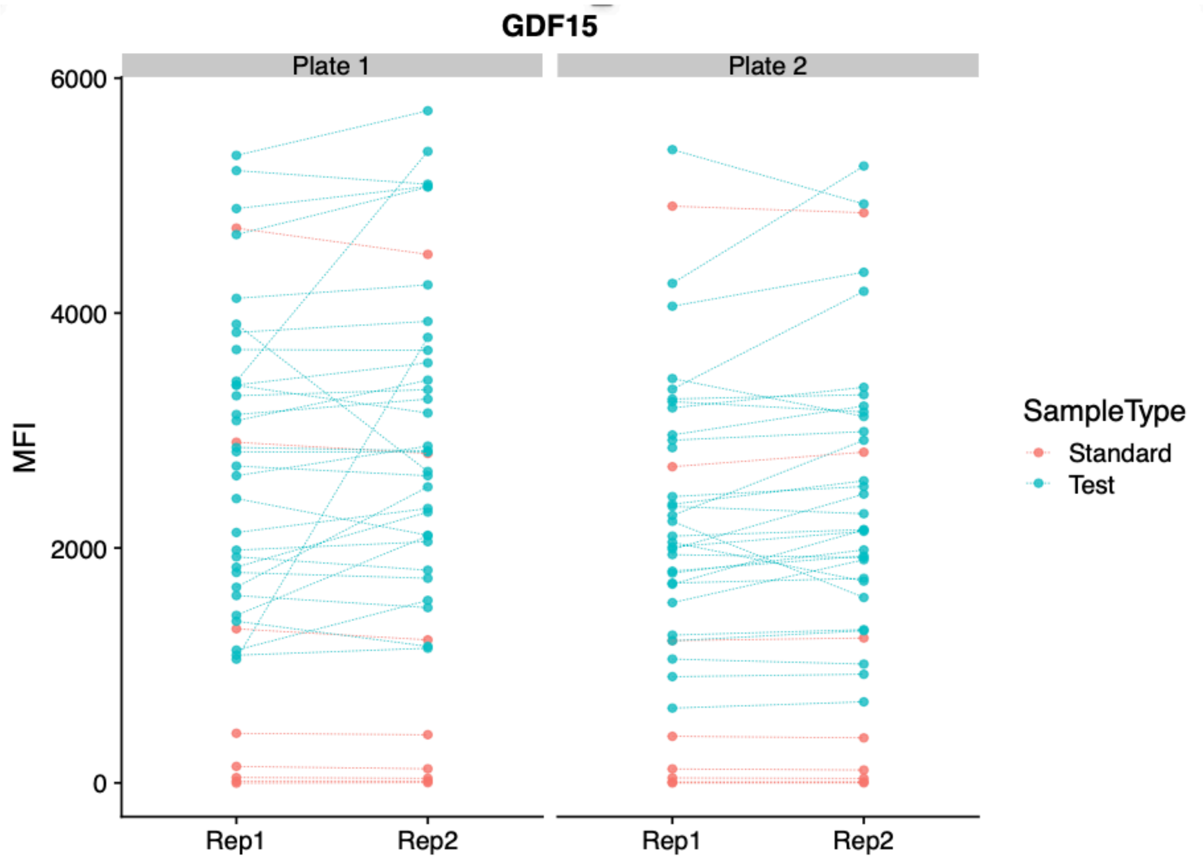
9.3 Appendix 3 – External validation QC

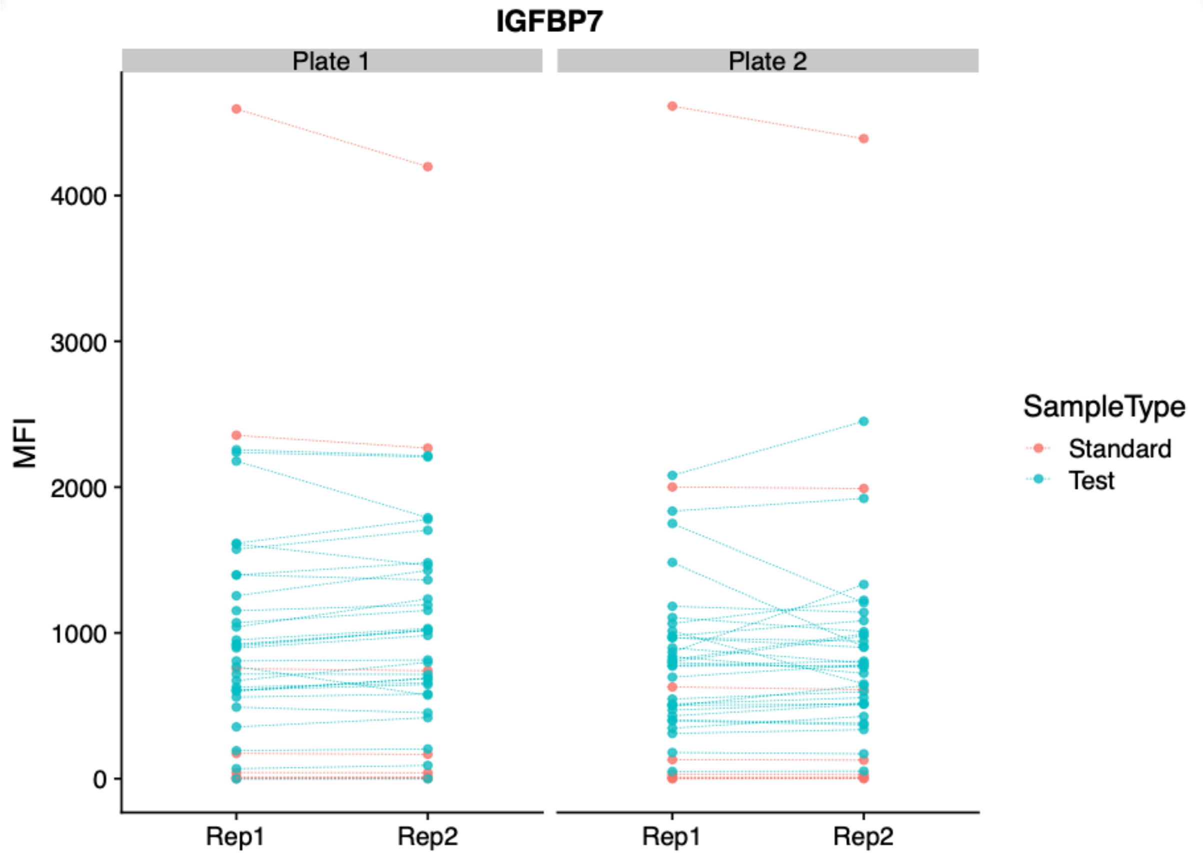
9.3.1 Sample replicate plots

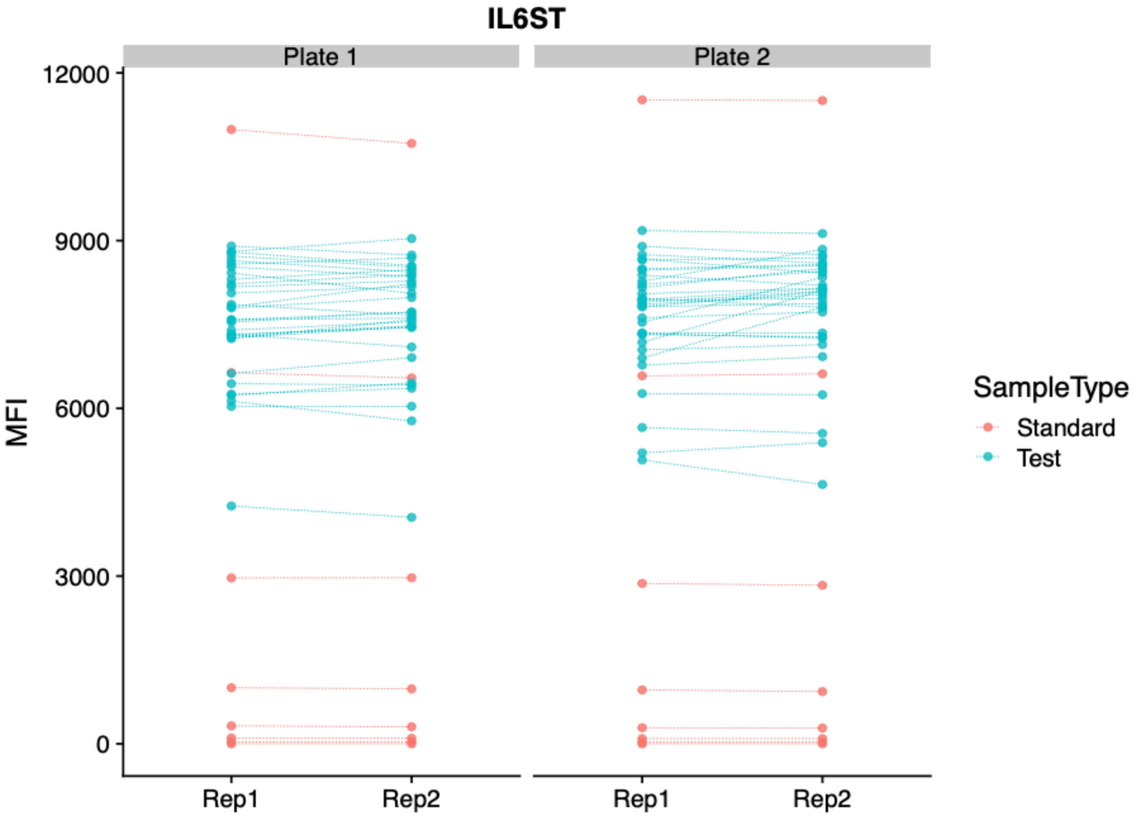
Stanford University sample replicate plots:

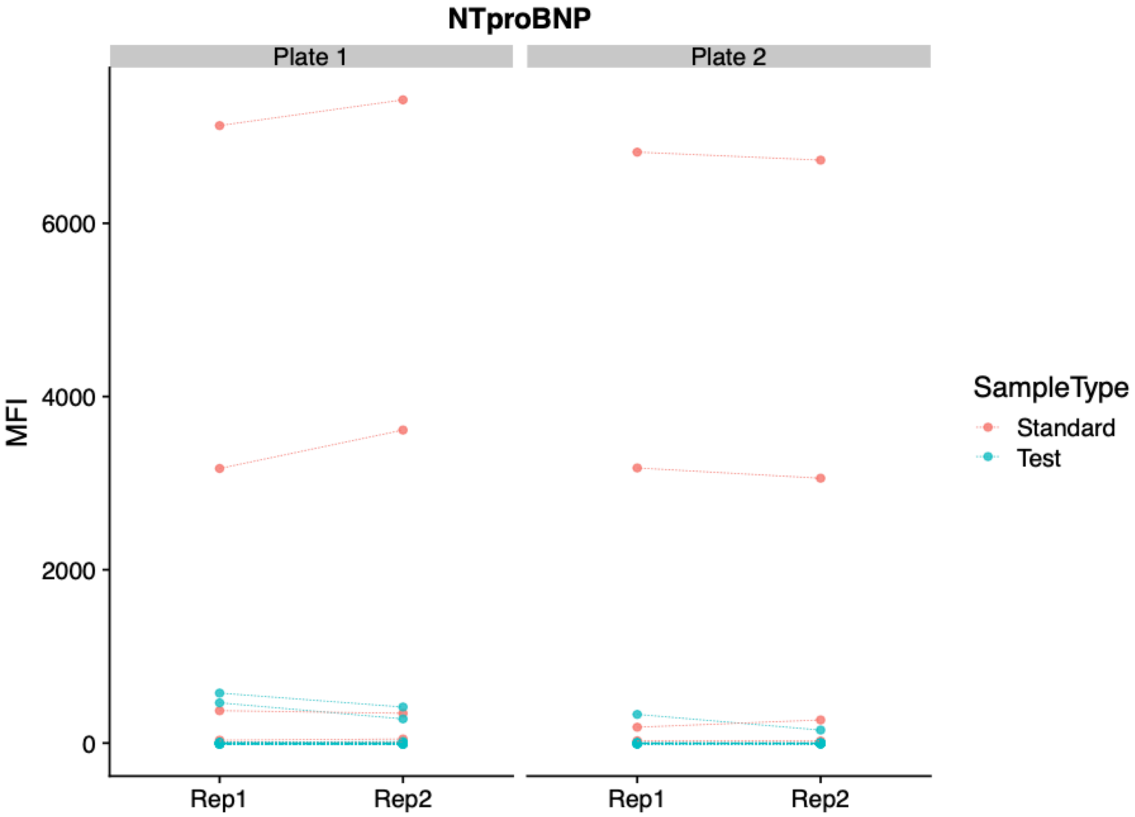


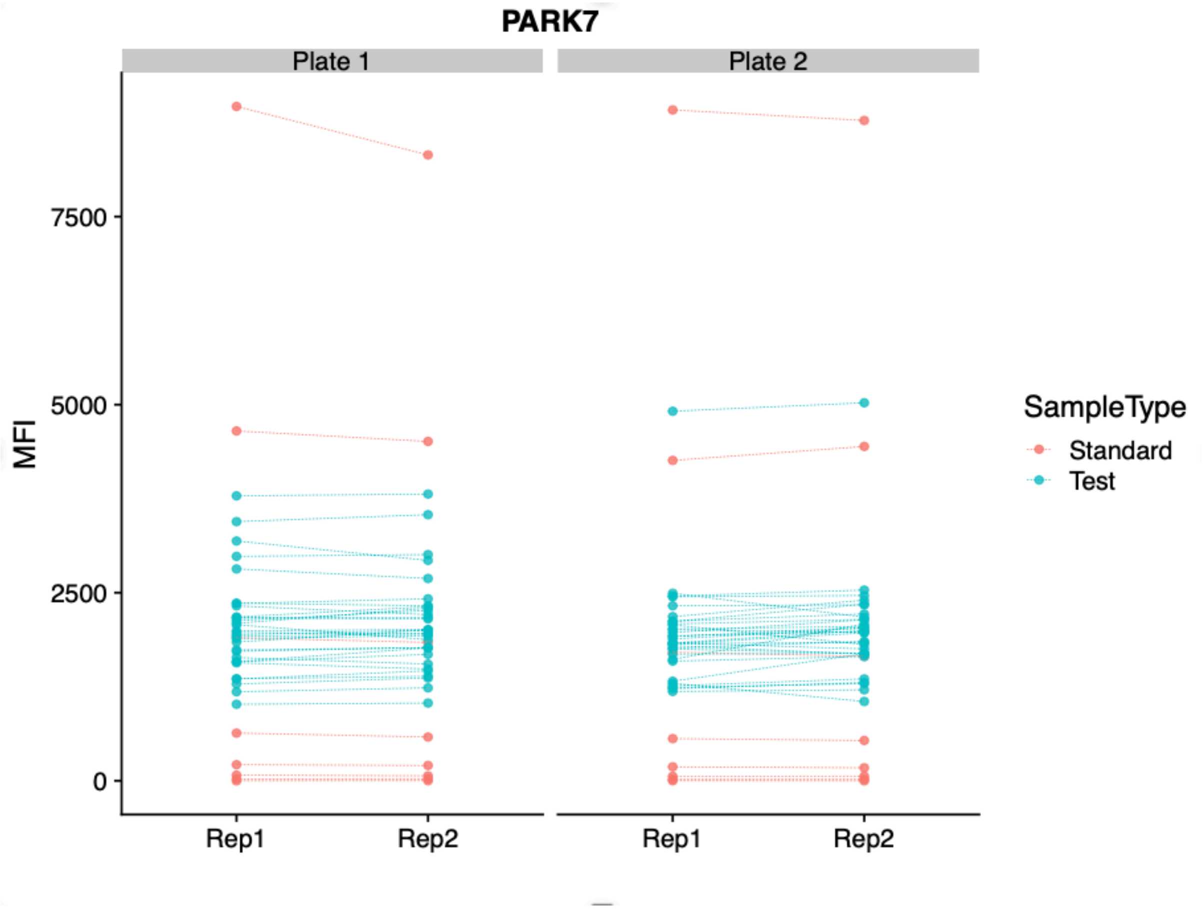



GDF15
Plate 1
Plate 2
6000
4000
2000
0
MFI
Rep1
Rep2
Rep1
Rep2
SampleType
Standard
Test


IGFBP7
Plate 1
Plate 2
4000
3000
2000
1000
0
MFI
Rep1
Rep2
Rep1
Rep2
SampleType
Standard
Test

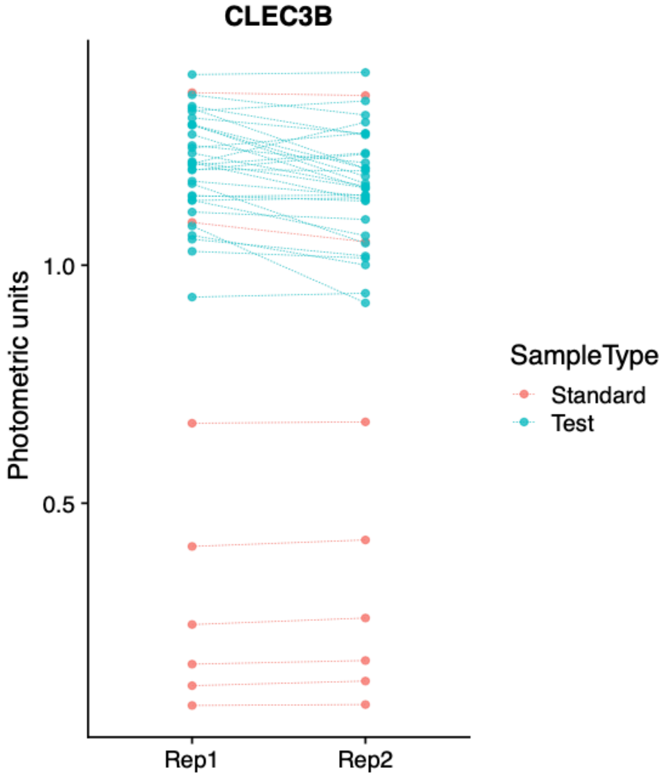

IL6ST
Plate 1
Plate 2
MFI
12000
9000
6000
3000
0
Rep1
Rep2
Rep1
Rep2
SampleType
Standard
Test

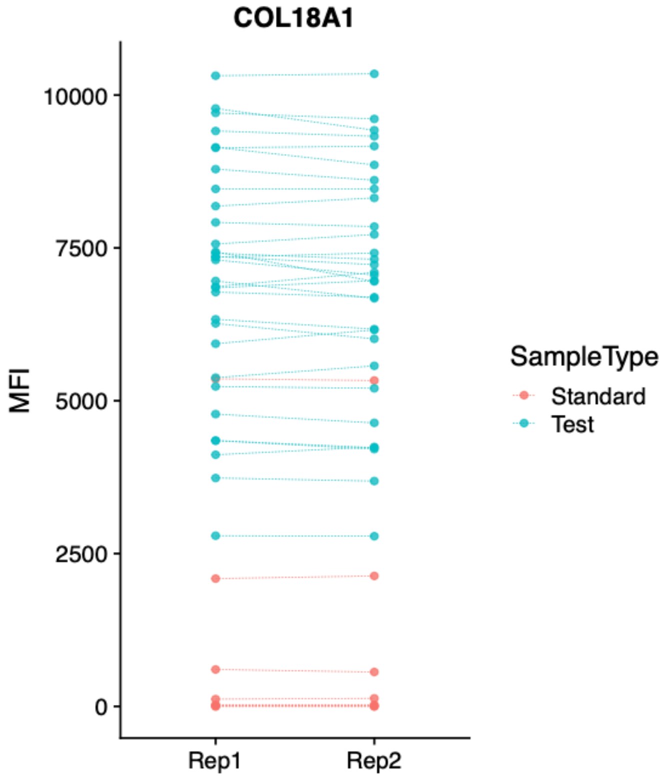

NTproBNP
Plate 1
Plate 2
MFI
6000
4000
2000
0
Rep1
Rep2
Rep1
Rep2
SampleType
Standard
Test

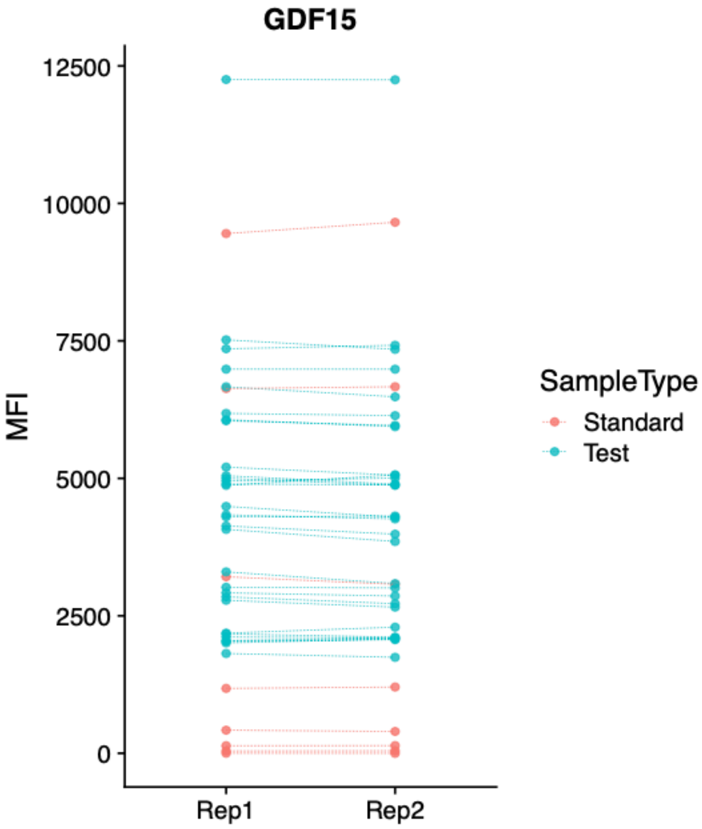


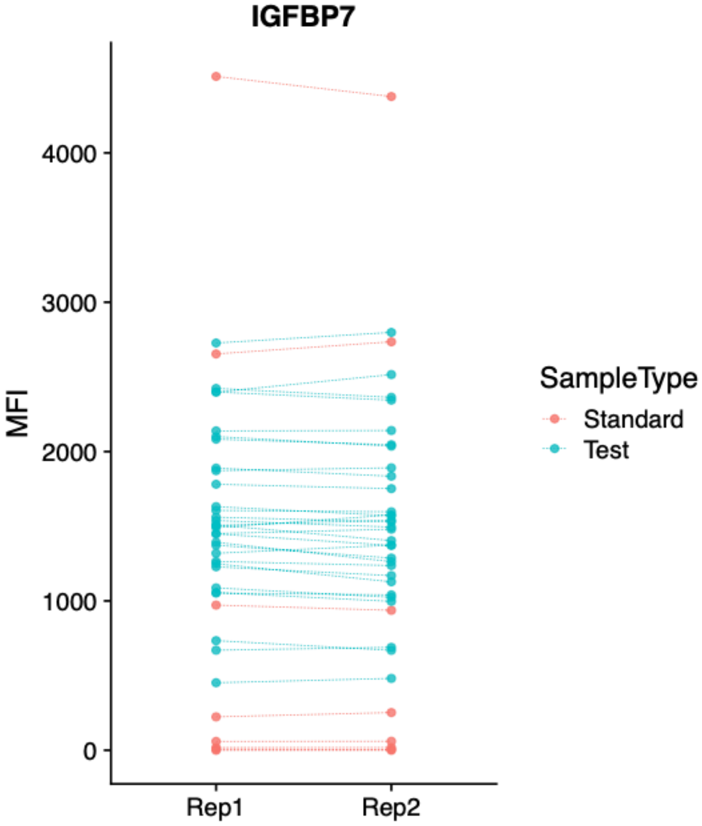
Dot plots showing raw MFI and photometric unit readings for replicates on each plate with patient samples linked by dotted lines. CLEC3B data from ELISA assay, other from Luminex assays. Stanford University patient samples.

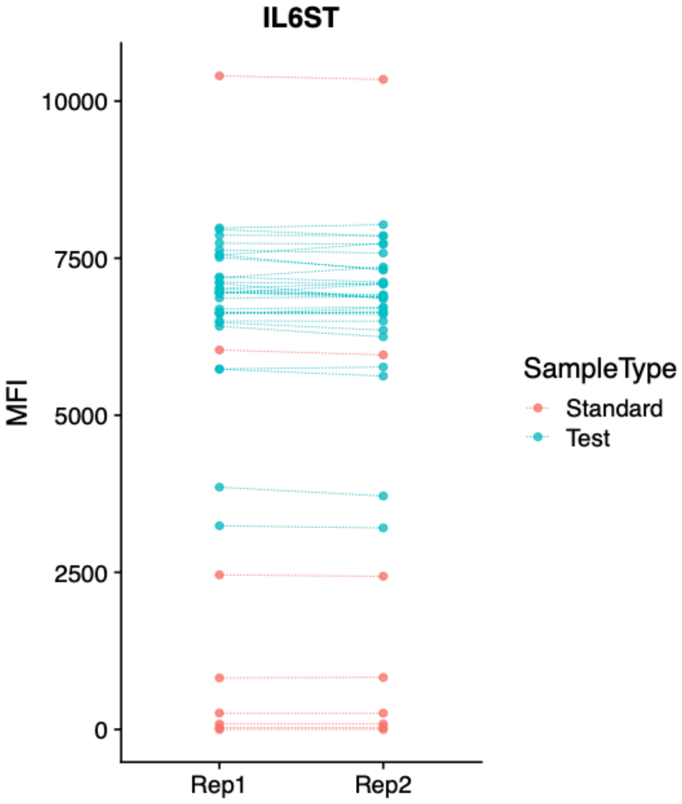
Vanderbilt University sample replicate plots:

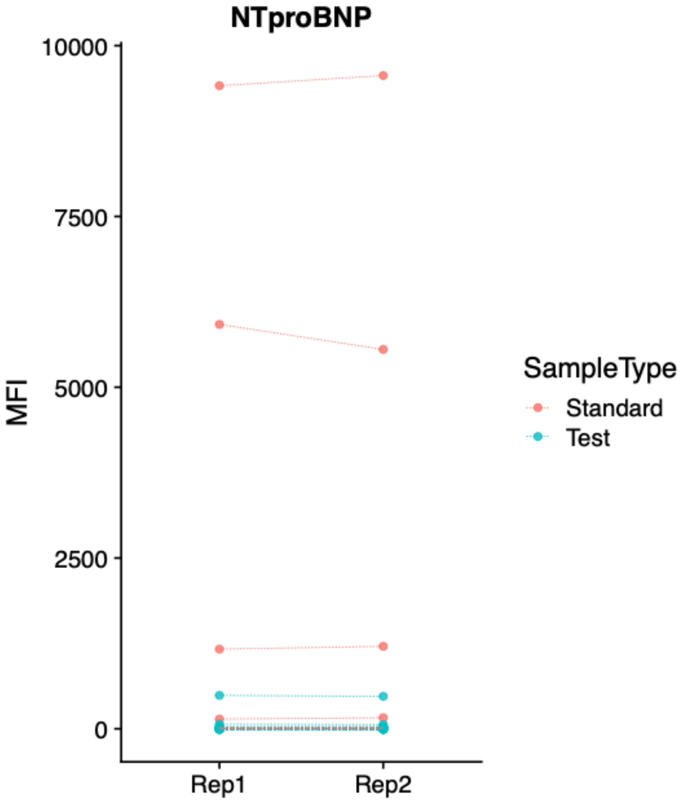


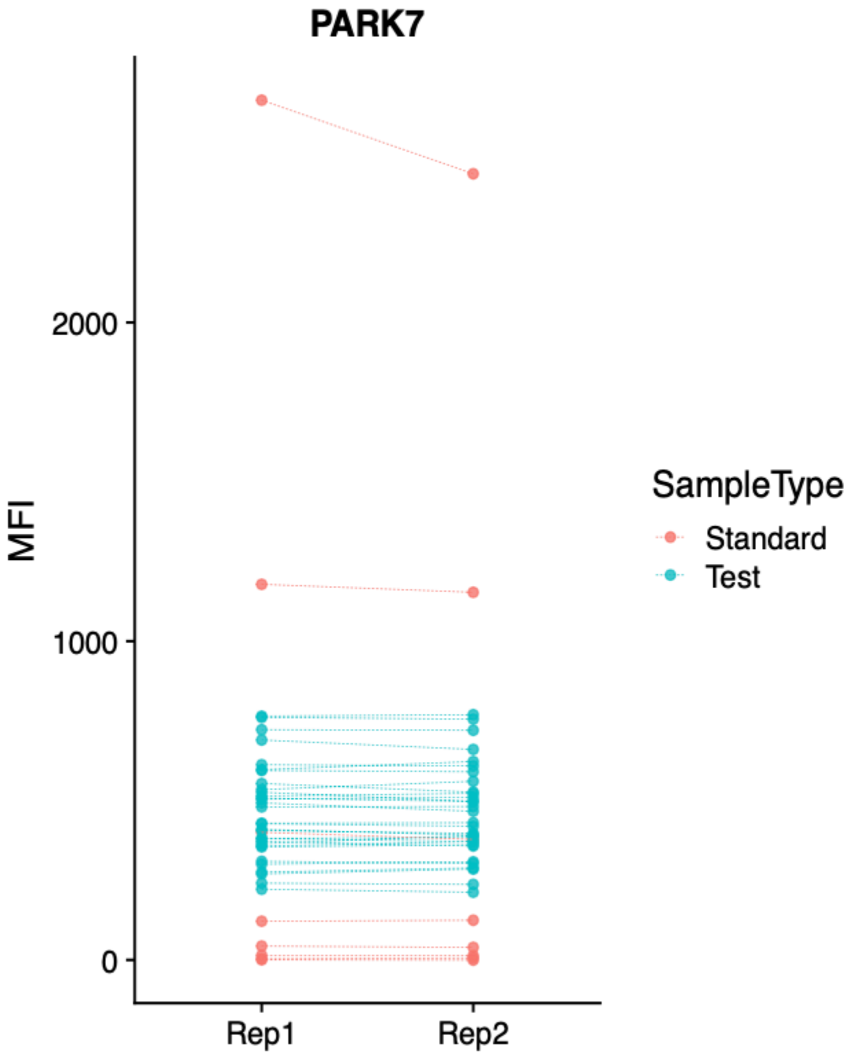



GDF15
12500
10000
7500
5000
2500
0
MFI
Rep1
Rep2
SampleType
Standard
Test


IGFBP7
4000
3000
2000
1000
0
MFI
Rep1
Rep2
SampleType
Standard
Test

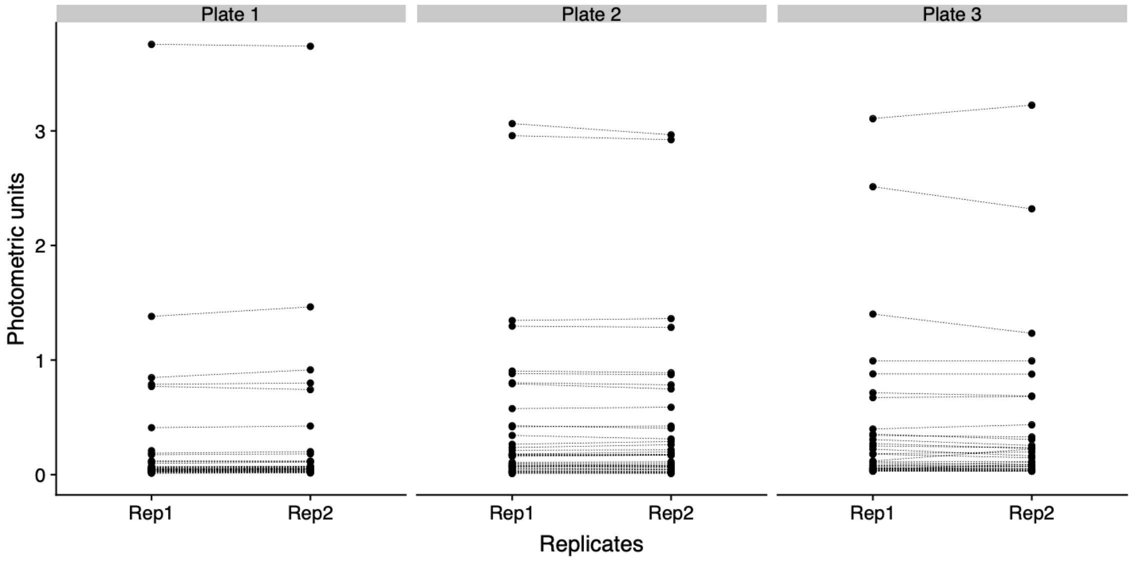

IL6ST
MFI
10000
7500
5000
2500
0
Rep1
Rep2
SampleType
Standard
Test


NTproBNP
MFI
10000
7500
5000
2500
0
Rep1
Rep2
SampleType
Standard
Test



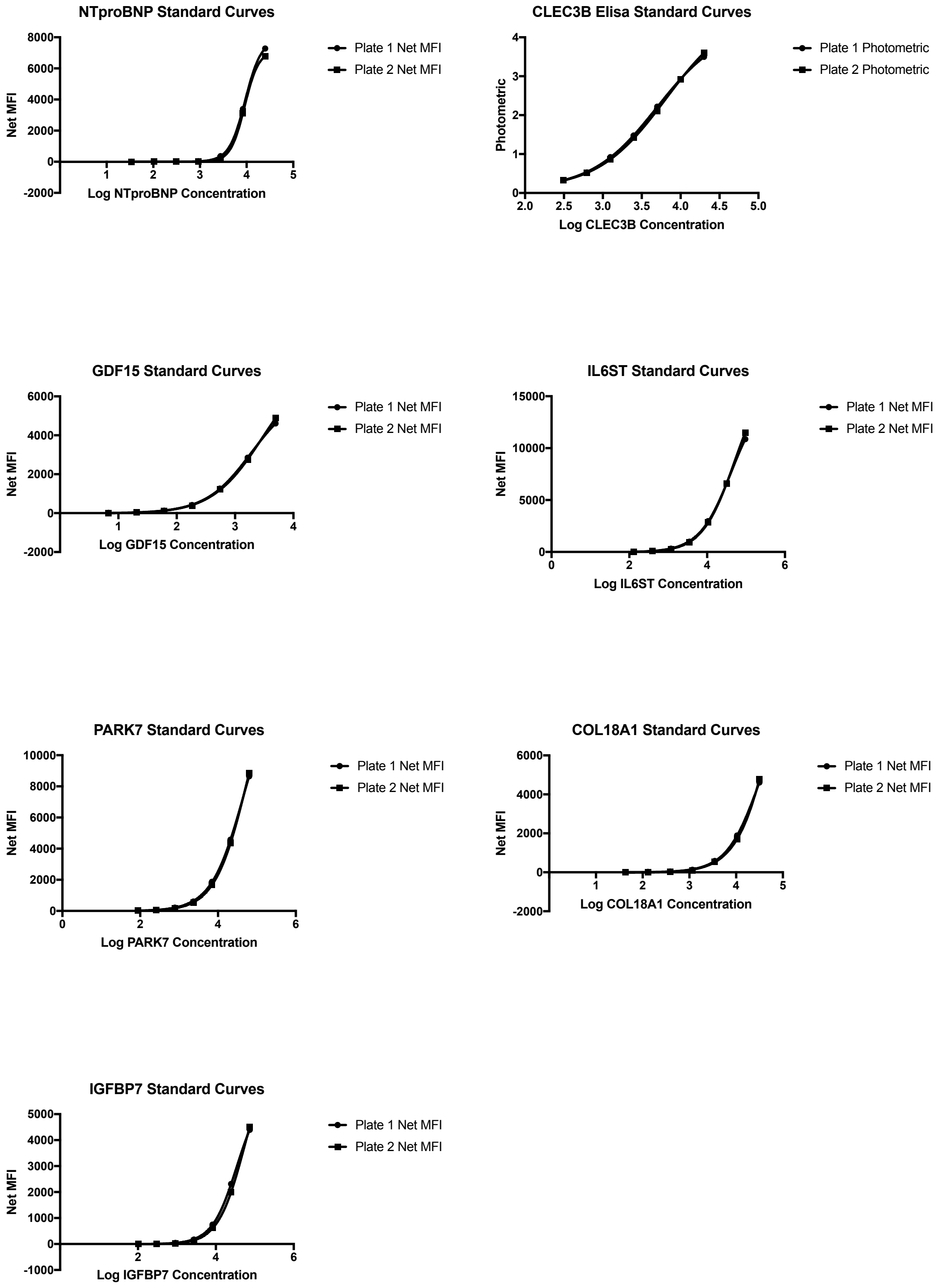
Dot plots showing raw MFI and photometric unit readings for replicates on each plate with patient samples linked by dotted lines. CLEC3B data from ELISA assay, other from Luminex assays. Vanderbilt University patient samples.

Repeat NTproBNP ELISA assay replicate plots:



Dot plots showing raw photometric unit readings for replicates on each plate with patient samples linked by dotted lines. Data from ELISA assay Plate 1 and 2 carried Stanford University samples, Plate 3 carried Vanderbilt University samples.

9.3.2 Stanford assay plate standard curves



NtproBNP Standard Curves

CLEC3B Elisa Standard Curves

GDF15 Standard Curves

IL6ST Standard Curves

PARK7 Standard Curves

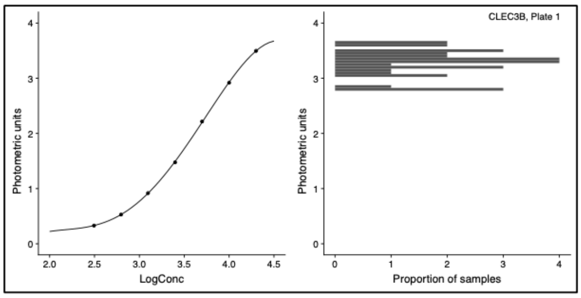
COL18A1 Standard Curves

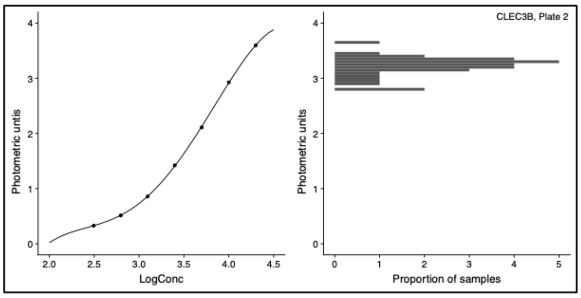
IGFBP7 Standard Curves

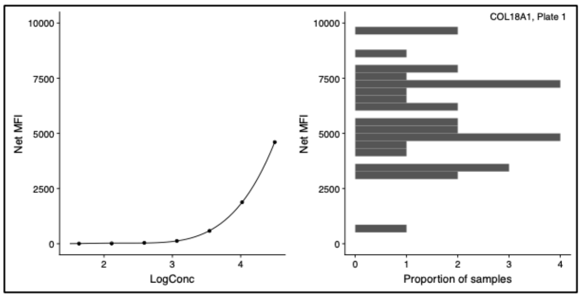
Plots showing standard curves from each plate in the Stanford sample analysis for each assay.

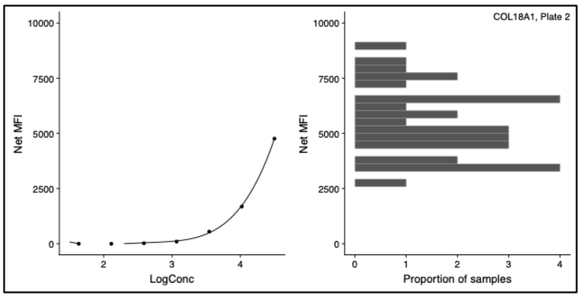
9.3.3 Raw data against standard curves

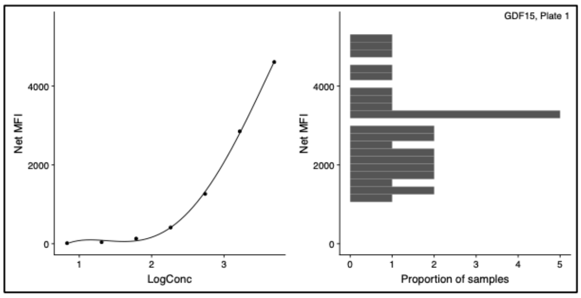
Stanford assays

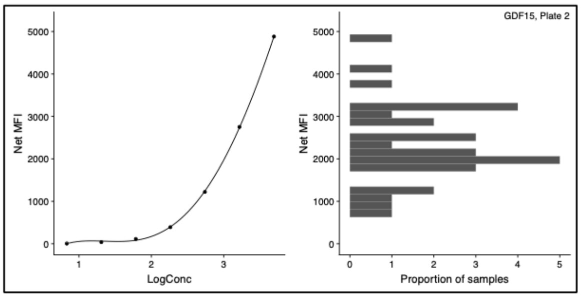


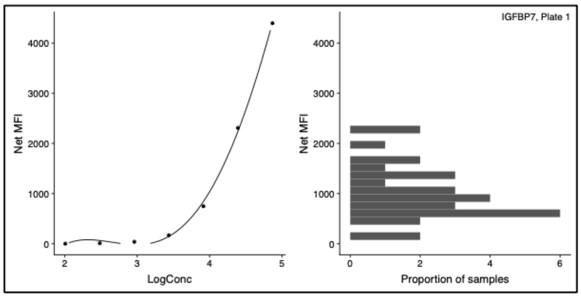


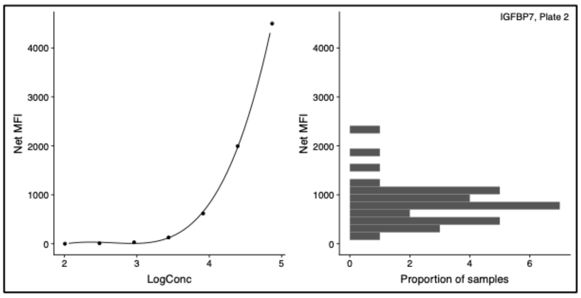


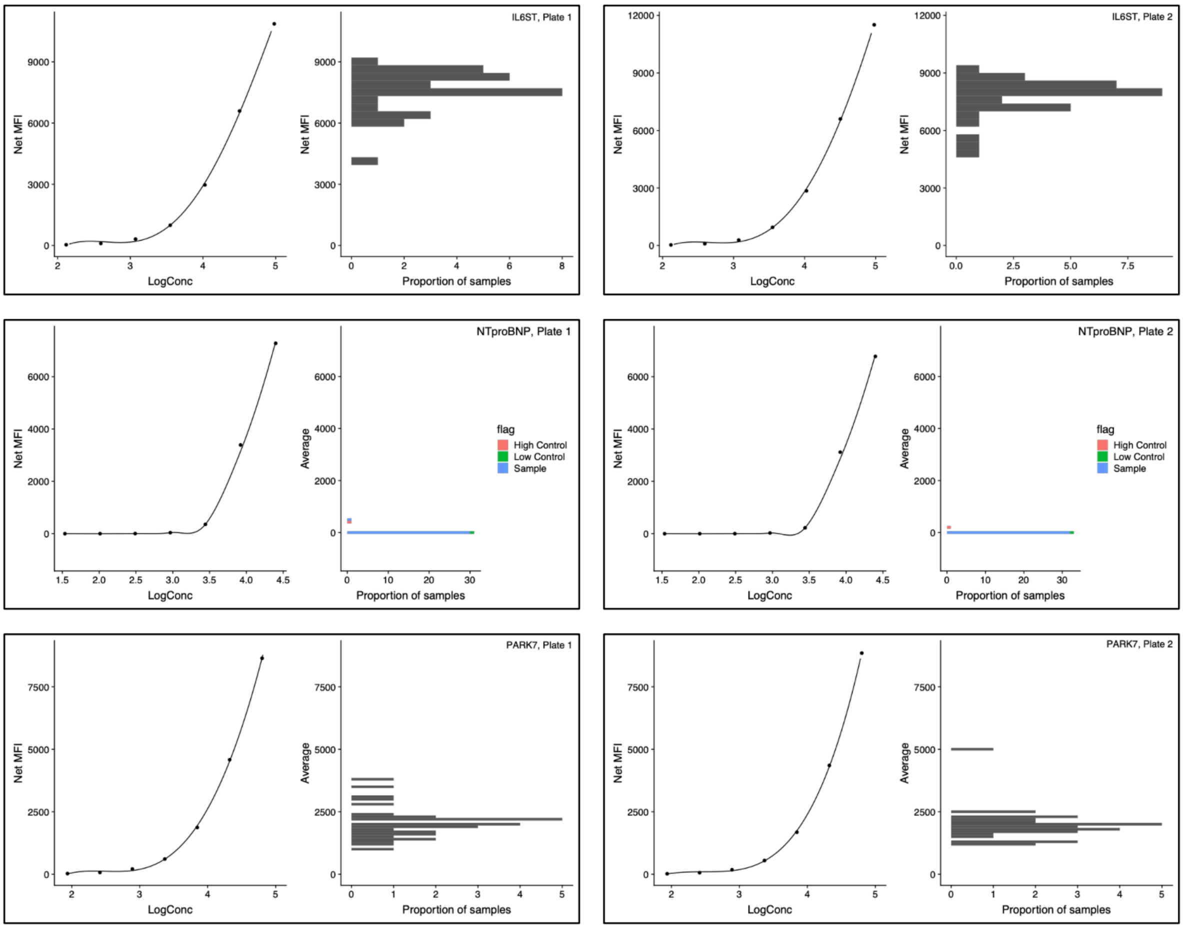






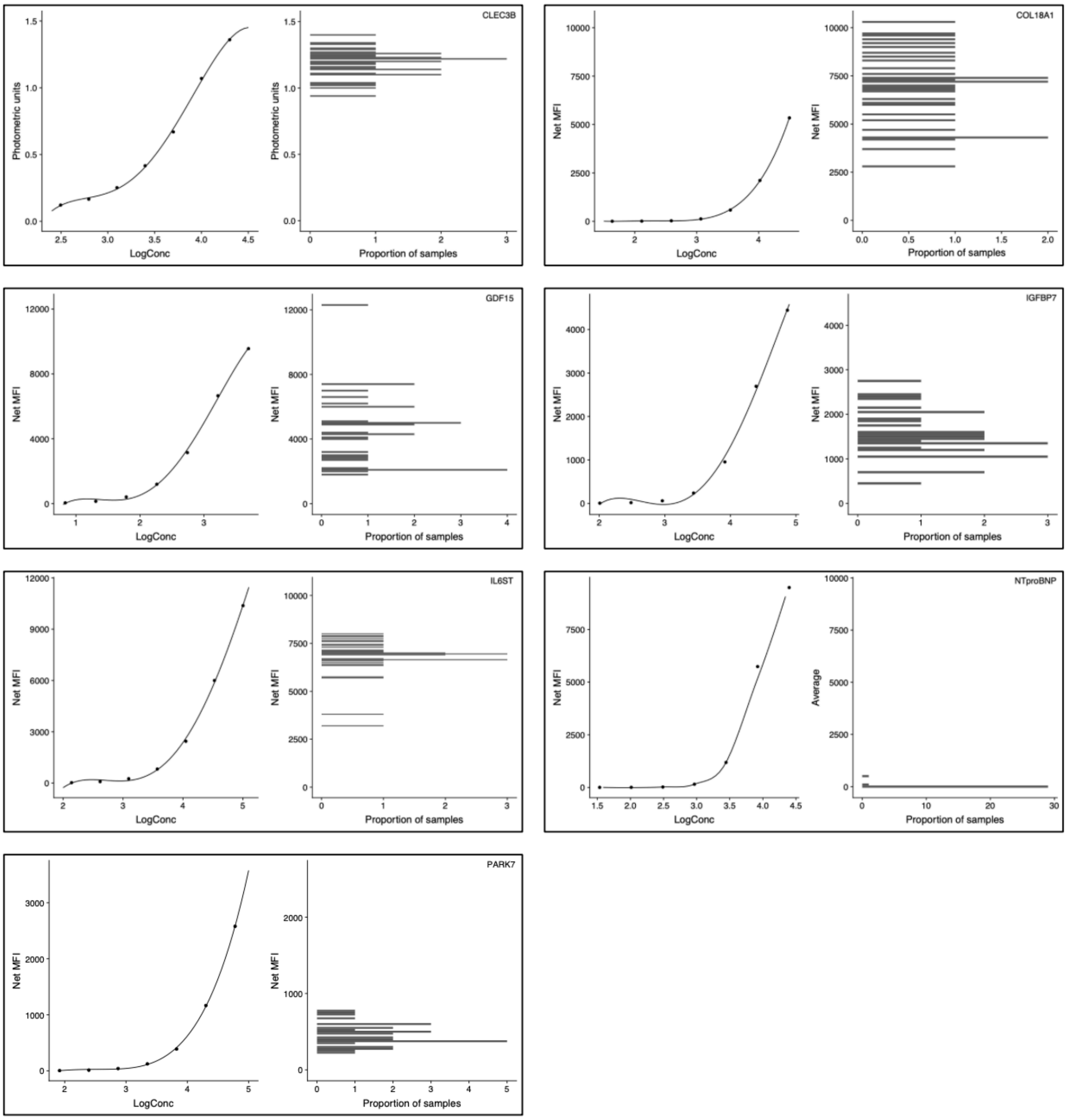






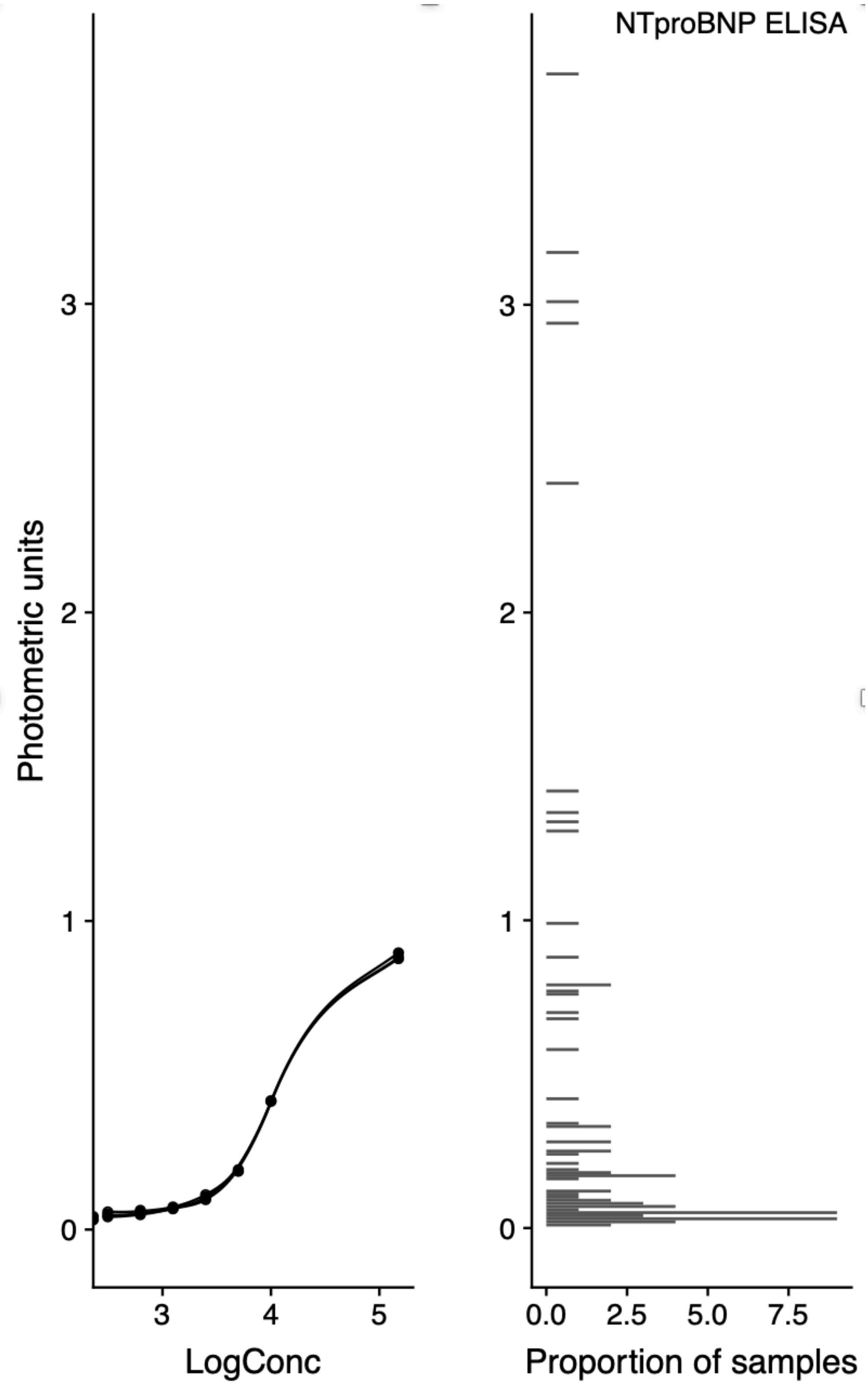
Raw data detection values plotted against corresponding standard curve. Stanford assays, each plate displayed separately.

Vanderbilt assays



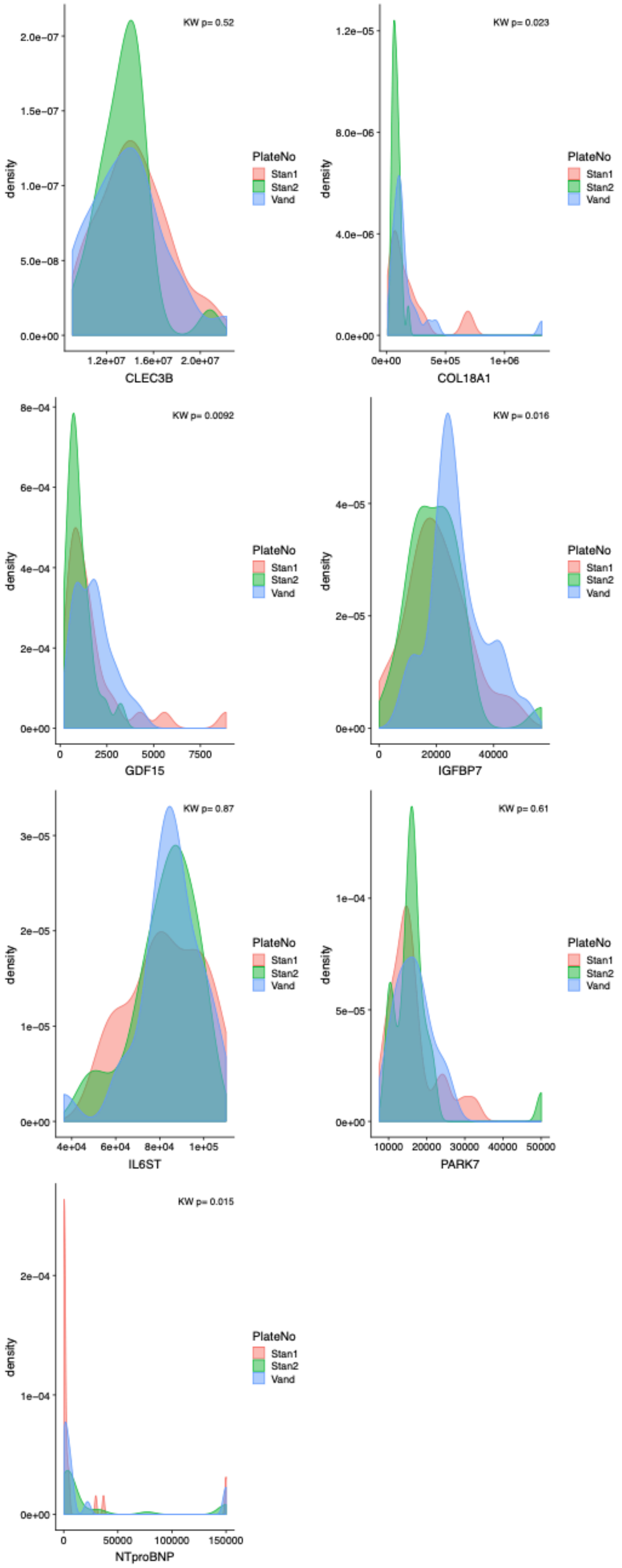
Raw data detection values plotted against corresponding standard curve. Vanderbilt assays.

Repeat NTproBNP ELISA:



Raw data detection values plotted against corresponding standard curve. Repeat NTproBNP ELISA assays. Standard curve graph also demonstrating repeatability of the standard curve measurement with all three standard curves overlaid on this single graph.

9.3.4 Protein concentration plate distribution plots



9.4 Appendix 4 – Publications

9.4.1 Publications arising from this work

Hickey PM, Iremonger J, Pickworth J, Del Rosario P, Hsi A, Casbolt H, Arnold N, Thompson AAR, Hemnes A, Zamanian R, Condliffe R, Lawrie A. Circulating proteomic biomarkers to screen for pulmonary arterial hypertension in systemic sclerosis. *Poster presentation at PVRI World Congress*. Barcelona. 2019. (Winner: Best Abstract).

Hickey PM, Lawrie A, Condliffe R. Circulating protein biomarkers in systemic sclerosis related pulmonary arterial hypertension: a review of published data. *Frontiers in Medicine.* 5:175. 2018.

9.4.2 Other publications

Hickey PM, Condliffe R, Lawrie A, Kiely DG. Pulmonary hypertension *in Foundations of Respiratory Medicine*. Eds: Hart S, Greenstone M. Pub: Springer. 2018.

Hickey PM, Thompson AAR, Charalampopoulos A, Elliot CA, Hamilton N, Kiely DG, Lawrie A, Sabroe I, Condliffe R. Bosutinib therapy resulting in severe deterioration of pre-existing pulmonary arterial hypertension. *European Respiratory Journal.* 48(5): 1514-16. 2016.

Pickworth J, Rothman A, Iremonger J, Casbolt H, Hopkinson K, Hickey PM, Gladson S, Shay S, Morrell NW, Francis SE, West JD, Lawrie A. Differential IL-1 signalling induced by BMPR2 deficiency drives pulmonary vascular remodelling. *Pulmonary Circulation.* 7(4): 768-76. 2017.

Charalampopoulos A, Lewis R, Hickey PM, Durrington C, Elliot CA, Condliffe R, Sabroe I, Kiely DG. Pathophysiology and diagnosis of pulmonary hypertension due to left heart disease. *Frontiers in Medicine.* 5:174. 2018.

Sweatt AJ, Hedlin HK, Balasubramanian V, Hsi A, Blum LK, Robinson WH, Haddad F, Hickey PM, Condliffe R, Lawrie A, Nicolls MR, Rabinovitch M, Khatri P, Zamanian RT. Discovery of distinct immune phenotypes using machine learning in pulmonary arterial hypertension. *Circulation Research.* 124(6): 904-19. 2019.

Lewis RA, Johns CS, Cogliano M, Capener D, Tubman E, Elliot CA, Charalampopoulos A, Sabroe I, Thompson AAR, Billings CG, Hamilton N, Baster K, Laud PJ, Hickey PM, Middleton J, Armstrong IJ, Hurdman JA, Lawrie A, Rothman AMK, Wild JM, Condliffe R, Swift AJ, Kiely DG. Identification of cardiac MRI thresholds for risk stratification in pulmonary arterial hypertension. *American Journal of Respiratory and Critical Care Medicine.* [epub ahead of print]. 2019.