1



The Diagnostic Potential of miRNAs in Pulmonary Arterial Hypertension

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Philosophy

The University of Sheffield

Faculty of Medicine

Department of Infection Immunity and Cardiovascular Science

Mr James Iremonger

Registration number: 140226973

Supervisor: Professor Allan Lawrie

2nd Supervisor: Professor Neil Lawrence

Acknowledgments

This has been a long 6-year journey in which my family, friends and colleagues have all helped me through.

I will always remain indebted to Professor Allan Lawrie, his guidance, support and kindness throughout has made this possible, I could not have wished for a better supervisor. As a staff candidate Allan made this opportunity happen. In doing so he knowingly created more work and anguish for himself and I am forever grateful for this grand act of selflessness.

Every member of the PVRG lab has always been there to support me and I would like to specifically thank Dr Josephine Pickworth, Dr Laura West, Dr Alex Rothman, Nadine Arnold, Kay Hopkinson and Helen Casbolt. Your dear friendship and skilled knowledge in science and life was a source of never-ending help and support where even the most stupid of questions were met with answers. In particular, Josephine and Laura, without your patient reading and comments my completed thesis would never have been possible. And to Alex, who introduced me to microarray analysis with R, giving me the programming bug!

My time in Sheffield filled me with experiences and knowledge I never thought possible and will never forget. The head of department whilst I was there was Professor Sheila Francis and I thank her for crafting the most welcoming, friendly and academically stimulating atmosphere, of which I feel privileged to have been part. The whole of M-floor deserves thanks, but that would take too long, however, an exception must be made Dr Mark Ariaans. Without his technical expertise and unending willingness to help I doubt anything would have been achieved! ­

And finally, a huge thank you to my families for helping me along when I needed it. My loving wife, Kayleigh, for supporting me and my desire to undertake this MPhil despite the hindsight of knowing me during my undergraduate years. And my three wild children Harriet, Annabel and Oscar who provided grounding and a welcome distraction at the least opportune times. And to my parents for continually asking “how’s the thesis going?” for 6 years, this reminded me to keep pushing on.

Abstract

**Rationale**

Pulmonary Arterial Hypertension (PAH) is a rare life limiting and difficult to diagnose orphan disease and a common complication of systemic sclerosis. Micro-RNAs (miRNAs) are post transcriptional regulators of gene expression detectable in circulating fractions and have been reported as potential biomarkers in multiple diseases. Circulating miRNA indicators of survival in PAH have been published, however, diagnostic miRNA panels have yet to be explored.

**Objective**

To determine if circulating miRNAs can be used to identify the presence of PAH in patients with systemic sclerosis who are at a known risk of developing PAH.

**Methods**

A miRNA microarray screen was used to determine if plasma or whole blood was more suitable for biomarker detection in a case/control cohort (n = 28/17) of healthy volunteers (HV) and PAH.

A further microarray screen in HV/PAH (n = 33/47) was used to develop a multiple biomarker panel discovery process using ROC determined cut-offs.

Finally, potential miRNA biomarkers were identified in a microarray screen of SSc / SSc-PAH (n = 10/18) and validated using a larger cohort (n = 14/29) by qPCR.

**Results**

Plasma was selected as the source of miRNAs to use in further investigations. Differential expression analysis identified 90 plasma miRNAs vs 1 whole blood miRNA as significantly differentially expressed (p < 0.05). Additionally, in a feature selection task 0 miRNAs from whole blood were identified by the Boruta algorithm compared to 56 miRNAs from plasma.

11 miRNAs were selected for qPCR verification as constituents of potential biomarker panels. Mir-483-5p, miR-584-5p and miR-638 had acceptable amplification profiles and were tested in the larger SSc/SSc-PAH verification cohort. From those, miR-483-5p correlated with the microarray findings, r = 0.55, p < 0.01 and expression was increased in disease p = 0.02, adjusted p = 0.056. Furthermore, when all 3 miRNAs were used in a panel an increase in predictive power was seen from a maximum individual AUC of 0.69 to a panel AUC of 0.72.

**Conclusion**

This work highlights the potential utility of miR signatures as a diagnostic tool in PAH. MiRNA-483-5p was identified as a miRNA with relevance to PAH in patients with SSc, this miRNA has not previously been reported in PAH literature but has links to lung cancer and warrants further investigation.

Table of Contents

Acknowledgments ii

Abstract iv

Table of Contents vii

List of Tables xiii

List of Figures xv

Abbreviations xvii

1 Introduction 21

1.1 Pulmonary Hypertension 21

1.2 Pulmonary Arterial Hypertension 21

1.3 Systemic Sclerosis Associated Pulmonary Arterial Hypertension 23

1.4 Non-invasive testing in PAH 24

1.4.1 Echocardiography 24

1.4.2 Pulmonary Function Tests 25

1.4.3 NT-proBNP 25

1.4.4 Measured Walk Distance Measures 26

1.5 MicroRNAs 27

1.5.1 Changes in MiRNAs - cause or effect 28

1.5.2 MicroRNAs in PAH 29

1.5.3 Circulating MicroRNA biomarkers 31

1.5.3.1 Prostate cancer 31

1.5.3.2 Acute myocardial infarction 32

1.5.3.3 Sepsis 33

1.5.3.4 Frontotemporal dementia 34

1.5.3.5 Non-Alcoholic Fatty Liver Disease 35

1.5.4 MicroRNA biomarkers in PAH 36

1.5.4.1 microRNA-204 36

1.5.4.2 microRNA-150 37

1.5.4.3 microRNA-23a 37

1.5.4.4 Let-7 family in SSc-PAH 38

1.5.4.5 microRNAs in the WNT/β-catenin pathway 38

1.5.4.6 microRNA 140-5p 39

1.6 Measurement of miRNAs 40

1.6.1 Measurement of miRNAs by real-time PCR 40

1.6.2 Measurement of miRNAs by microarray 41

1.6.3 Measurement of miRNAs by Next Generation Sequencing 42

1.7 Software for the Analysis of High Throughput Data 43

1.7.1 Statistical Discovery of Biomarkers 44

1.7.1.1 Logistic Regression 46

1.7.1.2 Decision Trees 48

1.7.1.3 Random Forests 49

1.7.2 Feature Selection 51

1.7.2.1 Filter Based Feature selection 51

1.7.2.2 Forward selection and Backward Elimination 52

1.7.2.3 Least Absolute Shrinkage Selection Operator Feature Selection 53

1.7.2.4 Boruta Feature Selection 54

1.8 Performance Measures of Biomarkers 55

1.8.1 Receiver Operator Characteristic Curves 55

1.8.2 Confusion Matrix 56

1.9 Hypothesis, Aims and Objectives 57

2 Materials and Methods 58

2.1 Ethics, Patient Cohorts and Whole Blood Processing 58

2.1.1 Ethical Approval 58

2.1.2 Patient Cohorts 58

2.1.3 Total RNA Isolation from Whole Blood 58

2.1.4 Total RNA Isolation from Plasma 59

2.1.5 Concentration of Plasma RNA 60

2.1.6 Assessment of RNA Quality by Nanodrop 61

2.2 Global microRNA Profiling; Array Processing, Quality Control, Data Analysis 64

2.2.1 Agilent microRNA Array Processing 64

2.2.2 Agilent microRNA Microarray Pre-processing and Quality Control 64

2.3 Differential Expression 66

2.4 Identification of Biomarkers 67

2.4.1 Filtering 67

2.4.2 Receiver Operator Characteristics 67

2.4.3 Identification of Correlating miRNAs 67

2.4.4 Feature Selection 67

2.4.5 Combining Features into Panels 67

2.5 Quantification of miRNAs by qPCR 68

2.5.1 Complementary DNA Synthesis 68

2.5.2 Quantification of miRNAs by Real-Time PCR 69

2.5.3 Data Analysis 70

2.5.3.1 Quality control 70

3 Determining the Best Source of Circulating Biomarkers of IPAH 73

3.1 Aims 73

3.2 Introduction 73

3.3 Sample Cohort 74

3.4 Expression Changes in Blood Fractions 74

3.5 Determining the Best Source of miRNAs for Classification 77

3.6 Variation of miRNA Expression Between Whole Blood and Plasma 77

3.7 Conclusion 79

3.7.1 Variability of miRNAs 79

3.8 Deconvolution of miRNA Expression Profiles 80

3.9 Differentially Expressed miRNAs 82

3.9.1 Whole Blood 82

3.9.1.1 miRNA199a-5p 82

3.9.2 Plasma 82

3.9.2.1 miRNA199a-5p 82

3.9.2.2 Remaining differentially expressed miRNAs 84

3.10 Platelet miRNA Contamination in Plasma 85

4 Identification of Biomarker Panels of IPAH 87

4.1 Introduction 87

4.2 Methods 87

4.3 IPAH from HV Biomarker Feature Selection 88

4.4 Combining Biomarkers with Decision Trees 89

4.5 Combining Biomarkers by ROC Determined Threshold 90

4.6 Comparison of Biomarker Performance 91

4.7 Conclusion 92

5 Identification of Biomarkers of PAH in SSc 93

5.1 Introduction 93

5.2 Methods 93

5.3 PAH in SSc Biomarker Feature Selection 94

5.4 Combing biomarkers with decision trees 95

5.5 Combining biomarkers by ROC determined thresholds 97

5.6 Comparison of Biomarker Performance 98

5.7 Discussion 99

6 Validation of Microarray Results by QPCR 100

6.1 Initial MiRNAs for Selected for Verification 100

6.2 Methods 101

6.3 Initial qPCR Probe Verification with Test Cohort 101

6.3.1 Results 103

6.4 Additional MiRNAs for Selected for Validation 103

6.5 Verification of additional SSc-PAH ROC based biomarker miRNAs 104

6.6 Validation of miRNAS in SSc and SSc-PAH 104

6.6.1 Cohort 105

6.6.2 Methods 105

6.6.3 Results 106

6.6.3.1 Concordance with microarray data 106

6.6.3.2 Univariate Comparisons 107

6.6.3.3 Performance in a panel 108

6.6.3.4 Comparison of Panel Performance 110

6.7 Discussion 111

6.7.1 Multivariate Biomarkers 111

6.7.2 miRNA-486-5p 112

6.7.3 miRNA-486-5p in Platelets 112

6.7.3.1 miRNA-486-5p in Pulmonary Fibrosis 113

6.7.3.2 miRNA-486-5p in Lung Cancer 113

6.7.4 Validation Performance 114

7 General Discussion 116

7.1.1 microRNA-436-5p 119

7.1.2 Inclusion of additional data in panels 121

7.1.3 Experimental improvements 122

7.1.4 Technical Variation in miRNA Expression 126

7.1.5 Can Microarrays be Quantitative? 128

7.1.6 Future directions 129

8 References 132

List of Tables

Table 1‑1 Recommended process to evaluate the probability of PH via echocardiographic methods taken from the 2015 European Society of Cardiology/European Respiratory Society guidelines (Galiè et al., 2016). 25

Table 1‑2 Standardised layout for confusion matrixes. Relevant performance calculations are shown in the lowest and furthermost right cells. 56

Table 2‑1 Numbers of samples included in each biomarker discovery cohort. 66

Table 2‑2 Universal cDNA master mix with example calculations for 5 and 10 reactions. 69

Table 2‑3 Real-time PCR reaction master mix with examples for 10 and 50 reactions including 10% overage 70

Table 3‑1 Summary demographics of IPAH and HV samples used in the sample source cohort. Sample distributions were tested using Shapiro-Wilk test, samples with a p-value <= 0.05 were compared with the Mann Whitney U-test whilst a p-value > 0.05 was compared using a Students t-test. Categorical data was testing with Chi-Squared. 74

Table 3‑2 Number of miRNAs with a p-value <0.05 both with and without adjusting for multiple testing Benjamini-Hochberg false discovery rate. 75

Table 3‑3 miRNAs significantly differentially expressed in RNA from plasma comparing IPAH vs HV (absolute log fold-change ≥ 1.5 & adjust p-value ≤ 0.05). 84

Table 4‑1Summary of patient and control numbers from Sheffield and Imperial College in the IPAH biomarker identification cohort. 87

*Table 4‑2 Summary demographics of the Sheffield IPAH biomarker identification cohort. Normally distributed data shown as mean + standard deviation and comparisons made using Students t-test. Non-normally distributed data shown as median + interquartile range and comparisons made using Mann Whitney U-test. Categorical data was tested with Chi-Squared.* 88

Table 4‑3 Biomarkers identified by both the Boruta and LASSO algorithms. 89

Table 4‑4 Rpart decision tree performance metrics summarising the diagnostic power for the classification of IPAH from HV. 90

*Table 4‑5 Top 5 miRNA panels for diagnosing IPAH from HV and metrics for their diagnostic power. Performance metrics are shown as the mean and, in brackets, standard deviation from the cross-validation.* 91

Table 5‑1 Summary of patient and control demographics in the SSc with PAH biomarker identification cohort Normally distributed data shown as mean + standard deviation and comparisons made using Students t-test. Non-normally distributed data shown as median + interquartile range and comparisons made using Mann Whitney U-test. Categorical data was tested with Chi-Squared. 94

Table 5‑2 Biomarkers identified by the Boruta feature selection algorithm. 95

Table 5‑3 Rpart decision tree performance metrics summarising the diagnostic power for the classification of PAH within SSc. 97

*Table 5‑4 Top two miRNA panels for determining PAH in SSc. Performance metrics are shown as the mean and, in brackets, standard deviation from the cross-validation.* 98

Table 6‑1 miRNAs chosen for validation and their association with previously described disease panels. 101

*Table 6‑2 Demographics table for initial qPCR probe verification cohort of IPAH and HV samples.* *Normally distributed data shown as mean + standard deviation and comparisons made using Students t-test. Non-normally distributed data shown as median + interquartile range and comparisons made using Mann Whitney U-test. Categorical data was tested with Chi-Squared.* 102

Table 6‑3 Summary of initial qPCR verification results following inspecting of amplification curves. 103

*Table 6‑4 Additional panel of miRNAs for determining SSc-PAH from SSc. Performance metrics are shown as the mean and, in brackets, standard deviation from the cross-validation.* 104

Table 6‑5 Demographics table for SSc-PAH from SSc qPCR diagnostic test. Normally distributed data shown as mean + standard deviation and comparisons made using Students t-test. Non-normally distributed data shown as median + interquartile range and comparisons made using Mann Whitney U-test. Categorical data was tested with Chi-Squared. 105

Table 6‑6 Results of Spearman’s Rho correlation of real-time PCR expression with microarray expression. 107

Table 6‑7 Summary results from Mann-Whitney U test comparisons of SSc miRNA panel adjustment for multiple comparisons by Bonferroni. 108

Table 6‑8 Rpart decision tree performance metrics summarising the diagnostic power for the classification of PAH within SSc. 109

*Table 6‑9 All possible panel combinations of the miRNAs used in qPCR validation and performance metrics summarising the diagnostic power in determining PAH. Performance metrics are shown as the mean and, in brackets, standard deviation from the cross-validation.* 110

List of Figures

Figure 2‑1 Metrics determined from Nanodrop data A) RNA concentration B) 260/280 ratio B) C)260/230 ratio 61

Figure 2‑2 Nanodrop absorbance spectra from plasma RNA samples showing the overall (A), high purity samples (B), high sloping samples (C) and low sloping samples (D). 63

Figure 2‑3 Boxplots demonstrating that arrays failing either due to under or over hybridization have clearly different data distributions that can be easily identified. 65

Figure 2‑4 Example dissociation curves for 4 miRs of varying quality. Individual qPCR reactions are shown in grey whilst the blue line is a loess best fit. 72

Figure 3‑1 Volcano plot visualising the differences between miRNA expression and p-value for both plasma (red) and whole blood (green) miRNA microarrays. Whole blood miRNAS with an adjusted p-value ≤ 0.05 are labelled. For clarity in plasma miRNAS with a log fold change ≥ 1.0 and an adjusted p-value ≤ 0.05 are labelled. 76

Figure 3‑2 Number of miRNAs confirmed (light blue) or tentatively (dark blue) identified as important in the classification between IPAH and HV in the Boruta feature selection model. 77

Figure 3‑3 Density distribution plots of the standard deviation of measured miRNAs for IPAH and HV groups and sample types. 78

Figure 3‑4 Bar plot showing the 16 miRNAs from the plasma microarray with a standard deviation ≥ 1.00. 79

Figure 3‑5 Search string template used to identify specific miRNA related PAH publications. 85

Figure 3‑6 Comparison of miRNA microarray relative expression of platelet related miRNAs in Mitchel et al., to the two most highly differentially expressed miRNAs in HV vs IPAH plasma samples. 86

Figure 4‑1 Rpart decision tree for separating IPAH from HV. Each terminal node represents the assigned classification with the fraction of correctly classified samples given below. 90

Figure 4‑2 ROC curves for top ROC threshold panel, Rpart model in black, their constituent miRNAs and NTproBNP reference in purple. 92

*Figure 5‑1 Rpart decision tree for separating SSc with PAH from SSc. Each terminal node represents the assigned classification with the fraction of correctly classified samples given below.* 97

Figure 5‑2 ROC curves for top ROC threshold panel, Rpart model in black, their constituent miRNAs and NTproBNP reference in purple. 99

Figure 6‑1 Real-time PCR concordance with microarray expression blue loess trendline with 95% confidence boundary shown in grey. 106

Figure 6‑2 miRNA expression analysis by real-time PCR. After correction for multiple testing there are no significant differences. Mann-Whitney U test with Bonferroni correction. 107

Figure 6‑3 Rpart decision tree for separating SSc with PAH from SSc. Each terminal node represents the assigned classification with the fraction of correctly classified samples given below. 109

Figure 6‑4 ROC curves for top ROC threshold panel, Rpart model in black, their constituent miRNAs and NTproBNP reference in purple. 111

Figure 7‑1 Publications tagged with the MeSH term biomarkers within a corpus of PAH publications. 116

*Figure 7‑2 Bar chart summarising disease MeSH terms used in publications mentioning miRNA-486-5p.* 120

Figure 7‑3 Variance between replicates shown at the mean expression of each replicate for both healthy volunteers and both extraction methods. 0.5 log fold-change is shown on the chart as a light grey line for reference. 127

Abbreviations

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 6MWT | 6 minute walk test | | AGO | Argonaught proteins | | AIC | Akaike information criterion | | AMI | Acute myocardial infarction | | AUC | Area under the curve | | BNP | Brain natriuretic peptide | | cDNA | Complementary deoxyribonucleic acid | | CF | Cystic fibrosis | | CFTR | Cystic fibrosis transmembrane conductance regulator | | CO | Carbon monoxide | | CPET | Cardio pulmonary exercise testing | | CT | Computed tomography | | DLCO | Diffusing capacity of the lungs for carbon monoxide | | eNOS | Endothelial nitric oxide | | ePH | Exercise induced pulmonary hypertension | | FN | False negative | | FP | False positive | | FTLD | Frontotemporal lobar degeneration | | g | Gravity | | ILD | Interstitial lung disease | | IPAH | Idiopathic pulmonary arterial hypertension | | ISWT | Incremental shuttle walk test | | LASSO | Least absolute shrinkage and selection operator | | LIMMA | Linear models for microarray analysis | | LNA | Locked nucleic acid | | miRNA | Micro ribonucleic acid | | ml | Millilitre | | NAFLD | Non-alcoholic fatty liver disease | | NASH | Non-alcoholic steatohepatitis | | NGS | Next generation sequencing | | NPV | Negative predictive value | | NS | Non-significant | | NT-proBNP | N-terminal pro hormone brain natriuretic peptide | | NTC | Non-template control | | ºC | Degrees centigrade | | PAH | Pulmonary arterial hypertension | | PASMC | Pulmonary artery smooth muscle cell | | PAWP | Pulmonary arterial wedge pressure | | PFT | Pulmonary Function Test | | PH | Pulmonary Hypertension | | PPP | Platelet poor plasma | | PPV | Positive predictive value | | qPCR | Quantitative polymerase chain reaction | | RHC | Right heart catheterisation | | RISC | RNA-induced silencing complex | | RMA | Robust multichip average | | RNA | ribonucleic acid | | ROC | Receiver operator curve | | rpm | Revolutions per minute | | Se | Sensitivity | | SS | Simple steatosis | | Sp | Specificity | | SSc | Systemic Sclerosis | | SSc-PAH | Systemic sclerosis with pulmonary arterial hypertension | | STH-Obs | Sheffield Teaching Hospitals Observational Study of Patients with Pulmonary Hypertension, Cardiovascular and Lung Disease | | TN | True negative | | TP | True positive | | TRV | Tricuspid peak regurgitation velocity | | TTE | Transthoracic echocardiography | | VR-PAH | Vaso-responsive pulmonary arterial hypertension | | WHO | world health organisation | | µl | Microlitre | |  |

# Introduction

## Pulmonary Hypertension

Pulmonary hypertension (PH) describes a rare and diverse group of disorders defined as having a resting mean pulmonary arterial pressure of > 20 mm Hg at right heart catheterisation (RHC) (Simonneau et al., 2019). According to various combinations of haemodynamic observations and risk factors PH is categorised by the World Health Organisation (WHO) into 5 categories; Group 1: Pulmonary arterial hypertension, Group 2: Pulmonary hypertension due to left heart disease, Group 3: Pulmonary hypertension due to lung diseases and/or hypoxia, Group 4: Pulmonary hypertension due to pulmonary artery obstructions, and Group 5: Pulmonary hypertension with unclear multifactorial mechanism (Simonneau et al., 2009).

## Pulmonary Arterial Hypertension

In addition to a mPAP ≥20 mm Hg group 1 PH, diagnosis of pulmonary arterial hypertension (PAH) requires increased pulmonary vascular resistance of ≥3 Wood units in the absence of left heart disease and pulmonary thrombi (Simonneau et al., 2019). A normal right atrial pressure indicates the absence of left heart disease and is measured indirectly by the pulmonary artery wedge pressure (PAWP). A PAWP of ≤15 mm Hg is required for a diagnosis of PAH. Thrombi present would indicate a diagnosis of CTEPH, this is typically determined from a chest computer tomography (CT) scan.

PAH affects around 1-2 people per million worldwide with a 2:1 female bias (Shanmugam, Jena, & George, 2015), (Ma et al., 2013). Age at diagnosis is typically 30-50 years with a 3yr survival of 63% (Hurdman et al., 2012). Symptoms initially presented include shortness of breath, fatigue, lethargy, dyspnea, and syncope or pre-syncope upon moderate exertion (Y.-C. Lai, Potoka, Champion, Mora, & Gladwin, 2014).

PAH is a complex multifactorial disease affecting the lung vascular system and heart. It is primarily characterised by progressive cellular remodelling within the lung vasculature leading to narrowing, occlusion and complete obliteration of the distal arteries and arterioles (Humbert et al., 2004), (Rabinovitch, 2012). Proliferation of pulmonary arterial smooth muscle cells (PASMCs) is arguably the most substantial cellular change. However, initial endothelial cell death and later proliferation along with fibroblast migration and proliferation into the intima are important in the disease processes (Wei, Schober, & Weber, 2013).

The remodelled vasculature results in a paradigm shift in the lungs, changing from a low resistance system of high flow to one of high resistance now resisting flow. To overcome the resistance hypertrophy of the right heart occurs, resulting in the high pulmonary artery pressure. This unnatural growth impedes the normal action of the heart which ultimately fails under the strain, causing death (Y.-C. Lai et al., 2014).

PAH can be further classified as heritable, idiopathic or associated. Heritable PAH makes up a small percentage of all PAH cases, between 3% and 8% and is predominantly seen in families, with only a small number of spontaneous mutations in known heritable genes also classified as heritable. PAH is deemed idiopathic (IPAH) when no family history or predisposing risk factors can be identified (Simonneau et al., 2013)

Without treatment, PAH has a median survival of 2.8 years and poor 5-year survival rates (Ma et al., 2013). Current treatments roughly double the median survival to around 5 years, but they are limited in effect as only the symptoms are targeted not the underlying pathology (Hurdman et al., 2012), (Lawrie, 2014). The only treatment offering potential longer-term survival is heart and lung or lung transplant, these are rare and infection and rejection are major complications and survival is still limited (Montani et al., 2013).

## Systemic Sclerosis Associated Pulmonary Arterial Hypertension

Systemic sclerosis (SSc) is a condition of unknown origin characterised by fibrosis of the skin and internal organs (Hachulla & Launay, 2011). In the UK there is a reported prevalence of 8.2 per 100,000 with a median age of 57.1 years (Allcock, Forrest, Corris, Crook, & Griffiths, 2004).

PAH is a major complication associated with SSc (PAH-SSc) with a not insignificant prevalence between 7.8–12% (Hachulla et al., 2005), (AVOUAC et al., 2010). Compared with IPAH, PAH-SSc presents with milder haemodynamics at diagnosis, however, the disease progression is more rapid with a 3 year survival of 52%, significantly less than IPAH 63% p<0.01 (Hurdman et al., 2012).

Due to the relatively high prevalence, and high mortality of SSc-PAH it is recommended in the Society of Cardiology (ESC)/European Respiratory Society guidelines that patients undergo annual screening (Galiè et al., 2016). It is difficult to assess whether improvements in survival from early detection are real or if it is lead-time bias as the diagnosis has been made earlier within the natural course of the disease. To discern the difference Humbert et al compared the survival of 32 incidental patients identified between 2002 and 2003, 16 were symptomatic upon referral and 16 were detected through echocardiography screening. The symptomatic cohort had a 3 year survival of 31% compared with 81% for the screened cohort (p <0.01) (Humbert et al., 2011). Despite the small number of patients studied this indicates that a successful screening program can benefit patient survival.

## Non-invasive testing in PAH

It is impractical to send every patient suspected of PH for RHC diagnosis, preliminary screening is typically undertaken by several non-invasive measures so that only patients with a high likelihood of PH are catheterised. However, several tests are typically undertaken, some by specialists, with varying degrees of success in identifying PH.

## Echocardiography

Transthoracic echocardiography (TTE) is recommended to be performed when a diagnosis of PH is suspected (Galiè et al., 2016). TTE can identify the effects of PH on the heart as well as estimating PAP. However, further investigations should not be based upon estimated PAP as there can be severe over or under estimation of PAP influencing clinical management (Augustine et al., 2018).

The primary measurement is tricuspid peak regurgitation velocity (TRV), A TRV of >3.4 ms is indicative of PH, requiring RHC. However, a TRV <3.4 ms does not preclude the diagnosis of PH and the likelihood of PH must then be assessed by combining the results of several other echocardiographic measurements, detailed in Table 1‑1.

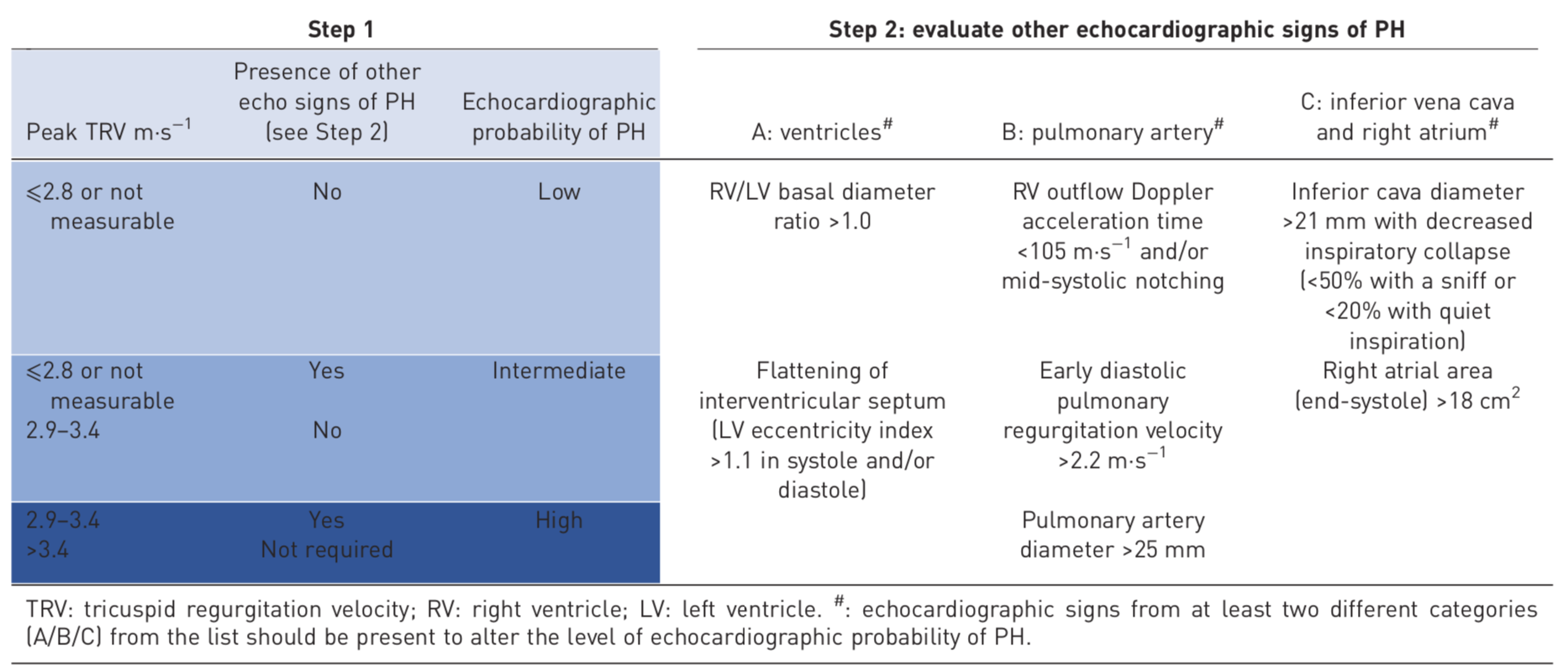


Table 1‑1 Recommended process to evaluate the probability of PH via echocardiographic methods taken from the 2015 European Society of Cardiology/European Respiratory Society guidelines (Galiè et al., 2016).

## Pulmonary Function Tests

Pulmonary function tests (PFT) measure the lung function and are economical. A large variety of measures are available during PFT the most useful in PAH is the diffusing capacity of the lungs for carbon monoxide (DLCO). This is an indicator of impaired diffusion of gases into the blood. Patients inhale a gas with known quantity of carbon monoxide (CO) and hold for 10 seconds, and then the quantity of CO is measured upon exhalation (MacIntyre et al., 2005). A DLCO of < 50 % is predictive of PAH whilst a DLCO < 45 % is associated with a poor 3 year survival of 38% compared to 80% (p = 0.002) ((Mukerjee et al., 2004), (Trip et al., 2013)).

## NT-proBNP

The N-terminal fragment from the protein pro-BNP (NT-proBNP) is currently the only guideline recommended blood-based biomarker and can be used to aid diagnosis and distinguish patients with poor prognosis (Galiè et al., 2009), (Galiè et al., 2016), whereby High levels are associated with poor survival (Williams et al., 2006), (Fijalkowska et al., 2006). NT-proBNP is the stable cleavage product of brain type natriuretic peptide (BNP) which is released in response to ventricular stress (Warwick, Thomas, & Yates, 2008). BNP relieves ventricular stress by lowering blood pressure, by signalling venous dilation.

## Measured Walk Distance Measures

Due to their simplicity walking tests measuring distance travelled are commonly used to assess patient exercise capacity. Currently there are two protocols used, most commonly the distance covered within a timed walk, the 6-minute walk test (6MWT), or the distance covered during an incremental shuttle walk test (ISWT).

In the ISWT a patient is required to walk between two targets 10-meters apart. The time given to walk between the targets is gradually reduced until the patient fails to reach a target within the allotted time. The ISWT is advantageous as it only requires a 10-meter corridor instead of the 30-meter corridor used it the 6MWT. The change in ISWT distance measured at 1-year following start of treatment has been shown to be prognostic in patients with PAH (Billings et al., 2017). With a distance ≤120 meters being associated with an increased 2-year mortality.

It was recently demonstrated in the 6MWT that patients who walked <165 meters was predictive of higher likelihood of mortality within 1 year and conversely a distance >440 meters being predictive of survival (Zelniker et al., 2018). However, the study noted that the change in 6MWT whilst prognostic was less predictive than the absolute distance travelled.

Research into translating walk tests into a form of active-monitoring has been suggested by replacing walk tests with accelerometer-based measurements. Sehgal et al., monitored 20 patients between follow-up visits 136.4 (± 47.3) days using wrist-based step counters. The change in daily step count correlated with the change in 6MWT, however, the accelerometer results did not correlate with measurements for NT-proBNP (Sehgal et al., 2019).

## MicroRNAs

Micro-RNAs (miRNA) are relatively recently described mediators of post-transcriptional regulation of gene expression. They are short (21-24nt) non-coding single stranded RNAs that negatively regulate gene expression (Lee, Feinbaum, & Ambros, 2004),(Flynt & Lai, 2008). Originally discovered in nematodes they are now known to be key regulators of gene expression with 48,860 mature sequences across 271 organisms reported in miRBase release 22 (Kozomara, Birgaoanu, & Griffiths-Jones, 2019).

MicroRNAs are synthesised from DNA by two distinct pathways (Winter, Jung, Keller, Gregory, & Diederichs, 2009) In the first pathway an initial larger primary miRNA (pri-miRNA) transcript is transcribed in the nucleus. The pri-miRNA has an average size of 33bp in humans and contains a hairpin stem and terminal loop with both 5’ and 3’ single stranded flanking regions (Winter et al., 2009). The pri-miRNA is then cleaved 11bp from the start of the single stranded flanking regions by a complex of the RNase III enzyme Drosha and DGCR8 producing a pre-miRNA (Winter et al., 2009). In the second pathway Drosha/DGCR8 processing of a pri-miRNA is not necessary. Known as mirtrons, the pre-miRNA is directly synthesised from introns during mRNA splicing (Ruby, Jan, & Bartel, 2007). Pre-miRNAs from either pathway are then transported from the nucleus to the cytoplasm by Exportin-5, and are subsequently cleaved by the RNase III enzyme Dicer (Winter et al., 2009). Dicer cleavage yields a mature double stranded miRNA:miRNA duplex containing a guide and passenger strand (Okamura, Hagen, Duan, Tyler, & Lai, 2007). Helicases separate the two strands and the guide miRNA is loaded with Argonaute (AGO) proteins and incorporated into the RNA-induced- silencing complex (RISC), whilst the passenger strand is degraded (Zhou, Chen, & Raj, 2015), (Winter et al., 2009). Once bound in the RISC complex miRNAs, through sequence homology, bind to specific 3’ targets of mRNA and by either translational repression or promotion of mRNA degradation reduce translation (Winter et al., 2009), (Zhou et al., 2015).

## Changes in MiRNAs - cause or effect

MicroRNAs have been associated with many common human disorders (Yu Li & Kowdley, 2012), however, the precise relationship between miRNA and disease is difficult to determine. Changes in miRNA expression may be driving disease or a resultant effect of disease processes.

Changes in miRNA expression have been detected in patients with sepsis where the root cause is a dysregulated immune response is a bacterial infection (Caserta et al., 2016). However, prokaryotes do not produce miRNAs indicating the miRNA changes seen reflect a non-specific effect (Furuse et al., 2014).

Mutations in cystic fibrosis transmembrane conductance regulator (CFTR) gene have been known to be causative of cystic fibrosis since 1989 (Kerem et al., 1989). Furthermore, CF can be induced in mouse models by mutation of the CFTR gene alone (Dorin et al., 1992), (McHugh et al., 2018). Whilst the genetic driver behind CF is a faulty ATP-gated anion channel miRNA dysregulation is still seen (JE, X, Rangaraj, S, & H, 2019), (Oglesby & McKiernan, 2017). Even though miRNA changes are not causative they can still have a negative effect, adding to disease processes. Bhattacharyya et al., identified that elevated expression of miR-155 was to blame for increased IL-8 activity, further exacerbating the pro-inflammatory phenotype of CF (Bhattacharyya et al., 2011).

Conversely, there are examples of miRNAs having strong links with the pathobiology and that altering the candidate miRNAs expression can ameliorate measures of disease. This has been demonstrated most clearly using miRNA conditional knockout mice. The role of miR-155 dysregulation in pulmonary fibrosis was investigated using an endothelial specific conditional knockout of miR-155 mice. In the bleomycin mouse model of pulmonary fibrosis Tang et al., found reported that conditional deletion of miR-155 resulted in a reduced EndoMT response and reduced lung fibrosis.

## MicroRNAs in PAH

Dysregulation of microRNA expression has been implicated as part of the pathophysiology of PAH (Chun, Bonnet, & Chan, 2017).

Reduced BMPR2 signalling is a key pathway in PAH pathogenesis, whereby loss of BMPR2 induces PASMC proliferation (NW et al., 2001). Additionally, BMPR2 haploinsufficiency is found within hereditary PAH as the result of mutations in the encoding gene(International PPH Consortium et al., 2000). BMPR2 expression has been shown to be indirectly moderated by miR-140-5p (Rothman et al., 2016). SMURF1 is a negative regulator of BMP signalling and expression increased in patients, over expression of miR-140-5p was shown to decrease SMURF1 expression, increasing BMP signalling in PASMCs. In further experimentation over expression of miR-140-5p was shown to reduce mPAP in rats and proliferation in PASMCs. Direct regulation of BMPR2 expression has been shown with miR-21-5p (Parikh et al., 2012). In hypoxic pulmonary arterial endothelial cells miR-21-5p expression was increased and increased expression repressed BMPR2 expression.

Reduced BMPR2 expression results in increased cellular proliferation, a hallmark of PAH pathobiology. In human lung tissue the expression miR-204-5p is reduced inducing increased NFAT, RUNX2 and HIF-1α (Courboulin et al., 2011), (J et al., 2015). This leads to metabolic dysfunction and an increase in cellular proliferation in PASMCs.

MiRNAs have also been shown to control elements of another hallmark of PAH, vasoconstriction. One animal model of PAH is the Sugen 5416 hypoxic mouse model, in this Sugen 5416, a vascular endothelial growth factor receptor antagonist, is injected and the animals develop PAH when subjected to chronic hypoxia for > 3 weeks (C. L et al., 2011). However, Bertero et al., found in this model that even in the absence of hypoxia Sugen 5416 and intrapharyngeal injection of liposomally encapsulated miR-130a alone could induce a significant increase in mPAP (Bertero et al., 2015). When the lungs from the animals were examined expression of the vasoconstrictor endothelin-1 was increased in the small pulmonary vasculature. Furthermore, a positive correlation of miR-130a expression to endothelin-1 concentration was reported in the plasma of patients with PAH, demonstrating that endothelin can be regulated by miRNA.

## Circulating MicroRNA biomarkers

MicroRNAs are promising potential candidate biomarkers, they have been isolated from bodily fluids including plasma, serum, cerebrospinal fluid, saliva, tears, amniotic fluid, breast milk and urine (J. A. Weber et al., 2010) and are resistant to degradation (Glinge et al., 2017). However, despite intense research as yet none have been successful in translating to clinic (Saliminejad, Khorram Khorshid, & Ghaffari, 2019). Some examples of successfully validated miRNA biomarkers utilising differing discovery strategies are detailed below.

## Prostate cancer

Cancers can have significantly altered genetic profile of which may potentially lead to dramatic miRNA profile changes that may be detectable in plasma. To identify a miRNA panel to diagnose prostate cancer Matin et al., used a focused qPCR array which profiled 372 cancer-associated miRNAs (Matin et al., 2018). Pooled plasma samples were used for the qPCR screen, 19 healthy volunteers and 42 patients with prostate cancer produced 2x pools of HV and 3x pools of patient samples. The qPCR array identified 11 candidate miRNAs which were validated individually in the samples in the pools. The miRNAs 98-5p, 152-3p, 326 and 4289 were significantly increased, by Bonferroni corrected Mann-Whitney U-test. These were successfully validated in an external cohort of 40 patients and 18 HVs.

Logistic regression was used to assess the 4 miRNAs as a panel, the internal cohort had an AUC of 0.82 and the external cohort an AUC of 0.95. It is highly unusual to have a predictive model to improve when going from train to test cohort and whilst possible it is more likely a new logistic regression model was built with the external validation data set. Ideally, a model built using the test data should have been applied to the external data. This would have given a measure of how stable the prediction would be in diagnostic use.

This study demonstrates an economical approach to biomarker identification through pooling initial samples and using a focused miRNA panel. However, if they had utilised a more traditional training/test dataset approach to the discovery and validation panels more insight into the diagnostic effectiveness of the 4 miRNA panel could have been determined.

## Acute myocardial infarction

To identify miRNA biomarkers of acute myocardial infarction (AMI) Adachi et al., first used a microRNA microarray screen across 15 different tissue types to identify miRNAs unique to cardiac tissue (T et al., 2010). MiR-499 was identified and then measured in plasma by qPCR in a validation cohort of patients with chronic heart failure (n = 15), AMI (n = 9) add healthy volunteers (n = 10). Whilst the n was small miR-499 was only above the limit of detection in AMI patients and the expression was large enough to be statistically significant.

This novel strategy of identifying tissue specific miRNAs before measuring in plasma it isn’t without precedent. Cardiac troponin is specific to heart muscle measuring and is measured in clinical practice, its release into the circulation caused by myocardial injury/necrosis during AMI (Thygesen et al., 2010).

Whilst there was no external cohort, however, the increase in miR-499 following AMI has been confirmed in sever subsequent publications (L. X et al., 2015), (Cheng, Wang, You, Chen, & Xia, 2014).

## Sepsis

To identify potential new therapies and improve outcomes in sepsis Visilescu et al., sought to identify new diagnostic markers (Vasilescu et al., 2009). A low-resolution miRNA microarray screen of 470 miRs in the peripheral blood mononuclear cells (PBMC) fraction in 8 patients with sepsis and 8 controls was performed. This identified 17 miRs that could differentiate disease, 4 of which had an absolute log fold-change > 2 which were further validated by qPCR in a larger cohort of 10 patients and 12 controls. The qPCR results validated the microarray performance confirming that miR-182 and miR-486 were upregulated in sepsis and miR-150 and 342-5p were down regulated. Further investigations found that the decrease in miR-150 expression in patients was also present in plasma. Importantly, this did not correlate with total white blood cell count and was therefore not a reflection of PBMC contamination.

The study could have demonstrated the predictive potential by combining the 4 miRs into a singular score and the samples numbers were also quite small, with no external validation cohort used. However, miR-150 serum levels have since been shown to predict survival in large cohort of 138 patients with sepsis and 76 controls (Roderburg et al., 2013).

## Frontotemporal dementia

As neurodegenerative diseases are typically challenging to diagnose Grasso et al., sought to identify a minimally invasive miRNA measure of frontotemporal lobar degeneration (FTLD). In plasma, 3 miRNAs were significantly downregulated miR-206, miR-502-3p, miR-663a in FTLD when compared with healthy controls (Grasso et al., 2019). Initial discovery was performed by profiling the expression 752 miRNAs using Exiqon qPCR arrays in 10 FTLD and 10 HVs. In the cohort a total of 221 miRNAs were found to be expressed following initial filtering of Ct < 37 and presence within ≥ 25% of samples. Following differential expression analysis 7 significantly differentially expressed miRNAs were chosen for validation in a larger cohort of 48 FTLD and 46 HVs, identifying 3 significantly downregulated miRNAs. The 3 miRNAs were then combined in their possible combinations using a Bayesian approach. This identified a panel of miR-502-3p and miR-633a that had a classification accuracy of 84.4%.

Good filtering and well balanced discovery and validation cohorts helped identify 3 miRNAs of interest. However, as a diagnostic biomarker it would have been good to have tested in diverse cohorts to establish whether the miRNAs could distinguish FTLD from other neurodegenerative disorders.

## Non-Alcoholic Fatty Liver Disease

Non-alcoholic fatty liver disease (NAFLD) is difficult to diagnose and can progress to the more serious non-alcoholic steatohepatitis (NASH). Pirola et al., sought to identify serum miRNA biomarkers that were diagnostic and followed disease progression (Pirola et al., 2015). Discovery was performed using a qPCR array on a cohort of 16 HVs, 16 NAFLD grade 1; simple steatosis (SS), and 16 NASH. The qPCR array was of custom design utilising 84 miRNAs identified in disease relevant literature. 6 miRNAs were identified by differential expression for validation in a larger cohort of 19 HVs, 30 SS, and 47 NASH. The validation confirmed significant upregulation in miR-122, miR-192 and miR-375. Additionally, these miRNAs were significantly associated with disease severity.

The miRNAs were assessed individually for their discriminatory power by ROC analysis, all 3 had similar ROC AUC values of 0.7. A large diagnostic panel was built with the miRNAs and common the common liver biomarkers and disease measures; aspartate transaminase (AST), alanine transaminase (ALT), cytokeratin 18 (CK18), body mass index, waist measurement. Unfortunately, it was not clear what method was used to combine all features in unison and strangely there was no increase in ROC AUC measurement to a single miRNA was seen. With additional features, even if it was overfitting, an increase in performance is likely.

By utilising prior knowledge this study was able to use a focused liver disease discovery panel that identified 3 serum miRNAs of NAFLD and its progression. Whilst validation with an external cohort was not performed the miRNAs were shown to correlate with disease progression demonstrating a promising relevant relationship with disease. Additionally, miRNA-122 is a well-known highly liver specific miRNA (L.-Q. M et al., 2002), (Thakral & Ghoshal, 2015) and the upregulation identified in patient serum is likely to be linked to deterioration of the liver.

The larger panel may have been more successful if different models had been tested, or if model stability was explored in a train and test scenario. Additionally, the AST/ALT ratio is an informative measurement of liver function and calculating this as an input in the model would have been appropriate and could have aided in predictive performance.

## MicroRNA biomarkers in PAH

The publication landscape regarding biomarkers in PAH was investigated by identifying, in PubMed, the subset of PAH publications which mentioned or were tagged with the MeSH term “biomarkers”. Only publications describing miRNA levels measured in patient samples of Group I PAH were interrogated and reported upon.

## microRNA-204

The first suggestion of a miRNA providing diagnostic biomarker in PAH was in 2011 when Courboulin et al., reported decreased levels if miR-204 in patients with IPAH (Courboulin et al., 2011). A significant decrease (p < 0.01) in expression was detected in buffy coat in IPAH when compared with control patients (hypertensive, not PH) (n = 13 IPAH, 7 Control) by qPCR. Whilst the paper explicitly described miR-204 as a “potent biomarker for PAH” and suggesting it “could become a reliable biomarker of PAH” further performance metrics beyond a significant t-test were not provided. Biomarkers not being the main topic of the study and the small sample numbers may in part be the reason for not exploring further.

Additionally, whilst not explicitly mentioned as a biomarker of severity miR-204 expression in PAH lung biopsy tissue was shown to negatively correlate with PVR. However, it was not reported if this correlation could be detected in the circulating fraction.

## microRNA-150

Rhodes et al., performed an unbiased screen on microRNA isolated from IPAH and control plasma, identifying miR-150 as a prognostic indicator of survival (Rhodes et al., 2013). Discovery was performed by microRNA microarray on 8 IPAH and 8 healthy volunteers, once miR-150 had been identified as the lead candidate miRNA survival analysis was performed on the qPCR expression of 175 patients with PAH. This identified that decreased miR-150 correlated with reduced 2-year survival (p <0.01)

## microRNA-23a

Increased microRNA-23a expression was seen following a microRNA microarray profiling serum from 12 IPAH and 10 healthy volunteers (Sarrion et al., 2015). Differential expression analysis identified 61 significant miRNAs of which 15 were qPCR validated. Following the qPCR results miRNA-23a was suggested as a potential biomarker as it had significant correlations with mPAP, CI and PVR.

Uncharacteristically for this type of study summary statistics of patient demographics were not detailed and p-values from the qPCR validation were not reported obfuscating the study and miRNA performance. Additionally, this study was not performed on treatment naïve patients it is therefore unknown whether miRNA-23a expression would remain diagnostic. Even so with further investigation and larger study size miRNA-23a may still be a biomarker of patient pulmonary function.

## Let-7 family in SSc-PAH

Skin biopsies taken from 6 SSc-PAH and 9 SSc patients identified 6 members of the let-7 family as being downregulated in PAH (Izumiya et al., 2015). This would provide a simple, albeit uncomfortable test for PAH for an at-risk group. The discovery was performed using qPCR arrays limiting the number of miRNAs that could be queried to 1066. Furthermore, statistical analysis was performed by multiple t-tests without correcting for multiple testing. The most significant change seen was in let-7d p = 0.020, it is unlikely that this would remain significant after correction for multiple testing.

## microRNAs in the WNT/β-catenin pathway

Reduced expression of miR-26a that correlated with 6-minute walk distance was suggested as a potential biomarker following microRNA microarray analysis comparing lung tissue from 8 IPAH too 8 healthy volunteers (Wu et al., 2016). 21 miRNAs were differentially expressed, 11 targeted genes in the WTN/β-catenin pathway and were validated by qPCR in a larger cohort of 22 IPAH and 22 healthy volunteers. 6 miRNAs had significant group differences and all demonstrated diagnostic potential with the ROC AUC between 0.71 and 0.86. However, only miR-26a was suggested as a potential biomarker due to its correlation with 6-minute walk distance.

As with miRNA-23a in Sarrion et al., this study was undertaken in patients with established disease and the biomarker may not be effective in a treatment naïve diagnostic cohort.

## microRNA 140-5p

Rothman et al., identified a reduction in miRNA 140-5p in whole blood RNA from patients with PAH when compared to controls (Rothman et al., 2016). The discovery screen for novel miRNA’s was performed using a Taqman Human microRNA V2.0 TLDA array (card A). This assayed the expression levels of 377 miRNAs from IPAH, SSc-PAH, CTEPH, HV and SSc without PAH (n = 4 each group). Upon further characterisation, reduced expression was also seen in both monocrotaline and sugen hypoxic animal model of PAH as well as in the lesions of IPAH patient’s lung tissue.

The study indicated the diagnostic potential of miRNA-140-5p with ROC curve analysis. An AUC of 0.74 was calculated using the qPCR results taken from the validation of cohort of 20 IPAH and 16 healthy volunteers. This is a small cohort and the diagnostic potential would ideally need to be further investigated in a larger international cohort. Additionally, as miRNA 140-5p’s direct link to disease has been shown follow-up investigations utilising a longitudinal cohort would be worthwhile and may identify prognostic and/or diagnostic utility.

## Measurement of miRNAs

MiRNAs are measured using adaptations of techniques used for measuring mRNA such as real-time PCR, microarray and next generation sequencing (NGS). For utility in biomarker discovery real-time PCR can be multiplexed, typically testing <1000 miRNAs across multiple assays on a single sample. Whilst microarray and NGS are massively multiplexed and unbiased techniques with the potential to detect all of the miRNAs present in a sample.

It is advantageous with real-time PCR that the discovery is being performed in using the validation technology, however, the number of miRNAs that can be tested is low. Unbiased approaches are a broader platform for biomarker discovery but require different technologies for validation.

## Measurement of miRNAs by real-time PCR

At roughly the size of a single primer, miRNAs are too short to be amplified by PCR and must first be extended in length. The typical strategy used to overcome this is to extend the miRNA by ligating a poly-A tail to the 5’ end. This enables a forward primer specific to the miRNA and a reverse primer to bind to the poly-A.

The Exiqon real-time PCR system uses locked nucleic acid (LNA) bases in the primers. LNAs are chemically modified nucleotides that offer significantly increased base pair affinity resulting in improved sensitivity/specificity and shorter primers (Koshkin et al., 2005). The LNA primers are short enough that the forward primer is fully miRNA specific, however, there is only enough room on a miRNA for a semi-specific reverse (3 bases) (Kauppinen et al., 1998). Real-time experiments are typically run in at least replicate therefore using a 384-well plate layout it is possible to assay up to 192 miRNAs at once, with larger numbers of miRNAs requiring multiple plates. Exploratory assays are typically limited to <1000 miRNAs therefore biasing experiments by the choices made when selecting miRNAs to assay. A large quantity of RNA is required and low-yield sources of RNA such as plasma/serum may not be concentrated enough to use.

## Measurement of miRNAs by microarray

Microarrays are capable of interrogating thousands to millions of genetic targets. Sequences complementary to both known and predicted miRNAs are deposited onto the surface of glass slide in a dense grid pattern. Each grid reference corresponds to a specific miRNA. Fluorescently labelled RNA is washed over the slide in conditions favouring specific binding of RNA to their complementary strand on the microarray. The microarray is then scanned at high resolution (2-10µm per pixel) and miRNA expression levels are inferred from the relative spot intensities of each probe.

To ensure specificity of the base pairing microarray probes are typically longer than 40 bases. As with qPCR miRNAs are too short for this approach and Exiqon miRNA microarrays have probes containing LNA nucleotides whilst Agilent use probes with a hairpin design. The hairpin limits the physical binding space to the size of the miRNA of interest prohibiting the binding of longer sequences (H. Wang, Ach, & Curry, 2007). Specificity is further increased by addition of a cyanine base at the 3’ end of the miRNA, extending their length by 1 base pair and creating a GC clamp further increasing binding stability.

Microarrays offer a significant increase in capacity from large scale real-time PCR providing unbiased miRNA profiling coving up to 100% of known and predicted miRNAs However, due to the nature of the technology dynamic range, sensitivity and specificity are sacrificed. Additionally, processing and analysis of microarray results requires specialist technical and bioinformatic skills.

## Measurement of miRNAs by Next Generation Sequencing

NGS involves the simultaneous sequencing of tens of millions of target miRNAs in a massively multiplexed reaction.

Whilst several competing NGS technologies have evolved the market leader is Illumina who’s platform builds up sequencing by synthesis (Quail et al., 2012). Sequencing is performed on a flowcell, a glass slide with oligonucleotides complementary to terminal sequences upon the sequencing adapters. MiRNAs are ligated to forward and reverse specific adaptors and hybridised to the flow cell. A single miRNA strand would not be detectable therefore clusters of clones are generated by bridge amplification. The sequencing reaction then proceeds by 1 base at a time by incorporating a single reversible fluorescent nucleotide terminator to the miRNA clusters. This is excited, read, then the terminator is chemically modified to allow further nucleotide addition and the fluorophore removed. The addition of one base, detection and resetting of a new terminator is then repeated until the whole strand has been read (Bentley et al., 2008). To determine expression levels the reads are aligned to a known reference sequence and the frequency of clones is used to determine expression levels.

By quantifying the expression in terms of frequency of RNA sequence there is little, albeit quantifiable, ambiguity in the species of miRNA detected. As miRNA detection is not based upon hybridisation of miRNAs to primers or probes NGS avoids issues with non-specific hybridisation. NGS of miRNAs (miRNA-seq) has increased dynamic range over microarrays, can detect novel miRs, and greater specificity (Mestdagh et al., 2014).

Access to miRNA-seq is becoming more common with the cost of reagents and machines reducing annually. However, there is a significant jump in technical and bioinformatic support required. It is rare that samples for NGS to be processed outside of a dedicated core-facility and analysis requires dedicated bioinformatic and high-performance computing support. This keeps associated costs high and therefore large-scale studies are typically impractical.

## Software for the Analysis of High Throughput Data

High throughput data poses unique analytical challenges, both practical and statistical issues arises with handling extremely large data sets. Practically, data is typically too large to be analysed in spreadsheet style software. Specialist analytical programs are available that are designed for big data analysis, the most prominent are R®, Python®, SAS®, SPSS®, Tableau®. All of the aforementioned packages are designed for big data analytics, but do so with varying degrees of versatility, cost and learning curve. Tableau® and SPSS® utilise graphical user interfaces that lend themselves to a shallow learning curve, however, both are moderately expensive and do not implement advanced analytical tools. SAS® uniquely offers both a relatively shallow learning curve and advanced analytical tools, however, the package is designed for the business market and prohibitively expensive to academic research. R® and Python® are both programming languages that are interfaced with through the command line and scripting, whilst this makes for a steep learning curve both are free open-source and highly versatile. Typically offering the latest cutting-edge analytical techniques through add on packages. R® and Python® are both perfectly suitable for biomarker discover, however, R® has a large bioinformatic community and class leading data manipulation tools. Additionally, R® has many packages specifically for the analysis of microarrays.

## Statistical Discovery of Biomarkers

Data containing hundreds to thousands of variables is commonly termed high-dimensional data. High-dimensional data has unique properties offering advantages and disadvantages in comparison to small focused data sets. Identifying biomarkers that separate two groups can be performed in two ways; 1) a classical statistical approach whereby each data measure, or feature, is analysed by multiple monovariate statistical tests or 2) a supervised statistical learning approach using a subset or all features available in combination.

In the classical approach, multiple t-tests or Mann-Whitney U-tests for parametric and non-parametric data respectively would be performed. As the probability of a statistically significant result occurring due to random chance increases with the number of monovariate tests performed a post-hoc adjustment to the p-value must applied to control for false positives. The Benjamini-Hochberg procedure is commonly used, this controls the Type I error rate, reducing Type II errors. For example for a 10% Type I rate means that in 100 significant features there is a high probability that 10 will be due to random chance (Benjamini & Hochberg, 2010). In spite of multiple correction approaches these strategies do not scale well and do not take advantage of any complex relationships that may be present within the data.

Supervised statistical learning methods are approaches that gain additional power by considering the relationships between multiple features and the classification groups. The statistical learning algorithms logistic regression, decision trees and random forests are well suited to biomarker discovery, however, the data and the projects aims dictate will guide the choice of algorithm. What follows is not an exhaustive list of classification algorithms, it is a highlight of models appropriate for biomarker discovery.

Not included are complex models such as gradient boosted machines or neural networks. These models are generally more flexible meaning that they can identify more subtle patterns and relationships which can lead to better predictive accuracy. However, they typically need impractically large sources of data otherwise overfitting is likely. Overfitting is where a model has learned patterns that are too specific, effectively describing the training data and not identifying the underlying patterns present. An overfitted model will have very high accuracy in the dataset it has been trained upon but is not generalised and will have very poor performance in future data. A well generalised model has determined the overall patterns in the sample, these are applicable to new data and accuracy in the new data can only marginally reduced. Conversely, underfitting is where a model is not flexible enough to fit the patterns present and the model has both poor accuracy in the sample trained upon and any future data.

More complex models have not been shown despite the fact that they can offer substantial improvements in accuracy are inherently unsuitable for biomarker discovery. These models require prohibitively large data sets and work like a “black box” where data is input, processed and a resulting classification is given.

Finally, complex models are avoided because as complexity increases interpretability decreases meaning the reason behind a given classification cannot easily be understood. This has arisen to many machine learning algorithms being dubbed “black boxes”. In the context of disease classification this is not an ideal and confidence in a black box algorithm can only be built through extensive validation.

## Logistic Regression

In medical research logistic regression is commonly used to model risk factors associated with patients (Sanada et al., 2019), (Radchenko, Zhyvylo, & Sirenko, 2019). However, logistic regression can also be used for the classification of a dichotomous dependant variable with one or more independent variables. It is popular due to the simplicity and robustness of the statistical formula and the ease of interpretation of the output. Interactions between features can be included modelling non-linear effects, however, they must be known a priori or empirically discovered.

Due to the simplicity of logistic regression, underfitting is a concern as the model may be unable capture the patterns within the data it is trying to classify. Conversely with high dimensional data there maybe too many features leading to overfitting.

Logistic regression was used to classify and identify diagnostic biomarkers in the peripheral blood mononuclear cells of 39 SSc-NoPH and 25 SSc-PAH (Moll et al., 2018). Microarray gene expression was initially filtered by identifying genes correlated with the presence of PAH to find sets of intra-correlated genes. From each gene set 1 or 2 genes were selected to represent the sets based upon a combination of c-statistic, largest fold-change, and biological plausibility. The cohort of SSc patients was split into those with interstitial lung disease (ILD) and those without, the filtering was performed separately to each subset. This identified 5 genes in SSc with ILD and 4 genes in SSc without ILD. Logistic regression with stepwise forward selection resulted in a best fit model of 4 genes AUC = 0.83 and 3 genes AUC = 0.80 for SSc ILD and SSc respectively.

Whilst the AUC values were very good for the tests the publication did not validate its findings by qPCR making it impossible to determine if the biomarkers identified offer a legitimate diagnostic panel. Instead, the publication concluded that as the panels consisted of different genes then SSc with ILD expressed a genomic signature different to SSc. However, this could have been explicitly tested by using both sets of input genes as the input to a model with the presence of ILD as the regressor.

## Decision Trees

Decision trees are a non-parametric statistical learning algorithm which require minimal tuning and have the ability to discover patterns that present in complex interactions hidden within the data. Additionally, the resulting model can be displayed as an easily interpretable graph.

Decision trees are built from the top down, first the feature that best divides the data into the purest possible child nodes is identified. If the child nodes contain significant misclassifications (impurities) the child node(s) are split again using the best feature that divides the child nodes into the purest possible additional child nodes. This process is repeated until the misclassification of the terminal child node reaches a selected minimum (Breiman, Friedman, Olshen, & Stone, 1984).

Decision trees are a simple statistical learning model and their accuracy is usually bettered by more modern algorithms. However, these are less interpretable their ease of interpretation can outweigh their performance.

To identifying a peripheral blood signature for vasodilator-responsive PAH patients Hemnes et al., used a decision tree model with expression levels measured by qPCR from whole blood RNA (Hemnes et al., 2015). Vaso-responsive PAH (VR-PAH) accounts for a small a minority of patients who respond extremely well to calcium channel blockers and have improved survival (Malhotra et al., 2011). Vaso-reactivity is identified invasively during RHC by measuring a response in PAP when challenged with an acute vaso-dilator such as nitric oxide. In this study 25 key RNAs were identified by microarray analysis. Once validated by qPCR in a cohort of 8 VR-PAH and 13 PAH there were 13 significantly different genes. These were used for building the decision tree. Unfortunately, the predictive accuracy in this cohort was not reported, however, predictive accuracy of 90% was reported in an external cohort consisting of 5 VR-PAH and 6 PAH. ­

Despite the small numbers of patients this study has demonstrated that a biomarker for treatment response specifically calcium channel blockers can be identified. The high degree of concordance within the external cohort is encouraging suggesting that a valid biomarker has been identified with a decision tree.

## Random Forests

The random forest algorithm is an extension of decision trees, as with decision trees random forests require little tuning and transformations to the data are unnecessary. However, due to the way they are constructed it is not possible to have a clear visualisation of the model.

The Random Forests classifier is built on the principle that an ensemble of multiple weak learners can come together to form a strong learner (Breiman, 2001). A Random Forest model uses thousands of weak decision trees together to make the forest. To ensure the decision trees that make up the model are weak learners the tree depth is limited, often to just a depth of two splits. Each tree is built from a random subset of features and a random subset of samples, roughly two thirds of the data. Further randomness is introduced by sampling the records with repetition (duplicates within the random sample) this is termed bootstrapping. Bootstrapping has been shown to aid generalisation, additionally it further enhances the randomness in a Random Forest as the chance of a bootstrap containing data and features that have already been learned from is reduced (Fitzgerald et al., 2013). Finally, each tree then votes on the classification, the final classification is the one with the most votes within the forest.

Random Forests has been used in PH to provide multiclass differentiation between healthy volunteers, PAH and PH on exercise (ePH) (Sanders, Han, Urbina, Systrom, & Waxman, 2019). An age and sex matched cohort of 78 people, 26 in each group, had plasma taken from the PA at rest, and during cardio pulmonary exercise testing (CPET). The Metabolon platform was used to quantify levels of 1013 metabolites. Samples at rest and peak exercise were paired and Random Forest classification accuracy between control, ePH and PAH was extremely high at 96%, 90% and 88% respectively.

­

Whilst Random Forests has many strategies to reduce overfitting, without testing upon a separate, and ideally external cohort it is impossible to know if the accuracy seen would be reflected in future studies­. Additionally, as the risk of overfitting increases with increasing numbers of features using all 1013 metabolites in the model has likely resulted in an over fitted model and it would be impractical to measure 1013 metabolites in a diagnostic tes­t.

## Feature Selection

Feature selection is the term given to the process of selecting the most minimal set of variables or features on which to build a model on.

A key reason for feature selection is reducing overfitting. A model is said to have overfitted when it has identified patterns unique to the training data alone. Such a model will have a very high accuracy on the training set but a very poor predictive power when used on any new test data.

Building on a smaller data set reduces the risk of overfitting due to the model being built with less redundant or even misleading data, therefore stopping modelling decisions being made on noise. Additionally, an added advantage is a reduction in computation time enabling more iteration and model evolution.

## Filter Based Feature selection

Filter based feature selection applies a relevant threshold to determine which features should be kept. Whilst not exhaustive, data may be filtered based upon feature variance, magnitude above background, resulting p-value from a t-test, u-test or upon its univariate predictive power by ROC curve analysis. Each feature is assessed independently therefore no consideration is given to how features may combine additively, therefore soft thresholds are typically chosen and filtering performed as a prepressing step prior to more sophisticated feature selection or model building.

Filter methods offer a simple interpretable way to reduce the initial number of variables low quality or weak predictors that would otherwise act as noise.

## Forward selection and Backward Elimination

Forward selection and backward elimination are two similar methods of feature selection where interactions between features are considered. The technique can be applied to a large majority of models. In forward selection the model is iteratively built by addition of the next best feature until the optimal model is identified. Conversely, backwards elimination starts with all features in a model and iteratively removes the worst feature until the optimal model is reached. The optimal model is typically identified as that with the lowest Akaike Information Criterion (AIC), (Akaike, 1974). AIC compares the performance of two different models by subtracting a penalty for the number of features used from the variance explained by the model. This approach balances increasing model accuracy against its complexity, when the addition of a new feature provides minimal reduction in explained variance the AIC score becomes smaller than the previous smaller model. This ensures that the model contains the enough features to capture the true variability whilst keeping the number of features small enough to reduce the risk of overfitting.

As both methods determine the best model by iteratively testing they are very computationally intensive which makes them unsuitable for extremely large data sets.

## Least Absolute Shrinkage Selection Operator Feature Selection

The least absolute shrinkage selection operator (LASSO) is a generalised regression technique that performs both feature selection and regularisation (Tibshirani, 1996). Regularisation is a technique used to reduce overfitting and describes the aspect of a learning algorithm that discourages complexity. LASSO controls how large the coefficients grow, as the features become less predictive the penalisation increases until the coefficient is reduced to 0. A coefficient of 0 removes the feature from the model, this is the feature selection aspect of LASSO.

LASSO regression is computationally less intensive than step forward/backwards selection. The reduced feature set is likely more generalisable than step forward selection as the algorithm is tuned over cross-validation therefore reducing internal validation bias.

Additionally, correlating features are also removed improving model performance as correlating features negatively affect linear model performance. However, knowledge of features that correlate is useful when interpretation of biological function is sought. In this case selecting correlating features for removal according to prior knowledge or a filter criterion may be more appropriate.

## Boruta Feature Selection

Boruta, the Slavic God of the forest, is a feature selection algorithm that utilises random forests (Kursa & Rudnicki, 2010). In a standard random forest model importance metrics are calculated for each feature, there are various metrics available depends on random forest implementation or prediction type. However, the importance value is not constrained, whilst higher magnitude indicates a feature has greater importance within the model a cut-off for the features that are relevant cannot be deduced or defined. A feature can be determined as relevant by comparing its importance to a reference value. In Boruta the reference values are generated from the source data, termed a shadow feature. Shadow features are generated by randomly shuffling all data points, therefore obfuscating the data entirely. The shadow features are combined with the original data and classification is performed on this extended data. A feature is deemed relevant if its importance exceeds that of the best performing shadow feature in the model. The process is repeated 100 times to reduce any bias created by the randomisation this process. Features that have importance consistently higher than the best shadow feature are considered relevant.

As Boruta is based upon random forests in inherits advantages typically associated with random forests. Other selection techniques identify the non-redundant features, whereas Boruta identifies any relevant feature. This is particularly relevant in medical biology where one wishes to identify all of the features that are connected to the dependant variable.

## Performance Measures of Biomarkers

The evaluation of a biomarker’s performance is critical in order determine suitability for the classification task. Generally speaking, a good biomarker is one that is easily sampled and correctly identifies disease from non-diseased. Quantifying a biomarkers ability to discriminate disease can be evaluated by the metrics sensitivity (Se) and specificity (Sp). Sensitivity is the percentage of positive samples classified correctly out of all positive cases present and specificity is the percentage of negative cases correctly classified out of all the negative cases. When assessing the performance of a binary outcome for monovariate or multivariate biomarkers ROC curves and a confusion matrix are applicable.

## Receiver Operator Characteristic Curves

Originally developed during World War II to correctly interpret signal to noise ratio from primitive RADAR early warning stations ROC curves are commonly used to visualise and evaluate the performance of tests with a binary outcome

(Hanley, Radiology, 1982, 1982). A ROC curve plot shows the performance of a diagnostic test or biomarker across all possible cut-off values for classification by plotting the sensitivity by 1 - specificity. The performance of a ROC curve can be assessed by its area under the curve (AUC), this is a value between 0 and 1 with 1 indicating perfect classification and 0.5 being equal to random chance. When the cut-off for a biomarker is not known *a priori* it can be derived from the ROC curve. A cut-off at the Youden Index (*J*), calculated as , this gives the best balance between sensitivity and specificity and therefore maximal differentiating effectiveness of a biomarker (“Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection,” 2008).

## Confusion Matrix

A confusion matrix is a special form of contingency table breaking down the resulting biomarker classification results into two dimensions, the gold standard and predicted classification. Correct classifications fall into the first two classes: true positives (TP) and true negatives (TN). Incorrect classifications fall into the remaining classes: false positives (FP) and negatives (FN).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Gold Standard | |  |
|  |  | Positive | Negative |  |
| Test Outcome | Positive | True Positive  (TP) | False Positive  (FP) | Positive Predictive value  = |
| Negative | False Negative  (FN) | True Negative  (TN) | Negative Predictive value  = |
|  |  | Sensitivity  = | Specificity  = | Accuracy  = |

Table 1‑2 Standardised layout for confusion matrixes. Relevant performance calculations are shown in the lowest and furthermost right cells.

The confusion matrix in Table 1‑2 demonstrates the layout and details of the calculations for the previously discussed metrics sensitivity and specificity, and the additional measures; positive predictive value (PPV), negative predictive value (NPV) and accuracy. PPV and NPV are valuable performance measures and details the percentage of correctly identified positive or negative sample respectively. The measure accuracy is a singular metric of model performance it is the percentage of all correctly classified samples. However, this is only useful in situations where the data is balanced. For example, prostate cancer has an incidence of 171 in 100,000 therefore randomly sampling 100,000 men and classifying all as normal would likely give a resulting accuracy of 99.829%, in this case a completely irrelevant (Cooper & Linch, 2016). In this scenario reviewing sensitivity and specificity separately would provide better interpretation of the model fit.

## Hypothesis, Aims and Objectives

I hypothesise that there will be miRNA differences between the disease groups of IPAH + HV and SSc-PAH + SSc detectable from a non-invasive sample of blood or plasma. Secondly, I hypothesise that a diagnostic panel of multiple miRNAs in combination has better diagnostic potential than one miRNA in isolation. Finally, I hypothesise that the likelihood of successfully validating a miRNA panel will be increased by identifying and validating the multiple miRNAs derived from the unique miRNAs of several panels.

Project aims are:

1. To identify the best source of circulating miRNAs.
2. To identify miRNAs belonging to several diagnostic panels and validate by qPCR.
3. To demonstrate that miRNAs in a panel can provide improved diagnostic potential to a singular miRNA.

# Materials and Methods

This chapter details the materials and methods used in this thesis.

## Ethics, Patient Cohorts and Whole Blood Processing

### Ethical Approval

Patient and healthy volunteer plasma samples and clinical data from the Sheffield Teaching Hospitals Observational Study of Patients with Pulmonary Hypertension, Cardiovascular and Lung Disease (STH-Obs) biobank was provided under the local Sheffield Teaching Hospitals Foundation Trust Observational Cardiovascular Biobank Ethics (08/H1308/193 and STH 15222). The PH scientific advisory board approved sample release.

### Patient Cohorts

Patient samples were taken during the first diagnostic contact with the Pulmonary Vascular Disease unit. For healthy volunteers this was at their time of enrolment into the STH-Obs biobank.

### Total RNA Isolation from Whole Blood

Total whole blood RNA was extracted from Tempus RNA storage vacutubes (Applied Biosystems, U.S.A.) using the Maxwell miRNA Tissue kit on the Maxwell 16 automated purification platform (Promega, U.S.A.) according to manufacturer’s instructions. Briefly, the contents of Tempus tube were decanted into a 50ml Falcon tube and topped up with 1x sterile PBS (VWR, U.S.A.) to a total volume of 12.5 ml, vortexed (high, 1 minute) and centrifuged (4000 g, 30 minutes). The supernatant was discarded and the tubes placed upside down on paper towels to dry (1 minute). Homogenisation solution is prepared in advance by adding 20 µl of thio-1-glycerol to 200µl homogenisation buffer and chilling. 200µl of prepared homogenisation buffer was added to the tube and vortexed. Sequentially 200µl of lysis buffer and 15µl of proteinase k were added, gently vortexed and incubated (15 minutes, room temperature).

To prepare the Maxwell extraction system; 75µl of RNase/DNase free water was added to the elution tube and placed in the Maxwell 16 loading rack, along with an extraction cartridge. The foil was removed from the cartridge and blue DNase 1 added to the well 4 on the cartridge (coloured yellow), and clean RNase/DNase free magnet sheaths added to well 7 on the cartridge.

Following sample’s proteinase K incubation, the entire 400µl of solution in the 50ml Falcon was transferred to well 1 on the extraction cartridge and the loading rack placed into the Maxwell 16 ready for extraction. The miRNA program was selected, after 1 hour 23 minutes the eluted RNA was stored at -80ºC.

### Total RNA Isolation from Plasma

Total RNA was extracted from plasma using the total RNA slurry format extraction kit (Norgen Biotek Corp. Canada). Prior to experimental start solution Slurry C2 and Lysis Buffer A were prepared by incubation (60ºC, 20 minutes). To a 50ml Falcon tube 200 µl of Slurry C, 1800 µl of Lysis Buffer A, 20 µl of 2 M DTT and 500 µl of plasma were sequentially added. The solution was mixed by vortex (high, 15 seconds) and incubated (60ºC, 10 minutes). 3 mL of 100% ethanol was added and mixed by vortex for 15 seconds and centrifuged (30 seconds, 1000 rpm). Supernatant was discarded and 300 µl of Lysis Buffer A was added to the slurry pellet, mixed by vortex and incubated (60ºC, 10 minutes). 300 µl of 100% ethanol was added and the mixture and mixed by vortex for 15 seconds. Up to a maximum of 600 µl of solution was transferred to a mini filter spin column with collection tube and centrifuged (10,000 rpm, 1 minute) and the flow-through discarded. This was repeated with the remaining solution. The column was washed by addition of 400µl of Wash Solution A and centrifuged (10,000 rpm, 1 minute) the flow-through was discarded. This wash was repeated for a total of 3 washes. The column was dried by centrifugation (14,000 rpm, 3 minutes). The column was transferred to a new DNase/RNase free 1.5ml Eppendorf. RNA elution was performed by addition of 100 µl of Elution Solution A and centrifugation (2,000 rpm, 2 minutes followed by 14,000 rpm, 3 minutes). The column was discarded and RNA stored at -80ºC.

## Concentration of Plasma RNA

The initial 100 µl eluate of RNA was concentrated to 11.5 µl with the RNA Clean and Concentrate-5 kit (Zymo Research Cop, U.S.A.). 200 µl of RNA Binding Buffer was added to the 100 µl eluate and mixed by vortex. 300µl of 100% ethanol was added, mixed by pipetting and transferred to the spin column and centrifuged (10,000 rpm, 30 seconds). 400 µl of RNA Prep Buffer was added and centrifuged (10,000 rpm, 30 seconds). The column was then washed twice by addition of 700 µl and centrifugation (10,000 rpm, 30 seconds). Flow-through was discarded, a final wash was performed with 400 µl of wash buffer followed by centrifugation (10,000 rpm, 30 seconds). The column was dried by centrifugation (10,000 rpm, 1 minute). 11.5 µl of DNase/RNase free water was then added directly to the membrane and eluted by centrifugation (10,000 rpm, 1 minute). The column was discarded and RNA stored at -80ºC.

### Assessment of RNA Quality by Nanodrop

RNA quality was measured by spectrophotometry (Nanodrop 1000, Thermo Scientific, UK). The Nanodrop was blanked with 1.5 µl of DNase/RNase free water and 1.5 µl of concentrated plasma or whole blood RNA eluate was added. Whole blood RNA concentration was recorded and quality assessed by 260/230 and 260/280 ratios. In general, pure RNA has both 260 ratios at 2.0, however, samples within 1.8-2.2 can be considered to be high quality. It can be seen from Figure 2‑1 that the majority of samples fall within accepted quality ranges as the 260/280 is the most important of the two metrics when considering downstream applications, as this is a measure of purity and higher numbers show that the samples have little contamination with proteins or phenols, the 3 samples with 260/230 ratios above 2.2 are not of concern as this is commonly accepted to read above the 260/280 ratio and low numbers indicate contamination with carbohydrates.

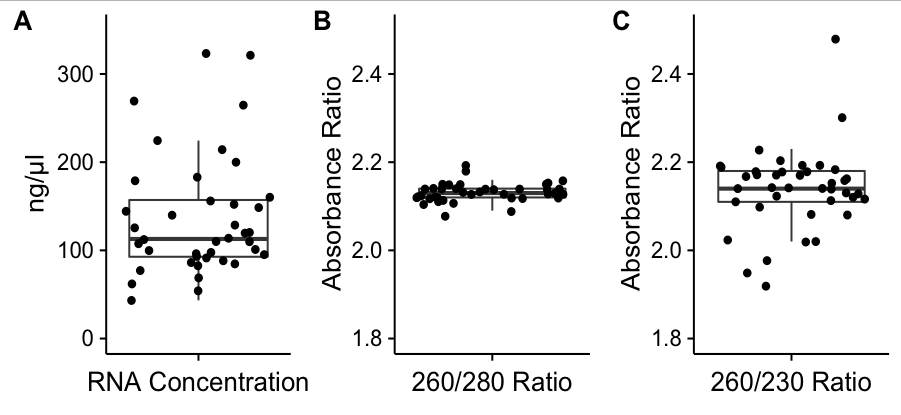


Figure 2‑1 Metrics determined from Nanodrop data A) RNA concentration B) 260/280 ratio B) C)260/230 ratio

The concentration of RNA from plasma is undetectable by absorbance spectrophotometry. However, during initial experiments it was noticed that the majority of samples had a characteristic curve where a high absorbance would quickly drop close to baseline, arrays from these samples did not fail. A portion of samples had absorbance spectras with varying degrees of slope; these samples did not always have successful arrays. These slopes potentially indicating the presence of impurities, samples with a ‘high slope’ were highly likely to fail and samples with a ‘low slope’ were less likely. To retain more samples for testing only ‘high slope’ samples were removed from further testing. Examples of these traces are shown in Figure 2‑2

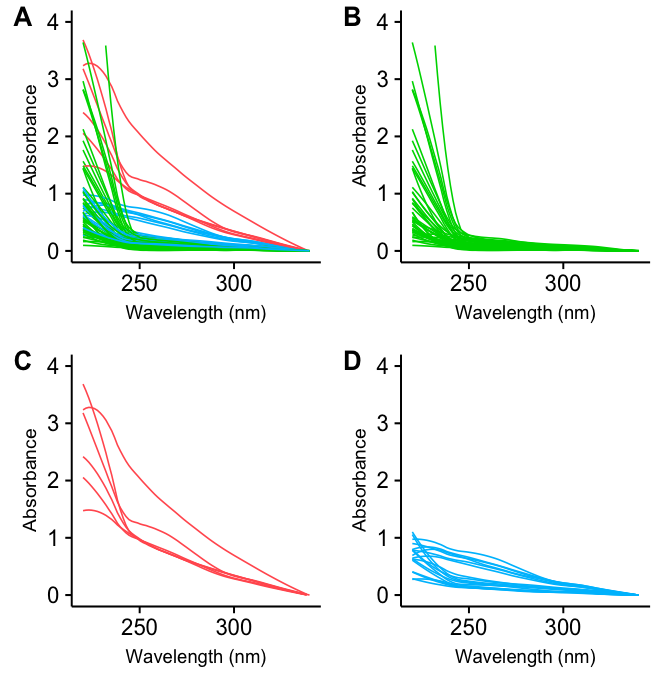


Figure 2‑2 Nanodrop absorbance spectra from plasma RNA samples showing the overall (A), high purity samples (B), high sloping samples (C) and low sloping samples (D).

### Global microRNA Profiling; Array Processing, Quality Control, Data Analysis

### Agilent microRNA Array Processing

Global miRNA expression was investigated using Agilent single colour miRNA arrays miRbase v.19 (Agilent Technologies, UK) according to manufacturer’s instructions. To adjust for the low concentration of RNA in plasma samples the protocol was modified accordingly; the 1.3 µl of spike in control was not added, instead an additional 1.3 µl of RNA was added. This gave a total of 3.3 µl of RNA were labelled and hybridised, instead of the standard 2 µl of RNA. The spike in controls aid troubleshooting a failed hybridization reaction and have no impact on downstream analysis. Furthermore, a successful hybridisation can be determined through various metrics discussed in 2.2.2. Whole blood RNA was diluted to 50 g/µl and 2 µl was labelled in accordance with manufacturer’s instructions. Completed microarrays were scanned at 2 µm on an Agilent Technologies Scanner G2505C and spot intensities determined using Feature Extraction Software v12.0 (Agilent Technologies, UK)

### Agilent microRNA Microarray Pre-processing and Quality Control

Data analysis was performed in the R statistical programming environment (http://www.r-project.org/). Boxplots summarising the intensity distribution for each array were plotted and arrays with abnormally high or low signals were removed from further analysis Figure 2‑3.

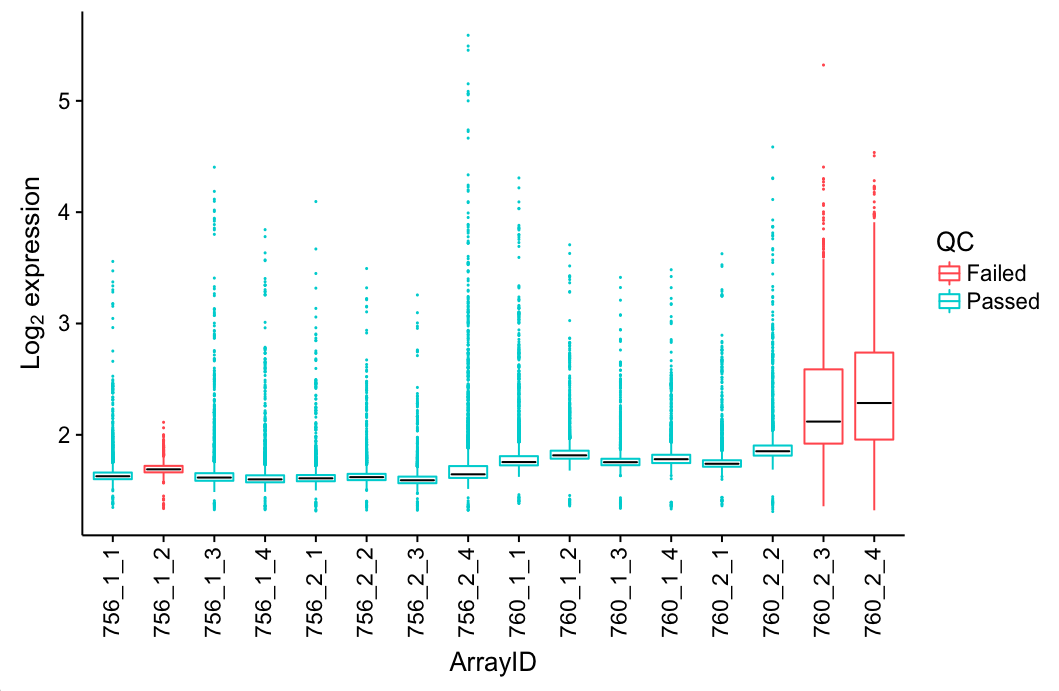


Figure 2‑3 Boxplots demonstrating that arrays failing either due to under or over hybridization have clearly different data distributions that can be easily identified.

Data was normalised using the robust multi-array (RMA) method from the R package affy (Gautier et al., 2004). Briefly, RMA is a two-step process where arrays are first quantile normalised minimising inter-array variation followed by a median-polish to remove any intra-array spatial variations. The package AgiMicroRna (López-Romero, 2011) was used to determine background expression from negative controls features and subtracted from all features. Preliminary filtering was performed using the default settings in the function *filterMicroRna()* this removed miRs with a low signal intensity, defined as requiring an expression level exceeding 1.5x the standard deviation of the array negative controls. Additionally, miRs with expression detectable above the negative controls that were only present in 10% or less of samples were also filtered out. Following filtering 393 plasma miRs and 352 whole blood miRs for downstream analysis. Additionally, miRs were then limited to those either qPCR wet-lab validated by Exiqon or deemed high confidence miRs in miRBase (v21.0). The pre-processing limits were in place so that subsequent analysis was performed upon miRs that would likely be present in the majority of samples and have good overall expression levels. Removing these left 210 plasma miRs and 229 whole blood miRs that were used in all downstream analysis. It should be noted that plasma and whole blood microarrays were normalised and processed separately.

For each set of analyses performed, subsets of data were then created containing only samples relevant to the specific classification problem being tested Table 2‑1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Comparison | Healthy Volunteer | IPAH | SSc with PAH | SSc without PAH |
| IPAH from HV | 33 | 46 | . | . |
| SSc-PAH from SSc | . | . | 18 | 10 |

Table 2‑1 Numbers of samples included in each biomarker discovery cohort.

### Differential Expression

Differential expression was performed in R using the package Linear Models for Microarray Analysis (LIMMA) (Ritchie et al., 2015).

Furthermore, to remove miRs with extremely low expression miRs with upper quartile expression <3 were filtered out so that the majority of miRNAs had detectable expression levels.

### Identification of Biomarkers

### Filtering

Identification of biomarkers was performed by a multistep approach. In order to minimize any misleading biomarkers from over-fitting, the number of potential targets was reduced by filtering and feature selection as discussed in 1.7.2. Initial filtering by ROC AUC and Spearman correlation was performed followed by feature selection and finally biomarker panel building.

### Receiver Operator Characteristics

The diagnostic potential of each individual miR was determined by ROC analysis using the R package OptimalCutpoints (Lopez-Raton et al., 2014). miRs with a ROC AUC <0.65 were removed from further analysis.

### Identification of Correlating miRNAs

Co-linear miRs were identified with an absolute Spearman correlation coefficient >0.75. Of the correlated miRs, wet-lab validated miRs were prioritised to be retained. However, if further separation was required then the miR with the largest AUC was retained for final feature selection.

### Feature Selection

Final feature selection was performed on the filtered data using Boruta and, where sufficient sample numbers allowed, the LASSO feature selection algorithms. Boruta features marked as “confirmed” were retained and from LASSO features with a coefficient was >0.00 were retained.

### Combining Features into Panels

miRs were constructed into panels based upon their univariate diagnostic thresholds as determined by ROC curve analysis. For each miR an expression threshold was determined at the point of maximum sensitivity and specificity. A patient score was generated as the number of miRs within the panel exceeding this threshold.

To find the best diagnostic panel all selected miRs were combined into exhaustive possible panel combinations containing between 2-5 miRs. Panels were not built greater than 5 miRs due to the increased computation time required, this made larger panels impractical to investigate. Panels with a sensitivity and specificity of >0.8 were subjected to k-fold cross-validation repeated 10 times. The value of k was selected as appropriate for the size of each of the cohorts tested, for HV and IPAH a k of 10 was used and for the smaller SSc cohort a k of 3 was used. The resulting panels were assessed based on their sensitivity, specificity, positive predictive value and negative predictive value.

### Quantification of miRNAs by qPCR

Quantification of target miRNAs levels was performed using the Exiqon (Denmark) miRCURY LNA SYBR Universal RT PCR system in a multi-step protocol detailed below.

### Complementary DNA Synthesis

MicroRNAs were reverse transcribed to complementary DNA (cDNA) with the Universal cDNA Snythesis Kit II (Exiqon, Denmark). On ice, cDNA master mix was prepared according to Table 2‑2 and 8µl added to 2µl of RNA in a 0.2µl microfuge tube, mixed gently followed by brief centrifugation. The reaction was thermal cycled (Verity Thermal Cycler, Life Technologies UK) at 42ºC for 60 minutes, followed by 95ºC for 5 minutes and immediately chilled and held at 4ºC. An initial 1:8 stock dilution was performed by addition of 70µl of nuclease free water before being stored at -80ºC.

|  |  |  |  |
| --- | --- | --- | --- |
| **Reagent** | **Volume (µl)** | **x5 (10% overage)** | **x10 (10% overage)** |
| 5x Reaction Buffer | 2 | 11 | 110 |
| Water | 4.5 | 24.75 | 49.5 |
| Enzyme mix | 1 | 5.5 | 11 |
| RNA spike in | 0.5 | 2.75 | 5.5 |
| Total volume (µl) | 8 | 44 | 88 |

Table 2‑2 Universal cDNA master mix with example calculations for 5 and 10 reactions.

### Quantification of miRNAs by Real-Time PCR

Upon first use lyophilised LNA primers were reconstituted by addition of 20µl of nuclease free water and incubation on ice for 30 minutes prior to first use.

A working dilution of cDNA was made by diluting the stock cDNA 1:10 with nuclease free water as appropriate for the number of qPCR reactions required, for example an experiment of investigating 5 miRNAs ran in duplicate requires 4.5µl cDNA per reaction x 2 (duplicate) x 5 (miRs) + 10% = 49.5 total working dilution. Therefore, 5µl of initial dilution would be added to 45µl of nuclease free water for a total of 50µl of cDNA at working dilution.

Each miRNA investigated required an individual real-time PCR master mix. These were made according to Table 2‑3, gently vortexed and centrifuged and stored on ice and shielded from light.

|  |  |  |  |
| --- | --- | --- | --- |
| **Reagent** | **Volume (µl)** | **10 rxns + 10% overage (µl)** | **50 rxns + 10% overage (µl)** |
| LNA primer Set | 0.5 | 5.5 | 27.5 |
| 2x ExiLENT SYBR Green Master Mix | 5 | 550 | 275 |
| Total volume (µl) | 5.5 | 555.5 | 302.5 |

Table 2‑3 Real-time PCR reaction master mix with examples for 10 and 50 reactions including 10% overage

4.5µl of cDNA at working dilution was added to 5.5µl of miRNA specific real-time PCR master mix in duplicate to a MicroAmp Optical 384-well reaction plate (Applied Biosystems, UK) and sealed with MicroAmp Optical Adhesive Film (Applied Biosystems, UK), mixed and briefly spun to ensure the solution was at the bottom each well.

The real-time PCR reaction was performed using an ABI 7900HT Fast Real-Time PCR System (Applied Biosystems, CA) using the following conditions; initial denaturation at 95ºC for 10 minutes followed by 40 cycles of 95ºC for 10 seconds, 60ºC for 1 minute. After the final cycle a melt curve is performed, samples are denatured at 95ºC for 5 minutes then cooled at a rate of 1ºC per minute and the fluorescence recorded at each decreasing temperature.

### Data Analysis

Quality control and data analysis was performed in R using tools from the packages ggplot2, reshape2, and tidyverse.

### Quality control

Non-template controls (NTC) were examined for any amplification. Amplification within the NTC is indicative of contamination or primer-primer amplification that may mask true expression values by artificially inflating them. As a guide NTCs with amplification within 5 cycles have little to no effect on the true expression value. Only miRNAs with clear NTC’s or those >5 cycles from any NTC were used.

Melt curves were analysed for the presence of more than one peak as an indicator of undesired amplification. Example melt curves are shown in 

Figure 2‑4 where plot A and B show acceptable melt curves, C shows low amplitude products with varying maxima and D shows a peak to the left indicating the presence of secondary undesired amplification products.



Figure 2‑4 Example dissociation curves for 4 miRs of varying quality. Individual qPCR reactions are shown in grey whilst the blue line is a loess best fit.

# Determining the Best Source of Circulating Biomarkers of IPAH

### Aims

The Sheffield Teaching Hospitals Observational Study into patients with Pulmonary Hypertension, Cardiovascular disease and Lung disease (STH-Obs) collects and preserves various samples from patients and healthy controls, including plasma, serum, whole blood, urine and DNA. This chapter aims to determine whether whole blood or a separated fraction, plasma, is the best source for disease relevant biomarkers.

### Introduction

Blood based biomarkers offer advantages over current detection methods for PAH. Phlebotomy is easily performed at medical centres and sample preparation is typically straightforward. Tempus RNA collection tubes are advantageous as the RNA profile is preserved near-instantaneously during collection due to the action of the lysis and preservation agent present in the collection tubes. Tempus RNA can be extracted robotically offering simpler processing and minimising variation due to sample preparation. However, whole blood contains the miRNA from variety of leukocytes which may become a source of noise, masking disease specific signatures. Plasma is a source of miRNA with a potentially cleaner signal, however, RNA extraction yields are very low and sample preparation both following collection and for RNA extraction are more involved.

### Sample Cohort

Patient cohort for determining between plasma and tempus samples is detail below Table 3‑1. The patients are well matched, with non-significance, in age, gender and ethnicity. IPAH patients are shown to have PAH from the appropriate values in the catheter based diagnostic measures.

|  |  |  |  |
| --- | --- | --- | --- |
|  | HV | IPAH | p-value |
| **N** | 19 | 28 |  |
| **Age** | 47.4 (35.2 - 55.8) | 53.1 (46.7 - 65.1) | NS |
| **Gender (% Female)** | 58% | 57% | NS |
| **Ethnicity (% Caucasian)** | 100% | 100% | NS |
| **mPAP (mmHg)** | - | 57.3 (10.8) | - |
| **PVR (Dynes.s/cm5)** | - | 847.5 (421.2) | - |
| **Cardiac Output (L/min)** | - | 5.2 (1.7) | - |
| **PA Wedge Pressure (mmHg)** | - | 10.6 (3.4) | - |

Table 3‑1 Summary demographics of IPAH and HV samples used in the sample source cohort. Sample distributions were tested using Shapiro-Wilk test, samples with a p-value <= 0.05 were compared with the Mann Whitney U-test whilst a p-value > 0.05 was compared using a Students t-test. Categorical data was testing with Chi-Squared.

### Expression Changes in Blood Fractions

Differential expression of miRNA between HV and IPAH in whole blood and plasma fractions was determined using the Bioconductor packages LIMMA and AgiMicroRAN as described in 2.3. 160 plasma miRNAs and 179 whole blood miRNAs were detected following initial filtering by the function filterMicroRna(). This was used at the default settings removing miRNAs with a low signal intensity, defined as probes with an expression level less than 1.5x the standard deviation of the array negative controls, and miRNAs present in 10% or more of samples. The null hypothesis tested by the differential expression analysis is “that there are no differences in expression between HV and IPAH”. Ultimately, the results of both differential expression analyses will be used to test the null hypothesis of “there are no differences in abundance of miRNAs between the circulating fractions of plasma and whole blood”. It will determine any differences between the magnitude and group variability of miRNA expression. The number of miRNAs over both unadjusted p-value and adjusted p-value of 0.05 is summarised in Table 3‑2. Differences between the two sources are visualised in the volcano plot in Figure 3‑1.

|  |  |  |
| --- | --- | --- |
| miRNA Source | p-value <0.05 | fdr adjusted p-value <0.05 |
| Whole Blood | 45 | 1 |
| Plasma | 95 | 90 |

Table 3‑2 Number of miRNAs with a p-value <0.05 both with and without adjusting for multiple testing Benjamini-Hochberg false discovery rate.



Figure 3‑1 Volcano plot visualising the differences between miRNA expression and p-value for both plasma (red) and whole blood (green) miRNA microarrays. Whole blood miRNAS with an adjusted p-value ≤ 0.05 are labelled. For clarity in plasma miRNAS with a log fold change ≥ 1.0 and an adjusted p-value ≤ 0.05 are labelled.

### Determining the Best Source of miRNAs for Classification

The matched whole blood and plasma miRNA microarray expression data was pre-processed as detailed in 2.3 and combined into one singular dataset so that each patient had expression values of both plasma and whole blood miRNAs present simultaneously. Feature selection using the R Boruta package was used to determine the number of miRNAs from each fraction that were significantly important in classification of HV from IPAH. The number of features with a significant variable importance in the model and their sample source are shown in Figure 3‑2.

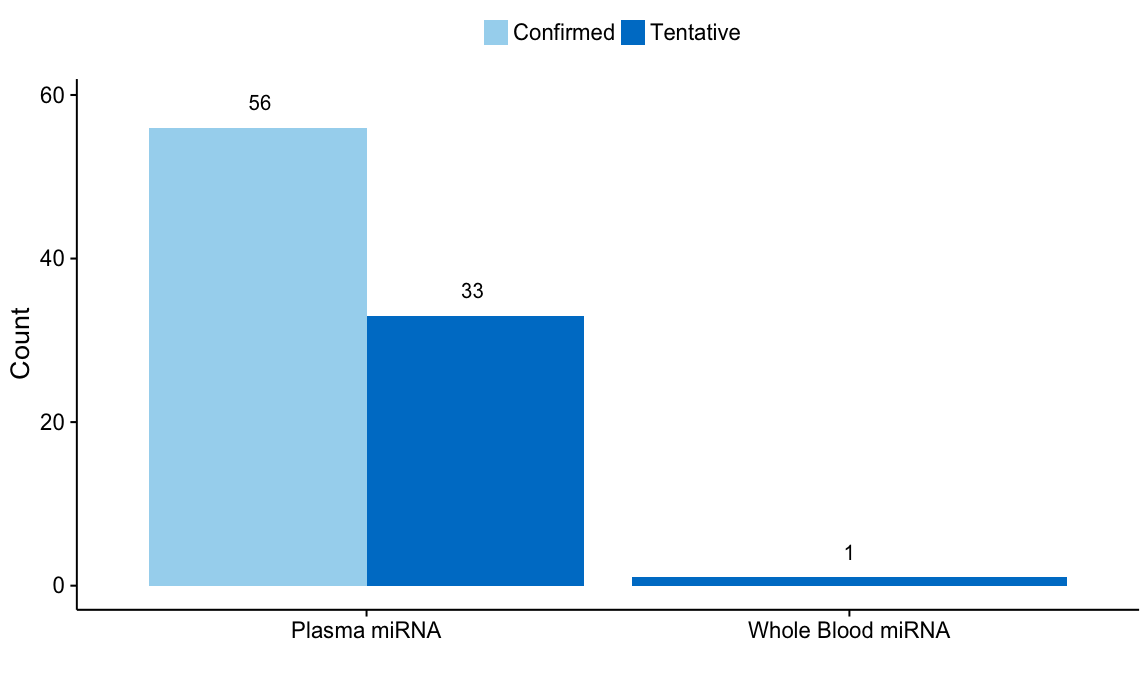


Figure 3‑2 Number of miRNAs confirmed (light blue) or tentatively (dark blue) identified as important in the classification between IPAH and HV in the Boruta feature selection model.

### Variation of miRNA Expression Between Whole Blood and Plasma

To interpret the variation between measured miRNAs the standard deviation for each miRNA was calculated within each group. This was done so that any groupwise differences didn’t affect the assessment of variation. Standard deviation was chosen as this demonstrates the distribution of the data in the units measured whereas variance is in the units-squared which makes interpretation more difficult.

The density plot, Figure 3‑3, shows an overview of the distribution of values of measured miRNAs within disease and sample groups. Overall, this shows little difference in variation between patient groups. However, it can be seen that there are more miRNAs with a greater variation in the plasma microarrays, seen as the higher density at higher standard deviations.



Figure 3‑3 Density distribution plots of the standard deviation of measured miRNAs for IPAH and HV groups and sample types.

There are 16 miRNAs with a standard deviation greater than 1 in the plasma samples compared to 0 in the whole blood sample, these are shown in more detail in Figure 3‑4.



Figure 3‑4 Bar plot showing the 16 miRNAs from the plasma microarray with a standard deviation ≥ 1.00.

### Conclusion

From the analyses preformed it was seen that plasma miRNAs provide a richer source of information. In a comparison between IPAH and HVs, plasma miRNAs yielded 90 significantly (BH adjusted p-value <0.05) differentially expressed miRNAs to 1 from whole blood. In an all-relevant feature selection task resulted in the selection of 56 miRNAs from plasma and 0 from whole blood, indicating that the plasma miRNAs offered the better source of input data for classification of IPAH from HVs.

### Variability of miRNAs

Analysis of the variability of measured miRNAs was performed retrospectively after plasma RNA had been chosen and used for further experiments. This wasn’t performed initially due to the link between p-value and standard deviation/variance. The assumption being that a p-value is indicative of variation. This test appears to contradict the findings that plasma is the better source. Whilst whole blood provided a smaller number if miRNAs of interest in the small sample cohort it is possible that in a larger cohort a greater number of differences could have been seen and should have been explored further.

Whilst the differential expression results indicate the best source of miRNAs to be from plasma RNA due to the large number of statistically significant miRNAs between IPAH and HV it does not necessarily mean that plasma miRNAs provide the best combined source of information to perform disease classification. Boruta is machine learning algorithm performing feature selection for classification problems, robustly identifying the features that in combination significantly add accuracy to a classification model. Having simultaneously entered both whole blood and plasma miRNAs expression levels it was determined that in plasma there were 56 significant (confirmed) and 33 border-line significant (Boruta classification; tentative) and only 1 borderline significant whole blood miR. From these results it was concluded that plasma is the best source of miRs, further experiments were performed on plasma RNA.

### Deconvolution of miRNA Expression Profiles

Whole blood consists of multiple different cell populations all with their own unique expression profiles. The very low number of significantly different miRNAs may be attributable to this heterogeneity. Correcting for the cell counts for each sample may have led to more significant miRNAs being detected in the differential expression analysis. Unfortunately, cell counts were not performed on samples in the STH-Obs biobank so this adjustment cannot be directly made. However, it may have been possible to estimate the contribution of RNA from different cell types within the sample and thereby the cell types present using deconvolution techniques. These decompose the bulk gene expression profile into the sum of the contributions of the different cell types that are in that mixture. Several approaches for gene expression deconvolution exist, least squares regression (Abbas, Wolslegel, Seshasayee, Modrusan, & Clark, 2009), quadratic programming (Gong et al., 2011) and v-support vector regression (CIBERSORT) (Newman et al., 2015).

Whilst deconvolution has historically been applied to gene expression experiments Srinivasan et al., recently demonstrated that deconvolution is possible with miRNA (Srinivasan et al., 2020a). Whilst the source of the tissue specific miRNA expression reference levels is not clear they were able to use this with the CIBERSORT package to identify miRNAs associated with the placenta in the serum of maternal mothers with and without pre-eclampsia. This demonstrates that it is theoretically possible, given an appropriate reference expression dataset, to quantify leukocyte proportions within the whole blood miRNA arrays. These proportions could have been used within the linear regression framework of LIMMA to adjust for any effect related to the heterogeneity of the blood samples. Accounting for this variability may have enabled a larger number of differentially expressed miRNAs to have been identified. However, this work was published long after the decision to use plasma miRNAs was made, therefore these methods could not be tested.

### Differentially Expressed miRNAs

### Whole Blood

The only significant miRNA in whole blood, miR-199a-5p1, was not found to be expressed in plasma. Therefore, there was no detectable similarity in differentially expressed miRNAs between whole blood and plasma.

### miRNA199a-5p

In the literature reduced levels of miRNA 199a-5p have been shown in lung cancer (A. A, B, S, M, & SJ, 2017a) and over expression by lentivirus vector suppressed non-small lung cell cancer growth by targeting MAP3K11 (Yanli Li et al., n.d.). In PAH activation of mitogen-activated protein kinases stimulate the hallmarks of disease, apoptosis, inflammation and fibrosis and inhibition of this pathway can reduce mPAP in the MCT rat model of PAH (Budas et al., 2018). The detected decrease of miR-199a-5p and the links to lung cancer and a known PAH pathway may potentially show MAPK activation in a circulating biomarker. However, it is important to note that targets for further investigations from microarray experiments use cut-offs of at least an adjusted p-value of ≤0.05 and an absolute log-fold change of between 1 – 2. Mir-199a-5p was had a log fold-change of -0.17 this is below the threshold that typically warrants further investigations.

### Plasma

### miRNA199a-5p

In the plasma microarray miR-630 had the largest fold change. In the literature reduced plasma exosomal levels of circulating miR-630 were detected in obese children with obstructive sleep apnea (Khalyfa et al., 2016). The publication noted that these children’s cardiovascular systems had endothelial dysfunction including reduced endothelial nitric oxide (eNOS). In PAH reduced endothelial expression of eNOS is seen in pulmonary arteries of patients (Chester, Yacoub, & Moncada, 2017). In the study artificially increasing miR-630 increased the levels eNOS detected. Whilst at odds with the direction found in the array it is possible that the increased plasma miR-630 in PAH may be part of a feedback loop attempting to upregulate eNOS production. However, in lung adenocarcinoma it was found that increased miR-630 reduced tumour metastasis by decreasing expression of SLUG providing anti-apoptotic activity {Kuo:2013by. Pulmonary artery smooth muscle cells (PASMCs) in PAH exhibit a pro-proliferative and anti-apoptotic profile, if the source of the increased plasma miR-630 could be attributed to production in PASMCs this could be both a biomarker of disease and extent. Here we have to opposing mechanisms where in the context of PAH high or low expression of miR-630 could potentially beneficial or detrimental, further investigation would be required to determine which, if any are valid.

### Remaining differentially expressed miRNAs

The plasma miRNAs exceeding the typical cut-off of absolute fold change ≥ 1.5 & adjusted p-value ≤ 0.05 are detailed in Table 3‑3.

|  |  |  |  |
| --- | --- | --- | --- |
| miRNA | Average Expression | logFC | Adjusted p-value |
| miR-630 | 7.4 | 2.3 | < 0.0001 |
| miR-320c | 4.3 | 2.3 | < 0.0001 |
| miR-642a-3p | 4.3 | 2.0 | < 0.0001 |
| miR-1207-5p | 5.2 | 1.8 | < 0.0001 |
| miR-940 | 7.5 | -1.7 | < 0.0001 |
| miR-485-3p | 4.6 | -1.5 | < 0.0001 |

Table 3‑3 miRNAs significantly differentially expressed in RNA from plasma comparing IPAH vs HV (absolute log fold-change ≥ 1.5 & adjust p-value ≤ 0.05).

Each miRNA, except the previously discussed miR-630, was investigated in PubMed for associations with PAH using the search string, the miRNA of interest replacing the parameter place holder <mirna\_of\_interest> Figure 3‑5. However, none of the miRNAs had any previous association with PAH, with all but miR-642a-3p returning 0 results. MiR-642a-3p returned 257 results, however, no abstracts were found to mention miR-642a-3p upon screening the first 30 abstracts as ordered by the PubMed “Best Match” algorithm. Concluding that this is likely a quirk of the PubMed search and there has been no relevant publications in PAH for miR-642a-3p.

A picture containing text

Description automatically generated

Figure 3‑5 Search string template used to identify specific miRNA related PAH publications.

### Platelet miRNA Contamination in Plasma

Contamination with platelet derived miRNAs is a concern when investigating plasma miRNAs. In a case-control study Mitchel et al., identified significant correlation of miRNA plasma expression to platelet count in archival plasma (Mitchell et al., 2016). In a qPCR array identifying miRNAs associated with peripheral artery disease the 5 most differentially expressed miRNAs were investigated for their relationship to platelets. High correlation amongst themselves was seen, suggestive of a common cellular origin and all 5 had highly significant correlation to platelet count. These findings together this suggest that the miRNAs identified were related to platelet contamination.

Whilst correlations with platelet count could not be made in our samples, as platelet counts were not recorded, the 5 miRNAs in Mitchel et al., can be looked at. None of the 5 miRNAs were present in the expression set following filtering. However, in the unfiltered results, the low expression of the 5 miRNAs can be seen, the two most differentially expressed miRNAs miR-630 and miR-320c are also shown to provide context Figure 3‑6. This is not conclusive evidence of a lack of platelet derived miRNA contamination, however, it shows that the 5 platelet correlated miRNAs in peripheral artery disease were not found to be expressed in the plasma miRNA microarrays in this study.



Figure 3‑6 Comparison of miRNA microarray relative expression of platelet related miRNAs in Mitchel et al., to the two most highly differentially expressed miRNAs in HV vs IPAH plasma samples.

It should be noted that Mitchel et al., stated that in order to achieve a platelet poor plasma (PPP) sample before freezing samples must be centrifuged twice. Once to remove most of the cellular matter and the second to reduce the number of residual platelets. Unfortunately, the STH-Obs protocol only has 1 centrifugation step which could lead to unwanted plasma miRNAs in our plasma samples. This may mean our samples were not optimally processed and if the data were available it would be even more prudent to identify miRNAs with a high correlation to platelet levels. In the absence of a direct measurement deconvolution of the plasma results like Srinivasan et al., may give an indication of the presence of miRNAs coming from platelets.

# Identification of Biomarker Panels of IPAH

## Introduction

Diagnosis of IPAH is a long process requiring invasive procedures such as right heart catheterisation (RHC) at a specialist treatment Centre to confirm diagnosis. A panel of biomarkers could potentially be used as a screening tool in suspected cases of IPAH, reducing unnecessary catheterisation. Detailed in this chapter are the demographics of patient groups, experiments and analysis performed to identify circulating biomarkers that identify IPAH from HV.

## Methods

Human plasma in citrate buffer was acquired from the Sheffield Pulmonary Hypertension Biobank and from the Imperial College London Pulmonary Hypertension Biobank Table 4‑1. RNA was extracted as described in 2.1 and miRNA microarray performed as described in 2.2.

|  |  |  |
| --- | --- | --- |
| Sample Source | Healthy Volunteer | IPAH |
| Sheffield | 21 | 34 |
| Imperial College | 12 | 12 |
| Total | 33 | 47 |

Table 4‑1Summary of patient and control numbers from Sheffield and Imperial College in the IPAH biomarker identification cohort.

Patient demographics in *Table 4‑2* show patients well matched, with non-significance, in age, gender and ethnicity. IPAH patients are shown to have PAH from the appropriate values in the catheter based diagnostic measures.

|  |  |  |  |
| --- | --- | --- | --- |
|  | HV | IPAH | p-value |
| **N** | 21 | 34 | - |
| **Age** | 47.4 (35.6 - 60) | 53.7 (44.8 - 65.7) | NS |
| **Gender (% Female)** | 57% | 59% | NS |
| **Ethnicity (% Caucasian)** | 100% | 85% | NS |
| **mPAP (mmHg)** |  | 58.7 (12.3) | - |
| **PVR (Dynes.s/cm5)** |  | 919.8 (463.7) | - |
| **Cardiac Output (L/min)** |  | 5 (1.7) | - |
| **PA Wedge Pressure (mmHg)** |  | 10.3 (3.2) | - |

*Table 4‑2 Summary demographics of the Sheffield IPAH biomarker identification cohort. Normally distributed data shown as mean + standard deviation and comparisons made using Students t-test. Non-normally distributed data shown as median + interquartile range and comparisons made using Mann Whitney U-test. Categorical data was tested with Chi-Squared.*

## IPAH from HV Biomarker Feature Selection

Identification of biomarkers of IPAH was performed on the data subset containing IPAH and HV created from the whole miRNA microarray data set of 210 miRNAs. Analysis was performed as described in 2.2 and 2.4. 180 miRNAs with a ROC AUC <0.65 were identified and removed from the initial 393 miRNAs. Spearman correlation analysis on the remaining miRNAs identified 198 correlating miRNAs that were removed leaving a filtered data set containing 15 miRNAs for feature selection. Feature selection was performed with both Boruta and LASSO algorithms identifying a combined total of 10 unique miRs; All 10 were identified in Boruta and 9 by LASSO (Table 4‑3).

|  |  |  |
| --- | --- | --- |
| microRNA | Boruta Feature Significance (%) | LASSO Rank |
| miR-187-5p | 100 | 6 |
| miR-331-3p | 91.9 | . |
| miR-451a | 98 | 5 |
| miR-630 | 79.8 | 3 |
| miR-636 | 100 | 1 |
| miR-671-5p | 94.9 | 4 |
| miR-877-3p | 100 | 7 |
| miR-1229-3p | 85.9 | 8 |
| miR-1237-3p | 69.7 | 9 |
| miR-1972 | 98 | 2 |

Table 4‑3 Biomarkers identified by both the Boruta and LASSO algorithms.

## Combining Biomarkers with Decision Trees

An Rpart decision tree model was used to discern a panel of biomarkers. The model was built using only the features selected in section 4.3. To minimise over-fitting 5-fold cross-validation was employed. The resulting decision tree is shown in Figure 4‑1 and the diagnostic power is summarised in the metrics detailed in Table 4‑4. Note, unlike the ROC based panels the standard deviation of performance metrics across the cross-validations is not available from the Caret cross-validate implementation of the Rpart algorithm.

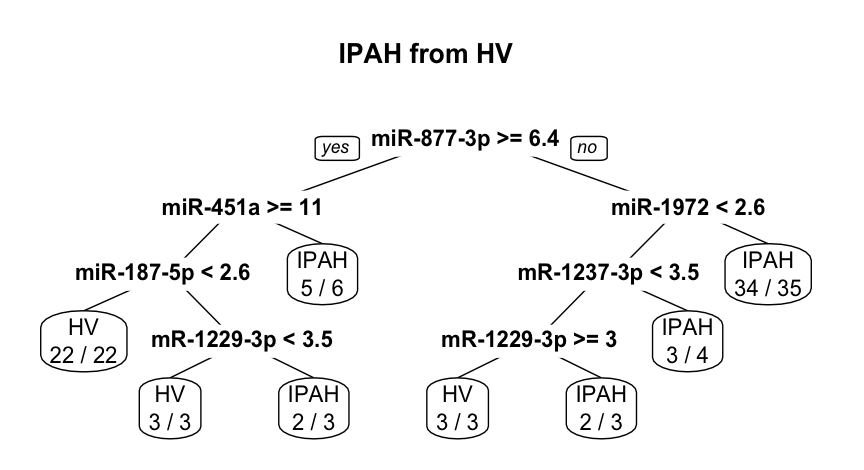


Figure 4‑1 Rpart decision tree for separating IPAH from HV. Each terminal node represents the assigned classification with the fraction of correctly classified samples given below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | ROC  AUC | Sensitivity | Specificity | Positive Predictive Value | Negative Predictive Value |
| Rpart DecisionTree | 0.97 | 1.00 | 0.84 | 0.90 | 1.00 |

Table 4‑4 Rpart decision tree performance metrics summarising the diagnostic power for the classification of IPAH from HV.

## Combining Biomarkers by ROC Determined Threshold

Using combined ROC thresholds as described in 2.3 biomarker panels were built using the 11 miRNAs from section 4.3, choosing a panel size containing from 2 to 5 miRNAs a total of 1102 combinations were possible. When tested, 139 panels had a sensitivity and specificity greater than 0.8 and were subjected to 10 repeats of k fold cross-validation (k=10). Further filtering and ordering of the cross-validated panels described in 2.4.5 reduced the final list of panels to 79, the top 5 are shown in *Table 4‑5*. The small standard deviation of the performance metrics shown is indicative of stability within the performance of the panel.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| miRNAs in Panel | Se | Sp | Positive Predictive Value | Negative Predictive Value |
| miR-331-3p, miR-451a, miR-630,  miR-636, miR-877-3p | 0.91  (0.014) | 0.82  (0.010) | 0.87  (0.007) | 0.87  (0.018) |
| miR-451a, miR-630, miR-636, miR-671-5p, miR-877-3p | 0.89  (0.010) | 0.85 (0.000) | 0.89  (0.001) | 0.85  (0.012) |
| miR-1972, miR-451a, miR-636,  miR-671-5p, miR-877-3p | 0.89  (0.014) | 0.87(0.016) | 0.90 (0.011) | 0.85  (0.018) |
| miR-1237-3p, miR-451a, miR-636  miR-671-5p, miR-877-3p | 0.88  (0.011) | 0.85 (0.010) | 0.89  (0.006) | 0.84  (0.013) |
| miR-187-5p, miR-1972, miR-451a,  miR-636, miR-877-3p | 0.88  (0.011) | 0.84  (0.020) | 0.88  (0.013) | 0.83  (0.014) |

*Table 4‑5 Top 5 miRNA panels for diagnosing IPAH from HV and metrics for their diagnostic power. Performance metrics are shown as the mean and, in brackets, standard deviation from the cross-validation.*

## Comparison of Biomarker Performance

Performance of the biomarker panels was compared with ROC curves. Additionally, individual miRNA ROC curves are plotted to demonstrate any additional benefit of combining miRNAs into a panel. NTproBNP, as a clinically relevant biomarker, has been plotted for comparison Figure 4‑2.

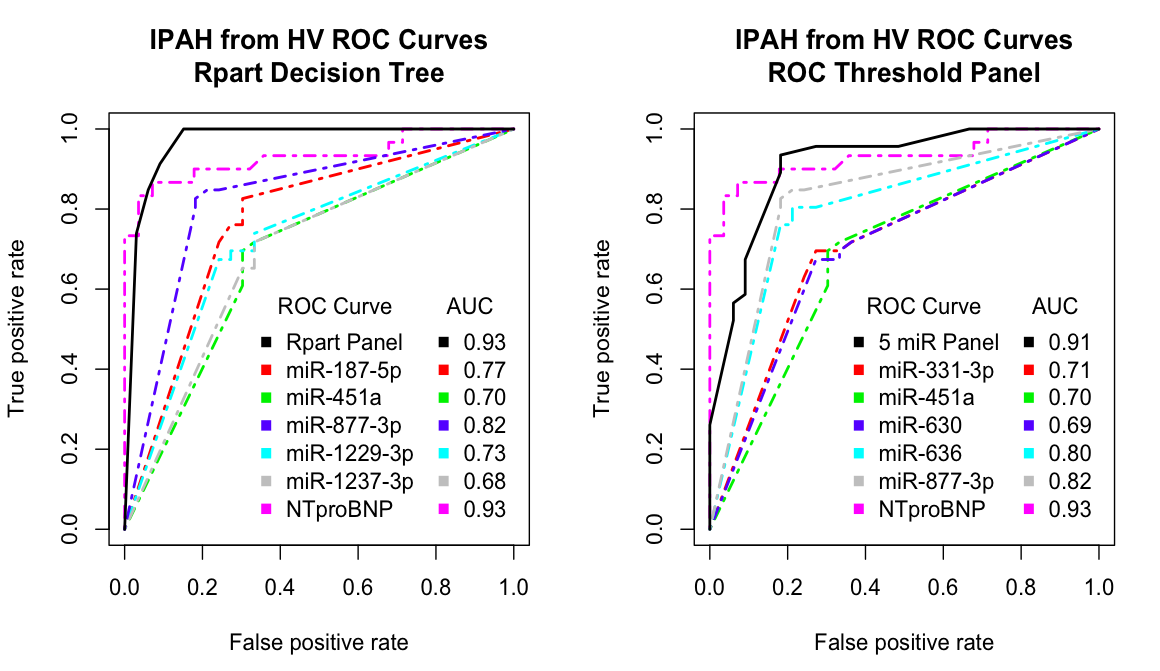


Figure 4‑2 ROC curves for top ROC threshold panel, Rpart model in black, their constituent miRNAs and NTproBNP reference in purple.

## Conclusion

This chapter has demonstrated the diagnostic utility of plasma miRNAs in a phenotypically diverse cohort. A multi-level process for selecting the most relevant features identified potential predictors that are likely to give interesting results when combined.

Panels built either by the Rpart decision tree or the ROC thresholds have shown increased performance by ROC AUC than their component miRNAs. However, the Rpart panel equalled NTproBNP and the ROC threshold panel did not.

With the exception of miR-331-3p the feature selection methods Boruta and LASSO both identified the same miRNAs. Indicating that, in this data set, irrespective of feature selection algorithm the features with diagnostic utility will be identified. This is important as due to large cost, global screening methods such as microarrays often have a sample number smaller than is useable by LASSO. In cases such as these feature selection by Bortua would be the preferred option.

# Identification of Biomarkers of PAH in SSc

## Introduction

PAH this is a serious complication affecting around 10% of patients with SSc therefore early detection is critical to long term survival. As previously discussed the DETECT study has endeavoured to address this with a diagnostic score determined from the results of at least 5 different tests (Coghlan et al., 2014). The tests require equipment and expertise that is only available at specialist centres, echocardiography for example. Additionally, the test is optimized to maximize positive predictive value at the detriment of the negative predictive value. This limits the algorithms practicality generating unnecessary numbers of patients where PAH has to be ruled out by RHC. A blood-based test has the potential to be faster, simpler and cheaper than DETECT and could potentially reduce unnecessary RHC.

## Methods

Plasma in citrate buffer was acquired from the Cardiovascular disease and Lung disease (STH-Obs) as detailed in 2.1. RNA was extracted as described in 2.1 and miRNA microarray performed as described in 2.2.

Patient demographics shown in Table 5‑1 show patients well matched, with non-significance, in age, gender and ethnicity. As expected, diagnostic measures of PAH are significantly different between groups. Absences of left heart disease is shown for both groups by non-significance and a PAWP ≤15 mmHg.

|  |  |  |  |
| --- | --- | --- | --- |
|  | SSc | SSc-PAH | p-value |
| **N** | 10 | 18 |  |
| **Age** | 62.1 (56.5 - 72.6) | 69.6 (65.1 - 72) | NS |
| **Gender (% Female)** | 90% | 67% | NS |
| **Deaths (%)** | 10% | 67% | 0.013 |
| **Ethnicity (% Caucasian)** | 90% | 100% | NS |
| **mPAP (mmHg)** | 21 (19 - 21) | 31 (30 - 45) | <0.001 |
| **PVR (Dynes.s/cm5)** | 130 (122 - 164.1) | 381.4 (270.3 - 783.9) | <0.001 |
| **Cardiac Output (L/min)** | 6.5 (1.1) | 5.1 (1.4) | 0.015 |
| **PA Wedge Pressure (mmHg)** | 9.8 (4) | 10.9 (4) | NS |

Table 5‑1 Summary of patient and control demographics in the SSc with PAH biomarker identification cohort Normally distributed data shown as mean + standard deviation and comparisons made using Students t-test. Non-normally distributed data shown as median + interquartile range and comparisons made using Mann Whitney U-test. Categorical data was tested with Chi-Squared.

## PAH in SSc Biomarker Feature Selection

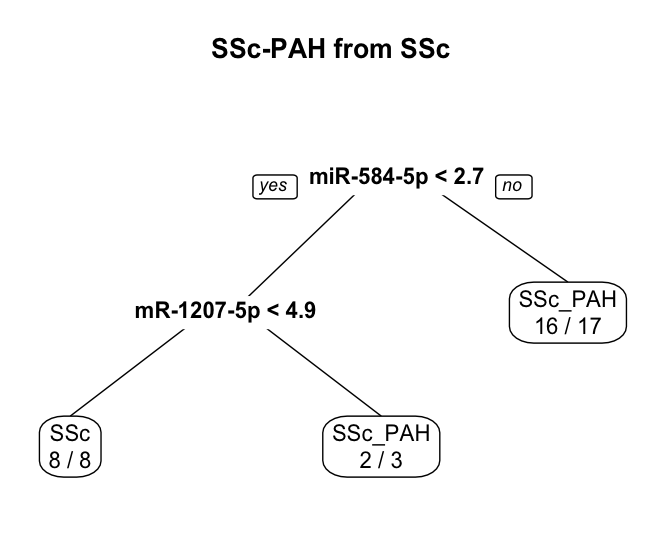
Identification of biomarkers of PAH within patients with SSc was performed on the data subset containing SSc both with and without PAH as determined by RHC, containing 210 miRNAs. Analysis was performed as described in 2.2 and 2.4. 73 miRNAs with a ROC AUC <0.65 were identified and removed from the initial 133 miRNAs. Spearman correlation analysis on the remaining miRNAs identified 41 correlating miRNAs that were removed leaving a filtered data set containing 19 miRNAs for final feature selection. Due to the small sample size feature selection was only performed with the Boruta algorithm, this identified 7 miRNAs (Table 5‑2).

|  |  |
| --- | --- |
| microRNA | Boruta Feature Significance (%) |
| miR-197-5p | 77.8 |
| miR-346 | 85.9 |
| miR-485-3p | 94.9 |
| miR-584-5p | 100 |
| miR-630 | 100 |
| miR-1207-5p | 89.9 |
| miR-3613-3p | 67.7 |

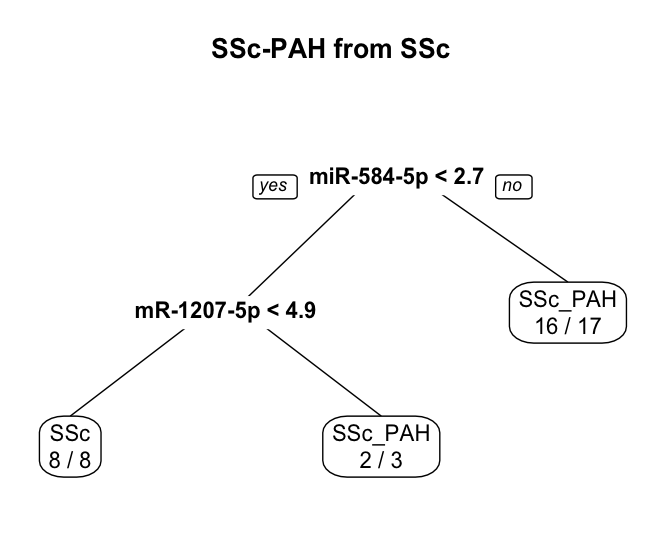
Table 5‑2 Biomarkers identified by the Boruta feature selection algorithm.

## Combing biomarkers with decision trees

An Rpart decision tree was used to discern a panel of biomarkers. Due to the smaller sample size a smaller cross-validation size of 3-fold was used to minimize over-fitting. The resulting decision tree is shown in Note, unlike the ROC based panels the standard deviation of performance metrics across the cross-validations is not available from the Caret cross-validate implementation of the Rpart algorithm.

**

*Figure 5‑1* and the diagnostic power is summarised in the metrics detailed in Table 5‑3. Note, unlike the ROC based panels the standard deviation of performance metrics across the cross-validations is not available from the Caret cross-validate implementation of the Rpart algorithm.

**

*Figure 5‑1 Rpart decision tree for separating SSc with PAH from SSc. Each terminal node represents the assigned classification with the fraction of correctly classified samples given below.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | AUC | Sensitivity | Specificity | Positive Predictive Value | Negative Predictive Value |
| Rpart Decision Tree | 0.89 | 0.89 | 0.90 | 0.94 | 0.82 |

Table 5‑3 Rpart decision tree performance metrics summarising the diagnostic power for the classification of PAH within SSc.

## Combining biomarkers by ROC determined thresholds

Combined ROC thresholds as described in 2.3 was used to build biomarker panels were built using the 7 miRNAs identified in 5.3 choosing a panel size containing from 2 to 5 miRNAs a total of 112 combinations were possible. 8 panels had a sensitivity and specificity greater than 0.8 and were subjected to 10 repeats of k fold cross-validation (k=3). Further filtering and ordering of the cross-validated panels, reduced the list of final panels to 2 (*Table 5‑4*). The small standard deviation of the performance metrics shown is indicative of stability within the performance of the panel.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| miRNAs in Panel | Se | Sp | Positive Predictive Value | Negative Predictive Value |
| miR-3613-3p, miR-584-5p, miR-630 | 0.92  (0.039) | 0.87  (0.048) | 0.93  (0.024) | 0.86  (0.055) |
| miR-485-3p, miR-584-5p, miR-630 | 0.88  (0.018) | 0.86  (0.052) | 0.92  (0.027) | 0.80  (0.021) |

*Table 5‑4 Top two miRNA panels for determining PAH in SSc. Performance metrics are shown as the mean and, in brackets, standard deviation from the cross-validation.*

## Comparison of Biomarker Performance

To compare the diagnostic performance of the panels as biomarkers of PAH ROC curves for both the Rpart decision tree and the top ROC threshold panel of 3 miRNAs were built. To demonstrate the added performance in a panel the constitutive miRNAs for each panel have been drawn along with NTproBNP as a reference, Figure 5‑2.

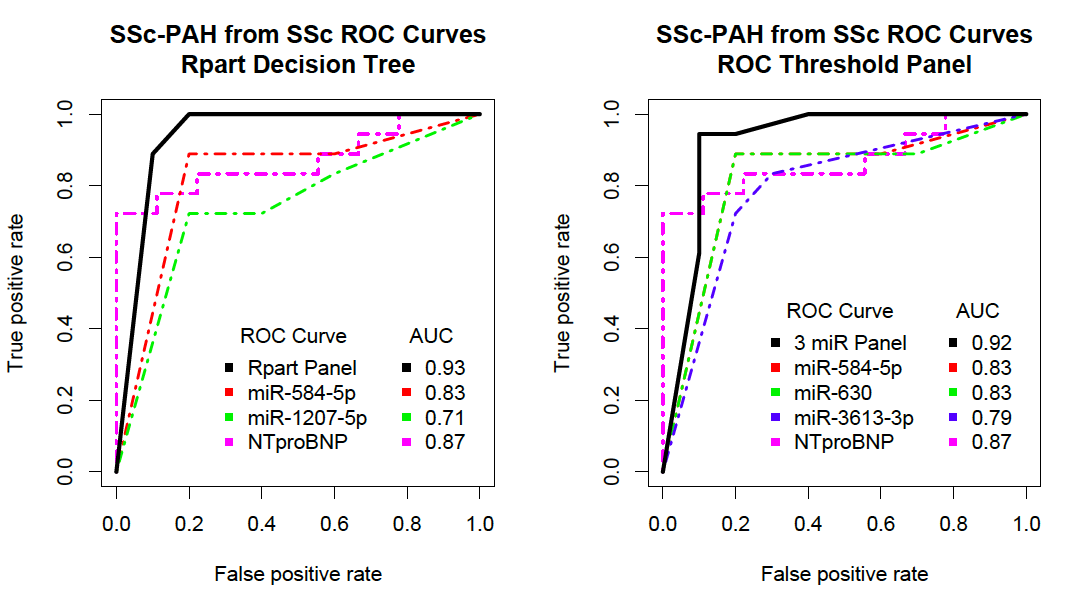


Figure 5‑2 ROC curves for top ROC threshold panel, Rpart model in black, their constituent miRNAs and NTproBNP reference in purple.

## Discussion

This chapter has applied the analytical method demonstrated in the larger and more phenotypically diverse IPAH HV cohort from chapter 0. The panels built by features selected by Boruta alone were successfully built into panels with superior performance to an individual miRNA and NTproBNP.

A direct comparison between the miRNA panel and DETECT score would have been preferable. Unfortunately, a DETECT score was not available for patients within this study as the Sheffield patient biobank data does not include the serum levels for NTproBNP, ACA, or urate nor the required echocardiograph measures. However, the 3-miRNA panel exceeds the overall positive predictive value from 0.35 to 0.93 of the DETECT study with only a minimal reduction in the negative predictive value 0.98 to 0.86 (Coghlan et al., 2014).

# Validation of Microarray Results by QPCR

Validation by an independent analytical technique is necessary to verify microarray findings as sensitivity and specificity are sacrificed in order to multiplex. This can lead to an increase in type I and type II errors. Additionally, the design of the of the miRNA microarray probes is unique to the platform and cannot be replicated, meaning that whilst a miRNA might be accurately profiled by microarray it may not be accurately determined by an alternate method. Validation by miRNA QPCR is a suitable technique for validation of miRNA microarray results.

Due to the lower availability of samples from patients SSc with and without PAH miRNA’s were first checked and validated in a small verification cohort of healthy volunteers and IPAH patients.

## Initial MiRNAs for Selected for Verification

MiRNAs from the top two ROC based panels determined from the miRNA microarray results for IPAH from HV and SSc-PAH from SSc was chosen for validation and detail in Table 6‑1.

|  |  |
| --- | --- |
| miRNA | Disease Panel |
| miR-331-3p | IPAH |
| miR-451a | IPAH |
| miR-485-3p | SSc |
| miR-584-5p | SSc |
| miR-630 | IPAH and SSc |
| miR-636 | IPAH |
| miR-671-5p | IPAH |
| miR-877-3p | IPAH |
| miR-3613-3p, | SSc |

Table 6‑1 miRNAs chosen for validation and their association with previously described disease panels.

## Methods

Reverse-transcription of RNA to cDNA was performed as described in 2.5.1 followed by real-time PCR as described in 0. Normalisation was performed to the mean of miRNA-320c and miRNA-142-3p described previously by (Schlosser, McIntyre, White, & Stewart, 2015)

## Initial qPCR Probe Verification with Test Cohort

Due to difficulties associated with amplifying miRNAs the qPCR probes were checked for clean amplification in a small cohort of HV and IPAH samples. Patients are well matched with non-significance in age, gender, and ethnicity *Table 6‑2*. IPAH patients are shown to have PAH from the appropriate values in the catheter based diagnostic measures.

|  |  |  |  |
| --- | --- | --- | --- |
|  | HV | IPAH | p-value |
| **N** | 5 | 7 |  |
| **Age** | 50.6 (14) | 41.1 (18.9) | NS |
| **Gender (% Female)** | 40% | 71% | NS |
| **Ethnicity (% Caucasian)** | 100% | 57% | NS |
| **mPAP (mmHg)** | - | 61.8 (7.5) | - |
| **PVR (Dynes.s/cm5)** | - | 1197.6 (519.1) | - |
| **Cardiac Output (L/min)** | - | 4.3 (1.5) | - |
| **PA Wedge Pressure (mmHg)** | - | 8.8 (1.9) | - |

*Table 6‑2 Demographics table for initial qPCR probe verification cohort of IPAH and HV samples.* *Normally distributed data shown as mean + standard deviation and comparisons made using Students t-test. Non-normally distributed data shown as median + interquartile range and comparisons made using Mann Whitney U-test. Categorical data was tested with Chi-Squared.*

## Results

R was used to interrogate the qPCR results of each miRNA. 5 out of 9 miRNAs failed due too extremely low to no sample amplification or NTC amplification close or within sample amplification, 2.5.3.1. These findings are summarised in Table 6‑3.

|  |  |  |
| --- | --- | --- |
| miRNA | qPCR Pass/Fail | Notes |
| miR-331-3p | Pass | - |
| miR-451a | Pass | - |
| miR-485-3p | Pass | - |
| miR-584-5p | Pass | - |
| miR-630 | Failed | Amplification in NTCs in-line with sample amplification |
| miR-636 | Failed | Amplification in NTCs and failed amplification within most duplicates |
| miR-671-5p | Failed | Amplification in NTCs |
| miR-877-3p | Failed | Amplification in NTCs |
| miR-3613-3p, | Failed | Amplification in NTCs and the within majority of samples failed to amplify |

Table 6‑3 Summary of initial qPCR verification results following inspecting of amplification curves.

## Additional MiRNAs for Selected for Validation

As only two ROC based biomarker panels had been previously determined in the SSc from SSc-PAH group additional biomarker targets were sought. The ROC based biomarker discovery methodology was repeated with the miRNAs that had experimentally failed removed from the input before processing. Only 1 panel exceeded the QC cut-offs described in 2.5.3.1 the panels predictive metrics are summarised in *Table 6‑4*. The small standard deviation of the performance metrics shown is indicative of stability within the performance of the panel.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| miRNAs in Panel | Se | Sp | Positive Predictive Value | Negative Predictive Value |
| miR-486-5p, miR-584-5p, miR-638 | 0.85  (0.027) | 0.83  (0.082) | 0.90  (0.040) | 0.76  (0.023) |

*Table 6‑4 Additional panel of miRNAs for determining SSc-PAH from SSc. Performance metrics are shown as the mean and, in brackets, standard deviation from the cross-validation.*

## Verification of additional SSc-PAH ROC based biomarker miRNAs

As miRNA-584-5p had previously been validated real-time PCR was performed only on miRNA-486-5p and miRNA-638 in the same cohort as 6.3. As previously described, R was used to interrogate the real-time PCR results for each miRNA.

­Both miRNAs validated according the criteria described in 2.5.3.1.

## Validation of miRNAS in SSc and SSc-PAH

Once successful amplification had been determined in the miRNAs were validate in a larger SSc and SSc-PAH cohort. This would be used to determine if the experimentally verified miRNAs could provide diagnostic utility between SSc-PAH from SSc.

## 

## Cohort

The larger SSc and SSc-PAH cohort consisted of 43 plasma samples taken at the time of diagnosis. Patients are well matched with non-significance in age, gender, and ethnicity Table 6‑5. As expected, diagnostic measures of PAH are significantly different between groups. Absences of left heart disease is shown for both groups by non-significance and a PAWP ≤15 mmHg.

|  |  |  |  |
| --- | --- | --- | --- |
|  | SSc | SSc-PAH | p-value |
| **N** | 14 | 29 | - |
| **Age** | 62.1 (56.5 - 72.6) | 69.1 (63 - 71.1) | NS |
| **Gender (% Female)** | 93% | 69% | NS |
| **Deaths (%)** | 7% | 55% | 0.007 |
| **Ethnicity (% Caucasian)** | 93% | 97% | NS |
| **mPAP (mmHg)** | 21 (20 - 21) | 37 (30 - 48.5) | <0.001 |
| **PVR (Dynes.s/cm5)** | 148 (123.9 - 164.1) | 381.4 (263.5 - 730.2) | <0.001 |
| **Cardiac Output (L/min)** | 6.8 (1.1) | 5.4 (1.7) | 0.003 |
| **PA Wedge Pressure (mmHg)** | 8 (7 - 11) | 11 (7.5 - 13.5) | NS |

Table 6‑5 Demographics table for SSc-PAH from SSc qPCR diagnostic test. Normally distributed data shown as mean + standard deviation and comparisons made using Students t-test. Non-normally distributed data shown as median + interquartile range and comparisons made using Mann Whitney U-test. Categorical data was tested with Chi-Squared.

## Methods

Reverse-transcription of RNA to cDNA was performed as described in 2.5.1 followed by real-time PCR as described in 0. Normalisation was performed to the mean of miRNA-320c and miRNA-142-3p described previously (Schlosser et al., 2015) and the mean of the SSc group was used for the comparator for delta delta Ct.

## Results

The performance of the miRNAs in the qPCR screen were characterised by testing for concordance with the microarray expression, univariate comparisons and together in a panel.

## Concordance with microarray data

Correlations were calculated for each microRNA between samples which were in the original microarray study. Expression values from the microarray and -deltaCt real-time PCR values were used in the Spearman correlation calculations, results are shown in Table 6‑6 and plotted in Figure 6‑1. The inverse of deltaCt was used as it has the same log2 relationship of the microarray expression and is normalised for experimental variation.

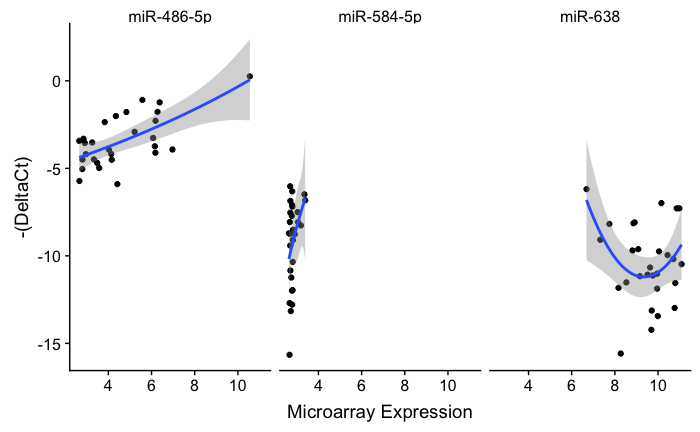


Figure 6‑1 Real-time PCR concordance with microarray expression blue loess trendline with 95% confidence boundary shown in grey.

|  |  |  |  |
| --- | --- | --- | --- |
| miRNA | Correlation Coefficient | p-value | Adjusted p-value |
| miR-486-5p | -0.55 | 0.0028 | 0.0083 |
| miR-584-5p | -0.28 | 0.1443 | 0.4329 |
| miR-638 | 0.05 | 0.8161 | 1.0000 |

Table 6‑6 Results of Spearman’s Rho correlation of real-time PCR expression with microarray expression.

## Univariate Comparisons

Relative quantification was calculated and pairwise comparisons expression differences were tested by Mann-Whitney U test with Bonferoni. Expression is shown in Figure 6‑2 and comparisons are summarised in Table 6‑7.

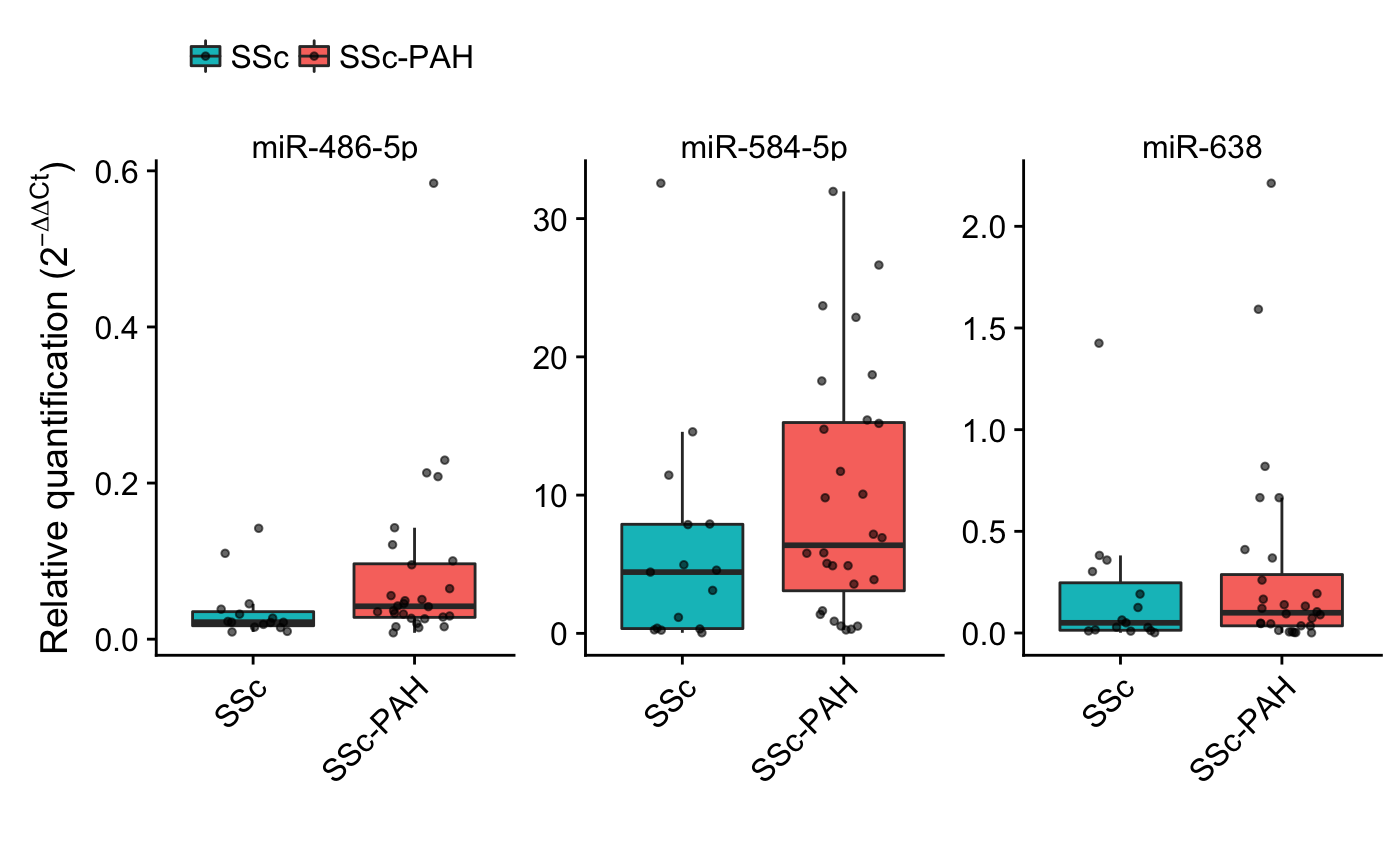


Figure 6‑2 miRNA expression analysis by real-time PCR. After correction for multiple testing there are no significant differences. Mann-Whitney U test with Bonferroni correction.

|  |  |  |
| --- | --- | --- |
| miRNA | p-value | Adjusted p-value |
| miR-486-5p | 0.020 | 0.059 |
| miR-584-5p | 0.090 | 0.271 |
| miR-638 | 0.472 | 1.000 |

Table 6‑7 Summary results from Mann-Whitney U test comparisons of SSc miRNA panel adjustment for multiple comparisons by Bonferroni.

## Performance in a panel

The original cut-offs used for each miRNA, either in a decision tree or ROC based panel could not be reused with the real-time PCR expression as the two data sets were not normalised in the same way. Therefore, new cut-offs were determined from the real-time PCR data as previously described.

An Rpart decision tree with 3-fold cross-validation was built using deltaCt as the input value. The resulting decision tree is shown in Figure 6‑3 and model metrics summarised in Table 6‑8. Note, unlike the ROC based panels the standard deviation of performance metrics across the cross-validations is not available from the Caret cross-validate implementation of the Rpart algorithm.



Figure 6‑3 Rpart decision tree for separating SSc with PAH from SSc. Each terminal node represents the assigned classification with the fraction of correctly classified samples given below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Method | AUC | Sensitivity | Specificity | Positive Predictive Value | Negative Predictive Value |
| Rpart Decision Tree | 0.71 | 0.82 | 0.60 | 0.79 | 0.64 |

Table 6‑8 Rpart decision tree performance metrics summarising the diagnostic power for the classification of PAH within SSc.

The ROC based panel was built using the deltaCt values with 3-fold cross-validation. All 4 possible panel combinations from the input miRNAs model metrics are summarised in *Table 6‑9*. In general, the standard deviation in the performance metrics is small, however, this variance is an order of magnitude greater than the performance metrics seen in the microarray results. This is suggestive of these panel being more unstable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| miRNAs in Panel | Se | Sp | Positive Predictive Value | Negative Predictive Value |
| miR-486-5p, miR-584-5p, miR-638 | 0.91  (0.048) | 0.62  (0.105) | 0.75  (0.056) | 0.46  (0.060) |
| miR-486-5p, miR-584-5p | 0.43  (0.133) | 0.87  (0.134) | 0.86  (0.060) | 0.45  (0.034) |
| miR-486-5p, miR-638 | 0.46  (0.186) | 0.76  (0.166) | 0.79  (0.077) | 0.43  (0.040) |
| miR-584-5p, miR-638 | 0.41  (0.002) | 0.69  (0.053) | 0.72  (0.070) | 0.39  (0.037) |

*Table 6‑9 All possible panel combinations of the miRNAs used in qPCR validation and performance metrics summarising the diagnostic power in determining PAH. Performance metrics are shown as the mean and, in brackets, standard deviation from the cross-validation.*

## Comparison of Panel Performance

To compare the performance of the Rpart decision tree and the panel ROC curves were plotted for both the model, constituent miRNAs and NTproBNP (Figure 6‑4).

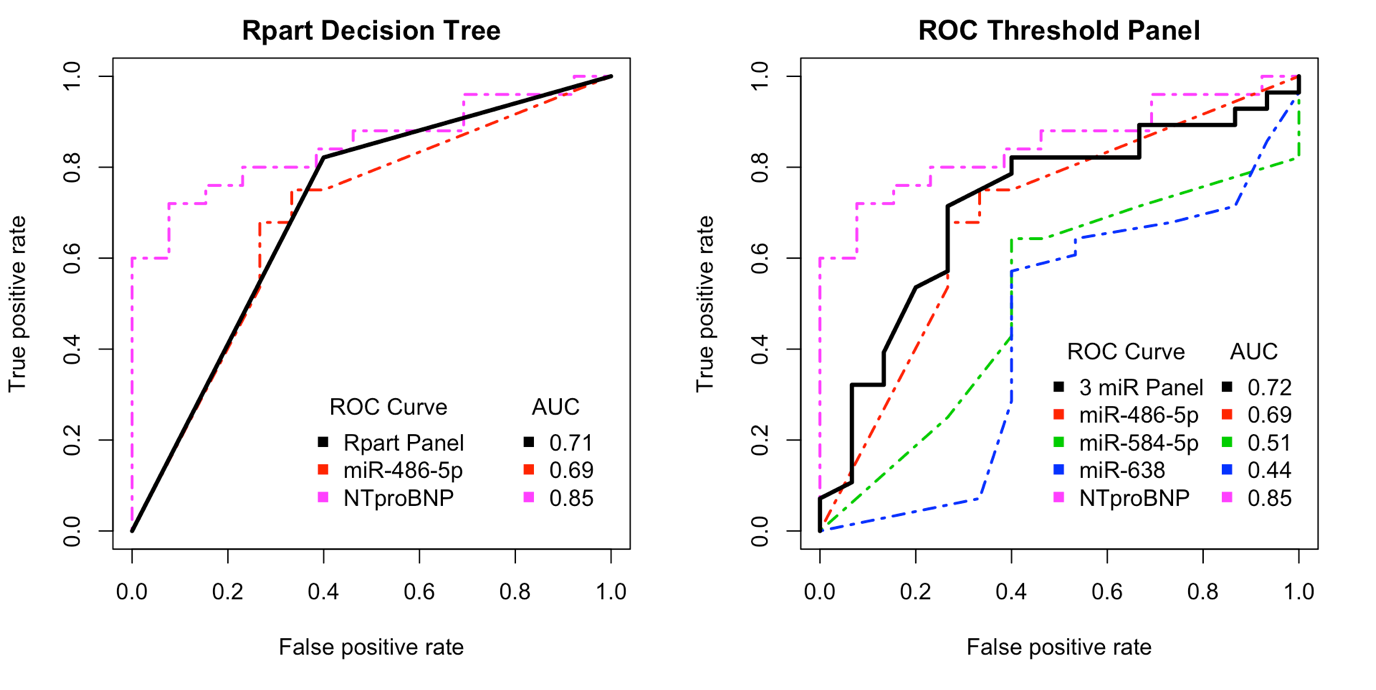


Figure 6‑4 ROC curves for top ROC threshold panel, Rpart model in black, their constituent miRNAs and NTproBNP reference in purple.

## Discussion

This chapter has demonstrated methodologies attempting to validate potentially diagnostic miRNAs discovered using an unbiased microarray screen. However, univariate differences between cohorts was only demonstrated with miR-486-5p, albeit only in the p-value prior to adjustment for multiple testing. However, this has the potential to become significant with a larger sample size.

## Multivariate Biomarkers

MiRNA-584-5p and miRNA-638 did not demonstrate significant univariate group differences or microarray correlations, however, when combined in a panel with miR-486-5p the 3 miRNAs still demonstrated superior performance to any single miRNA within the panel. Unfortunately, they did not offer any significant benefit to testing NT-proBNP alone, NT-proBNP ROC AUC was 0.85 compared to the panels 0.72. Their additive effect adds to evidence that combining multiple miRNAs is worthwhile strategy.

Whilst there are currently no reports in PAH of combining multiple miRNAs in panel in the literature (Tang, 2018) combined a single miRNA with echocardiogram results. In this study miRNA-509-3p had an AUC of 0.69 alone, this increased to 0.84 when used in combination with the echocardiogram.

## miRNA-486-5p

MiRNA-486-5p showed the most promising profile in qPCR validation, however, no reports of miR-486-5p in PAH or SSc literature were identified. However, there are reports in pulmonary fibrosis and lung cancer diseases areas close to aspects of PAH. However, there have also been associations of this miRNA with platelets.

## miRNA-486-5p in Platelets

4 publications mentioned platelets and expression of miR-486-5p, of these 2 publications specifically identified platelet derived miR-486-5p.

Platelet extracellular vesicles from stored healthy donors platelets were found to have in increased miR-486-5p expression following 5 days under standard blood banking conditions (P.-R. A et al., 2017b). Additionally, Maues et al., found reduced miR-486-5p in platelets following storage (Maués, Moreira-Nunes, & Burbano, 2020). Together, this work is suggestive of miR-486-5p being released from platelets in storage. However, the protocol for sample preparation in STH-Obs is to immediately centrifuge the collection tubes following phlebotomy. This time scale is far smaller than the 5 days used in both studies, reducing the risk of extracellular vesicles contamination the plasma sample. However, checking for any correlation with platelet count would be preferable.

## miRNA-486-5p in Pulmonary Fibrosis

Fibrosis in the lungs is a hallmark of systemic sclerosis, Xioming et al, found miR-486-5p expression reduced in animal models of lung fibrosis and humans with silicosis and idiopathic pulmonary fibrosis (Ji et al., 2015). In patients the reduction was seen within both serum and lung tissue samples. They went on to identify that over expression in the mouse model reduced lung fibrosis.

## miRNA-486-5p in Lung Cancer

Expression of miR-486-5p was found to be significantly reduced in lung squamous cell carcinoma (Yang et al., 2019) and non-small cell lung cancer (Tian et al., 2019). In Tian et al., further investigations found reduction of miR in serum and within NSCLC cell lines overexpression of miR-486-5p inhibited cellular proliferation and induced apoptosis. However, Gao et al., identified the opposite interaction in NSCLC (Gao, Yuan, Yuan, Yuan, & Wang, 2018). Observing in NSCLC cell lines that downregulation of miR-486-5p led to cellular proliferation and inhibition supressed growth.

The conflicting findings for miR-486-5p action in lung cancer and high miR levels in pulmonary fibrosis being anti-fibrotic make it difficult to interpret the high miR-486-5p seen in SSc-PAH plasma. It is clear that miR-486-5p has a role to play in lung pathophysiology and that dysregulation is undesired. However, in this case there is a possibility that the potential source could be from platelet contamination and only more data and research in PAH specific models will help to elucidate this.

## Validation Performance

Out of the 10 miRNAs tested by qPCR only 1 demonstrated desirable performance characteristics in validation (10% concordance). The miRQC cross-platform miRNA study sought to determine the performance of miRNA platforms this provides a bench mark for the likely concordance amongst platforms (Mestdagh et al., 2014). In the study it was reported that Agilient miRNA arrays had a 70-75% concordance with all tested technologies, which included Exiqon LNA and Taqman qPCR. However, hierarchal clustering of platform concordance placed the Agilent miRNA microarrays in a different cluster to the qPCR validation methods. The clustering indicated that LNA qPCR had the best concordance with Ilumina and Ion Torrent NGS systems.

The miRQC study was performed under ideal sample conditions, however, the concordance is also evidenced by a disease relevant study which profiled miRNAs from IPAH and HV explanted lungs (Wu et al., 2016). Here, discovery performed with Agilient microRNA microarrays saw 5 out of 7 target miRNAs validated by Taqman qPCR (71% concordance). The qPCR technology in this study is not the same LNA qPCR used in my work, but it is unlikely that this is a major cause of the lack of validation concordance.

The biggest difference with this work and that reported by Wu et al., is the source of RNA. Plasma miRNA concentrations very low and undetectable by spectrographic methods this means that when running an experiment RNA input is likely sub-standard and the quantity cannot be normalised, which is the approach used with total RNA as an input.

Microarray technology was well developed at the beginning of this project, however, more recently smallRNA sequencing technologies have matured enabling sequencing from low input sources such as in single cells, exosomes and plasma (Reyfman et al., 2019), (Hannafon et al., 2016), (Yeri et al., 2018). Given the advances and the miRQC study identifying NGS techniques in the same cluster as qPCR techniques NGS discovery followed by PCR validation is a valid consideration for future studies.

# General Discussion

Biomarkers in PAH still remain elusive and difficult to identify. Consequently, this is still an actively researched area within PAH. As of October 2019, there were 530 PAH publications categorised with the MeSH term for biomarkers, for 60 of those biomarkers was a major topic of the publication. Over the past 3 years the rate of publications has been 47 and 5 publications per year for all publication/major topics respectively Figure 7‑1.

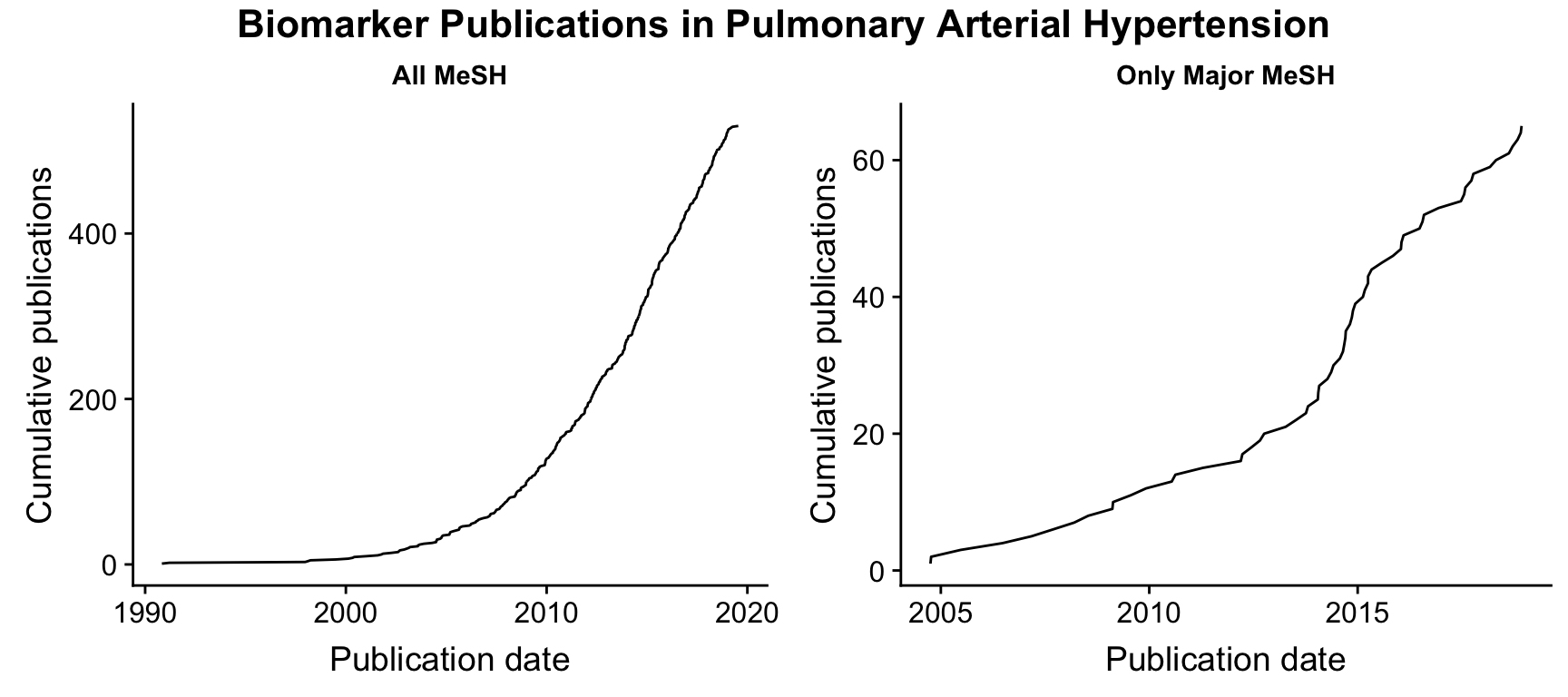


Figure 7‑1 Publications tagged with the MeSH term biomarkers within a corpus of PAH publications.

Recently Janssen who now own Actelion’s significant PAH drug portfolio began a large, international PAH biomarker study on the 10th December 2019 (clinical trial identifier NCT04193046). The study for the Identification of Biomarker Signatures for Early Detection of Pulmonary Hypertension (CIPHER) seeks to recruit 800 patients from centres in the USA and Europe. Recruitment from multiple sites across continents should help of any predictive mode to be more generalised and accurate. Blood and breath samples from incident and prevalent samples will be taken from patients who have undergone RHC. The diverse sampling of both volatile organic compounds in participants breath as well as blood miRNA will increase the potential for identifying useful biomarkers. However, the inclusion of CTEPH, Isolated pre-capillary, post-capillary and prevalent PH samples in the cohort risks it becoming too heterogenous which may negatively affect the primary outcome of “identifying biomarker signature(s) of participants at risk of PH”.

This thesis set out to identify a set of diagnostic biomarkers for subsequent validation from an unbiased microRNA microarray screen. Pre-empting a level of validation failure, typical of microarrays, a method was developed to identify a validation set containing not just the top miRNAs but, additional, next best targets too.

First investigated was the best source of circulating miRNAs. From the results of both differential expression and feature selection plasma was chosen as the best source for potential circulating miRNA biomarkers from matched sample whole blood and plasma miRNA microarray data. An analysis protocol was then developed in a larger IPAH and HV cohort and applied to the smaller SSc and SSc-PAH.

In total 10 miRNAs were identified for further investigation. Initial verification of performance by qPCR was performed in samples matched to the initial microarray study. This identified 3 miRNAs with satisfactory performance for validation within a larger SSc and SSc-PAH cohort.

The final miRNA panel of miRNA-436-5p, miRNA-584-5p and miRNA-638 with an AUC of 0.72 was inferior to 0.85 for NT-proBNP. Whilst this is indicative of NT-proBNP being a better diagnostic test it is only a marker of myocardial stress and not specific to PAH. The panel miRNAs may be linked to drivers or disease processes of PAH which could be deduced with further investigations. In-silico prediction of miRNA targets alone or in combination with proteomics, mRNA expression and pathway analysis may shed further light into as yet unknown diseases mechanisms. Additionally, it is unknown how the miRNAs will behave longitudinally or response to PAH therapies and with a sufficiently large longitudinal cohort miRNA response to survival and prognostic markers could be investigated.

The analysis method developed gave a ranked list of potential miRNA biomarker panels. Where the use of ROC cut-offs as panel constituents made it clear how individual miRNAs contribute to the final classification score. However, this may not make the best use of the continuous expression miRNA values. A logistic regression model with continuous data would likely provide superior diagnostic utility, however, by simplifying the data to categorical may result in a more generalised solution by reducing any negative effects of extreme or outlier values that would have more effect in a smaller sample size.

In this method, where possible, steps were taken to improve generalisation of any resulting panel as much as possible. At the sample level, discovery was performed with a mixed cohort with samples from the University of Sheffield and Imperial College and cross-validation was used. Unfortunately, the plasma samples were not processed in a block randomised design, this could have led to undesired batch effects. Whilst none were detected (data not shown) without randomised the sample processing it is extremely difficult to separate what is a sample effect from a batch effect.

The sample size in this study was small and a source of additional difficulty throughout the study. A small sample may have obfuscated the identification of more subtle but valid diagnostic patterns within the data that would have been clear in a larger cohort. Additionally, more samples may had led to more relationships being confirmed as statistically significant in the qPCR validation.

## microRNA-436-5p

As miRNA-436-5p had the strongest profile in the validation panel there is reason to investigate its relevance to PAH with further studies. As discussed in 6.7.2 miR-486-5p has a role within several many lung pathophysiology’s although conflicting reports make it difficult to build a hypothesis on mechanism of action from the literature alone. The relationship to the lung can also be seen in the disease (category C) MeSH term counts for the publications mentioning miR-486-5p, in particular lung cancer is the most frequent disease mentioned *Figure 7‑2*. This is relevant as PAH has been described as having a neoplastic phenotype with proliferation of apoptosis resistant pulmonary artery endothelial cells (Voelkel et al., 1998), {Rai:200fv}.

**

*Figure 7‑2 Bar chart summarising disease MeSH terms used in publications mentioning miRNA-486-5p.*

Despite high levels of miRNA-486-5p being seen in plasma it is unknown whether or not levels are increased in lung tissue. Comparing miRNA expression between lungs from PAH and control models would demonstrate changes in disease relevant tissue. In PAH pulmonary arterial and endothelial tissue has different roles, localisation of miRNA expression in lung tissue can be identified using fluorescent in-situ hybridisation with miRNA specific probes. Transfecting with either a pri-miRNA or miRNA inhibitor in PAH relevant cell lines, such as pulmonary artery smooth muscle cells would help determine if high or low expression of miR-486-5p caused a proliferative or apoptotic phenotype. Furthermore, mRNA expression array analysis following stimulation/inhibition could be undertaken to identify gene/pathway changes that further elucidate any role. Changes in mRNA expression can be linked to know PAH genes and miRNA targets to identify the potential mechanisms of action

In summary, despite limitations, this study has identified miR-486-5p as a target for further research and demonstrated a novel method for identifying biomarker panels from a small and heterogenous cohort.

## Inclusion of additional data in panels

This study set out to find independent miRNA biomarkers exclusive of additional data. Behind this was the hypothesis that a circulating miRNA biomarker should be a marker of the disease process or effects. As the diseased sample cohort all have PAH we hypothesised that this drastic change in lung and system biology would produce something detectable in the circulation. Furthermore, as the disease is histologically the same irrespective of age demographics then a disease related biomarker should be independent of these.

As a clear miRNA biomarker(s) did not appear future could be performed with additional phenotype data or protein biomarkers in. Initial candidates for this would be NT-proBNP, age and gender. Whilst the cohorts were well matched for age and gender there was still differences in these values that could aid classification.

The ROC based panel methodology could be used with age and NT-proBNP but would not be suitable for the binary variables such as gender. Analysis using logistic regression would be the most suitable as this can directly model the relationship between miRNAs and additional factors through the use of interaction terms. Additionally, the extent to which the additional data supports the classification accuracy can be determined from the model coefficients and AIC.

Unfortunately, in this study there was less than ideal number of samples with data on NT-proBNP values. Only 65% of samples had data on NT-proBNP and using these would further reduce the subset of data available for analysis. Imputation could have been used to overcome this, e.g. the MICE algorithm but this is typically frowned upon, especially in smaller datasets, when dealing with medical data (Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006).

## Experimental improvements

The poor validation rate maybe an indication of a poor step or steps in the methodology or an intrinsic difficulty in validating microRNA microarrays. Whilst I have not found evidence in the literature beyond the miRQC study, as negative results are rarely published, many groups have indicated in discussions that miRNA microarray results are very difficult to validate. Nevertheless, the early experimental process could have been improved to identify any part of the process that should have been performed differently and quantify the accuracy of microarray early on.

Prior to starting the cohort of samples from Imperial College London had already been extracted using the Norgen Plasma/Serum RNA slurry kit. This informed our decision to use this extraction method going forward. This kit was primarily designed for extracting large yield of RNA from large input volumes. As we were working with 500 µl aliquots of biobank plasma a kit optimized for smaller inputs may have been better.

The Norgen kit used an atypical 100 µl elution volume, by comparison the Qiagen miRNeasy and Zymo cfRNA Serum & Plasma Kit have an elution volume of 20 µl and 15 µl respectively. Due to the large elution volume and from an outlook of caution the Zymo clean and concentrate kit was used to concentrate RNA into an elution of 11.5 µl.

At this point two hypothesis should have been tested. Firstly, “that Norgen was a suitable extraction method for 500 µl of plasma” and secondly “that concentration of the 100 µl elute was required and that this did not introduce bias or additional variability in the results”. The possible RNA purification options should have been compared using a qPCR panel of miRNAs known to be detected at medium to high levels within plasma e.g. miR-23a-3p, miR-103a-3p (Donati, Ciuffi, & Brandi, 2019). The results of this to be used to determine yield and experimental variability.

Once extraction had been confirmed it would have been prudent to test that we could detect known plasma miRNAs in PAH. When this study began only a miR-150 had been reported (Rhodes et al., 2013). This was investigated on the microarray results and was found to be detected at very low levels. However, a fair comparison could not have been made as the samples were different, EDTA as opposed to citrate plasma, and the detection method was qPCR as opposed to microarray. With an n of 1 it was not possible to attribute its presence or even absence to a conclusion about performance. Confirming miR-150 expression in a qPCR experiment would have been more suitable, however, the question of sample source differences would still have been present. As previously, mentioned microRNAs known to be at a medium to high expression could have also been investigated by microarray. This was performed as, at the time, it was felt there was enough evidence of successful detection with differentially expressed miRNAs being detected and an on chip detected dynamic range of relative expression values between from 2.5 to 13.13.

Ultimately the choice was made to use Norgen for extractions and to concentrate the elute prior to microarray processing. Whilst the efficacy of this protocol was not tested as suggested above an initial test microRNA microarray proved successful in terms of a providing relative expression of a range of miRNAS at a variety of expression levels.

All this potential experimental testing comes with a cost, multiple extraction kits, cDNA synthesis, qPCR probes & master mix and potentially multiple microarrays. Budget constraints would not allow for such an exhaustive validation. Funding was only provided for the microarrays themselves meaning testing and validating multiple extractions and miRNAs was prohibitive. Additionally, the experimental steps used in this study have all been demonstrated in peer reviewed publications.

The Norgen plasma/serum kit has been successfully used in smallRNA Seq studies (El-Mogy et al., 2018) and in a comparison study between competitor kits from Qiagen, Zymo, Macherey-Nagel and Exiqon it was remarked that “PCR replicates were excellent” and that “The five kits gave comparable results in terms of Cq values” (Brunet-Vega et al., 2015). However, in the same study the 11 spike in controls had “unsatisfactorily high” variability between samples. Importantly, there was no clear best or worst kit and for this study the use of Norgen for extraction is justified.

The Zymo clean and concentrator has been used to plasma miRNA samples for microRNA microarray, smallRNA Seq and qPCR (Qin et al., 2015), (Danielson et al., 2018),

(Srinivasan, Treacy, Herrero, Reports, 2020, 2020b).

There have been many studies demonstrating the Agilent microRNA microarray platform. In particular Callari et al., demonstrated in RNA extracted from 350 µl of plasma reproducible results using the Agilient Human miRNA microarray platform (Callari et al., 2013). 11 patient samples were processed in triplicate across two batches intra-batch correlation between technical replicates higher than 0.99, and inter-batch was 0.89.

In summary, hindsight indicates that an exhaustive validation would have been appropriate potentially improving validation. However, at the outset of this project there was peer-reviewed evidence that the systems used work and there was no reason to doubt that it would be different in this study and budget constraints made exhaustive testing prohibitive.

## Technical Variation in miRNA Expression

As discussed in 7.1.3 Callari et al., found very low variation between technical replicates in plasma samples. Additionally, the miRQC study performed samples in replicate and found the Agilent microarray to be the most stable amongst replicates. Both of these studies set out to test the variation and reproducibility of miRNA technology and both concluded the Agilent platform performed well, even on RNA from plasma. Agilent may not be as well utilised in publications as the Affymetrix platform, but this is mostly down to the way microarrays are processed at universities. Typically, most departments or labs do not have the equipment or expertise and instead use core facilities for processing, the Affymetrix platform is most commonly used as it was designed for the higher throughput core facilities have.

Ideally, I would have set-up an experiment to test the variation between technical replicates between extraction methods and sample source. This was not considered at the time as it would have used a large number of valuable microarrays and samples. After the microarrays for this study had concluded the Lawrie lab worked with Promega on a prototype extraction for whole blood total RNA on the Maxwell 16 Robot. A small experiment of two healthy volunteers extracted and processed in replicate using a manual Norgen Tempus RNA kit and the prototype Maxwell and hybridised to an Agilent Human miRNA microarray. Each set of replicates is plotted at its mean expression on the x axis by the difference between replicates on the y Figure 7‑3. This demonstrates the difference between replicates in relation to the detected expression as there maybe relationship.



Figure 7‑3 Variance between replicates shown at the mean expression of each replicate for both healthy volunteers and both extraction methods. 0.5 log fold-change is shown on the chart as a light grey line for reference.

From the scatter plot it can be seen that the manual Norgen extraction displays a greater variance between replicates than the Maxwell extraction, with some probes exceeding a 0.5 logFC difference in measurements. Unfortunately this is not in the same extraction kit used in the plasma array nor the same sample type. However, one may draw a very tentative conclusion that a greater technical reproducibility is achieved with the robotic extraction. Additionally, it can be seen that the variance is not related to the expression level. If a heteroskedastic relationship was present miRNAs would require further pre-processing to ensure accuracy of downstream analysis.

## Can Microarrays be Quantitative?

After processing the probe intensities from a microarray experiment are transformed and returned as log2 relative expression levels. As a continuous variable this is suggestive of a quantitative result and was used as such in this study.

To understand whether a microarray has quantitative properties it needs to be compared to a known quantitative platform. Therefore if qPCR is quantitative then a microarray platform can be considered quantitative if it correlates with qPCR results.

The Microarray Quality Control (MAQC) project in 2006 was initiated to investigate the reliability of microarray platforms (S. L et al., 2006). The study systematically compared 7 leading microarray manufacturers at the time against 1,004 Taqman qPCR. In all cases correlations of R > 0.82 were found, the Agilent platform specifically had an R > 0.9 in all 3 different RNA test tissues. Whilst the Agilent gene expression arrays are not of the same probe design as the microRNA arrays, making comparisons and evaluation difficult, they both use Agilent’s proprietary Sureprint technology chip to print the DNA probes on a glass wafer.

Agilent miRNA microarray expression from 9 different tissues was compared to expression from Taqman qPCR (Ach, Wang, & Curry, 2008). To avoid complications with normalising miRNAs rations between tissues were calculated and directly correlated, for example the ratio of microarray expression of heart / liver was compared with the ratio of qPCR expression of heart / liver. In all 60 miRNAs were chosen for qPCR profiling and the study found “generally excellent” correlations between the platforms with 53/60 miRNAs with an R > 0.9.

In another study Sato et al., compared the Taqman qPCR results of 142 miRNAs to multiple platforms including the Agilent platform. Comparing all 143 qPCR miRs together they found the Agilent miRNA microarray to have good correlation of R of 0.85. However, they noted that the slope was < 1 demonstrating log-ratio compression

(Sato, Tsuchiya, Terasawa, & Tsujimoto, 2009). Whilst not specifically mentioned a slope of < 1 was also seen in 31 out of 36 microarray array/qPCR comparisons in Arch et al.

Whilst it is difficult to conclude whether microarrays are truly quantitative good correlation to the quantitative platform, qPCR, has been reported in several studies specifically looking at microRNA detection platforms. However, these studies show that microarrays are not directly comparable as they exhibit compression of dynamic range. However, it is clear that as miRNA expression increases there is a proportional increase seen in the qPCR results. Finally, the Agilent platform used in this study has consistently shown to be one of the best in performance as measured by correlation with qPCR.

## Future directions

Even within a single diagnosis it is clear that the miRNA expression is very heterogenous. To address this, I would plan to use a smaller, but more phenotypically similar discovery cohort for discovery. Firstly, a smaller cohort would ease the financial cost of using NGS (smallRNA SEQ). Secondly, it should follow that a more phenotypically similar cohort should have less patient-to-patient variation, helping identification of differentially expressed miRNAs. Alternatively, I would seek to utilise a discovery technology that matched the validation this reduces the technical differences and reduces technical complexity between discovery and validation. Fortunately, in miRNA discovery Taqman TLDA qPCR arrays exist for this very purpose. However, they only interrogate about a quarter of known or potential miRNAs. A strategy here may be to use TLDA arrays on a pilot feasibility study before expanding to a larger microarray or smallRNA SEQ study.

Many studies use global mRNA expression results in combination with global miRNA to identify target genes. Identifying an inverse relationship is supportive. however, this requires a confidence is the microarrays quantitative behaviour of the experiments.

When this study began a manual column-based miRNA extraction from plasma was the only available option. However, the Lawrie lab now has a robotic, magnetic bead based, nucleic acid processor the Maxwell 16 (Promega, Maddison WI). With columns manual miRNA extraction was limited to <8 samples per batch to keep extraction times to a minimum in an effort to preserve RNA integrity. However, the Maxwell system runs easily with up to 16 samples per batch. Utilizing robotic extraction also reduces batch to batch variation and in small scale tests on plasma this has shown clean and amplifiable miRNA (data not shown). Furthermore, to reduce batch effects block randomisation should be used throughout, from extraction batches, discovery batches and validation batches. Block randomisation can reduce batch effects from affecting the experimental outcome. If batch effects are present with a block randomisation design any these can be identified during exploratory data analysing and even corrected for by modelling the effect and correcting for it (Manimaran et al., 2016), (Leek, Johnson, Parker, Jaffe, & Storey, 2012).

In a future study I would first try to verify miRNA identified using selection methods from the discovery experiment, such as differential expression and feature selection techniques. This would almost certainly lead to greater than 10 miRNAs to be tested in the verification experiment. However, this has a better chance of leading identification of miRNA targets for validation in a larger cohort. The miRNAs tested in the verification experiment would be tested for concordance identifying miRNAs that show correlation with the discovery platform and ideally group differences. Results from either verification or validation experiment could then be confidently tested for multivariate diagnostic potential.

# References

A, A., B, K., S, H., M, G., & SJ, M. (2017a). miR-199a-5p and miR-495 target GRP78 within UPR pathway of lung cancer. *Gene*, *620*, 15–22. <http://doi.org/10.1016/j.gene.2017.03.032>

A, P.-R., T, K., MM, M., A, S., A, B., G, M., & G, S. (2017b). Transcriptomic profiling of platelet senescence and platelet extracellular vesicles. *Transfusion*, *57*(1), 144–156. <http://doi.org/10.1111/trf.13896>

Abbas, A. R., Wolslegel, K., Seshasayee, D., Modrusan, Z., & Clark, H. F. (2009). Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. *PloS One*, *4*(7), e6098. <http://doi.org/10.1371/journal.pone.0006098>

Ach, R. A., Wang, H., & Curry, B. (2008). Measuring microRNAs: Comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. *BMC Biotechnology*, *8*(1), 1–16. <http://doi.org/10.1186/1472-6750-8-69>

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. <http://doi.org/10.1109/TAC.1974.1100705>

Allcock, R. J., Forrest, I., Corris, P. A., Crook, P. R., & Griffiths, I. D. (2004). A study of the prevalence of systemic sclerosis in northeast England. *Rheumatology (Oxford, England)*, *43*(5), 596–602. <http://doi.org/10.1093/rheumatology/keh124>

Augustine, D. X., Coates-Bradshaw, L. D., Willis, J., Harkness, A., Ring, L., Grapsa, J., et al. (2018). Echocardiographic assessment of pulmonary hypertension: a guideline protocol from the British Society of Echocardiography. *Echo Research and Practice*, *5*(3), G11–G24. <http://doi.org/10.1530/ERP-17-0071>

AVOUAC, J., AIRÒ, P., MEUNE, C., Beretta, L., DIEUDE, P., CARAMASCHI, P., et al. (2010). Prevalence of pulmonary hypertension in systemic sclerosis in European Caucasians and metaanalysis of 5 studies. *The Journal of Rheumatology*, *37*(11), 2290–2298. <http://doi.org/10.3899/jrheum.100245>

Benjamini, Y., & Hochberg, Y. (2010). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B*, 1–13.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59. <http://doi.org/10.1038/nature07517>

Bertero, T., Cottrill, K., Krauszman, A., Lu, Y., Annis, S., Hale, A., et al. (2015). The MicroRNA-130/301 Family Controls Vasoconstriction in Pulmonary Hypertension. *Journal of Biological Chemistry*, *290*(4), 2069–2085. <http://doi.org/10.1074/jbc.M114.617845>

Bhattacharyya, S., Balakathiresan, N. S., Dalgard, C., Gutti, U., Armistead, D., Jozwik, C., et al. (2011). Elevated miR-155 promotes inflammation in cystic fibrosis by driving hyperexpression of interleukin-8. *Journal of Biological Chemistry*, *286*(13), 11604–11615. <http://doi.org/10.1074/jbc.M110.198390>

Billings, C. G., Hurdman, J. A., Condliffe, R., Elliot, C. A., Smith, I. A., Austin, M., et al. (2017). Incremental shuttle walk test distance and autonomic dysfunction predict survival in pulmonary arterial hypertension. *The Journal of Heart and Lung Transplantation*, *36*(8), 871–879. <http://doi.org/10.1016/j.healun.2017.04.008>

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. <http://doi.org/10.1023/A:1010933404324>

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees, Belmont, California: Wadsworth. *Biometrics*.

Brunet-Vega, A., Pericay, C., Quílez, M. E., Ramírez-Lázaro, M. J., Calvet, X., & Lario, S. (2015). Variability in microRNA recovery from plasma: Comparison of five commercial kits. *Analytical Biochemistry*, *488*(C), 28–35. <http://doi.org/10.1016/j.ab.2015.07.018>

Budas, G. R., Boehm, M., Kojonazarov, B., Viswanathan, G., Tian, X., Veeroju, S., et al. (2018). ASK1 Inhibition Halts Disease Progression in Preclinical Models of Pulmonary Arterial Hypertension. *American Journal of Respiratory and Critical Care Medicine*, *197*(3), 373–385. <http://doi.org/10.1164/rccm.201703-0502OC>

Callari, M., Tiberio, P., De Cecco, L., Cavadini, E., Dugo, M., Ghimenti, C., et al. (2013). Feasibility of circulating miRNA microarray analysis from archival plasma samples. *Analytical Biochemistry*, *437*(2), 123–125. <http://doi.org/10.1016/j.ab.2013.03.002>

Caserta, S., Kern, F., Cohen, J., Drage, S., Newbury, S. F., & Llewelyn, M. J. (2016). Circulating Plasma microRNAs can differentiate Human Sepsis and Systemic Inflammatory Response Syndrome (SIRS). *Scientific Reports*, *6*, 28006. <http://doi.org/10.1038/srep28006>

Cheng, C., Wang, Q., You, W., Chen, M., & Xia, J. (2014). MiRNAs as Biomarkers of Myocardial Infarction: A Meta-Analysis. *PloS One*, *9*(2), e88566. <http://doi.org/10.1371/journal.pone.0088566>

Chester, A. H., Yacoub, M. H., & Moncada, S. (2017). Nitric oxide and pulmonary arterial hypertension. *Global Cardiology Science & Practice*, *2017*(2), 14. <http://doi.org/10.21542/gcsp.2017.14>

Chun, H. J., Bonnet, S., & Chan, S. Y. (2017). Translational Advances in the Field of Pulmonary Hypertension. Translating MicroRNA Biology in Pulmonary Hypertension. It Will Take More Than “miR” Words. *American Journal of Respiratory and Critical Care Medicine*, *195*(2), 167–178. <http://doi.org/10.1164/rccm.201604-0886PP>

Coghlan, J. G., Denton, C. P., Grünig, E., Bonderman, D., Distler, O., Khanna, D., et al. (2014). 23687283. *Annals of the Rheumatic Diseases*, *73*(7), 1340–1349. <http://doi.org/10.1136/annrheumdis-2013-203301>

Cooper, S., & Linch, M. (2016). Prostate cancer. *InnovAiT: Education and Inspiration for General Practice*, *9*(5), 275–283. <http://doi.org/10.1177/1755738016639070>

Courboulin, A., Paulin, R., Giguere, N. J., Saksouk, N., Perreault, T., Meloche, J., et al. (2011). Role for miR-204 in human pulmonary arterial hypertension. - PubMed - NCBI. *Journal of Experimental Medicine*, *208*(3), 535–548. <http://doi.org/10.1084/jem.20101812>

Danielson, K. M., Shah, R., Yeri, A., Liu, X., Camacho Garcia, F., Silverman, M., et al. (2018). Plasma Circulating Extracellular RNAs in Left Ventricular Remodeling Post-Myocardial Infarction. *EBioMedicine*, *32*, 172–181. <http://doi.org/10.1016/j.ebiom.2018.05.013>

Donati, S., Ciuffi, S., & Brandi, M. L. (2019). Human Circulating miRNAs Real-time qRT-PCR-based Analysis: An Overview of Endogenous Reference Genes Used for Data Normalization. *International Journal of Molecular Sciences*, *20*(18), 4353–19. <http://doi.org/10.3390/ijms20184353>

Dorin, J. R., Dickinson, P., Alton, E. W., Smith, S. N., Geddes, D. M., Stevenson, B. J., et al. (1992). Cystic fibrosis in the mouse by targeted insertional mutagenesis. *Nature*, *359*(6392), 211–215. <http://doi.org/10.1038/359211a0>

El-Mogy, M., Lam, B., Haj-Ahmad, T. A., McGowan, S., Yu, D., Nosal, L., et al. (2018). Diversity and signature of small RNA in different bodily fluids using next generation sequencing. *BMC Genomics*, *19*(1), 1229. <http://doi.org/10.1186/s12864-018-4785-8>

Fijalkowska, A., Kurzyna, M., Torbicki, A., Szewczyk, G., Florczyk, M., Pruszczyk, P., & Szturmowicz, M. (2006). Serum N-terminal brain natriuretic peptide as a prognostic parameter in patients with pulmonary hypertension. *Chest*, *129*(5), 1313–1321. <http://doi.org/10.1378/chest.129.5.1313>

Flynt, A. S., & Lai, E. C. (2008). Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nature Reviews Genetics*, *9*(11), 831–842. <http://doi.org/10.1038/nrg2455>

Furuse, Y., Finethy, R., Saka, H. A., Xet-Mull, A. M., Sisk, D. M., Smith, K. L. J., et al. (2014). Search for MicroRNAs Expressed by Intracellular Bacterial Pathogens in Infected Mammalian Cells. *PloS One*, *9*(9), e106434. <http://doi.org/10.1371/journal.pone.0106434>

Galiè, N., Hoeper, M. M., Humbert, M., Torbicki, A., Vachiery, J.-L., Barbera, J. A., et al. (2009, October). Guidelines for the diagnosis and treatment of pulmonary hypertension: the 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 the International Society of Heart and Lung Transplantation (ISHLT). *European Heart Journal*. <http://doi.org/10.1093/eurheartj/ehp297>

Galiè, N., Humbert, M., Vachiery, J.-L., Gibbs, S., Lang, I., Torbicki, A., et al. (2016, January 1). 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). *European Heart Journal*. <http://doi.org/10.1093/eurheartj/ehv317>

Gao, Z.-J., Yuan, W.-D., Yuan, J.-Q., Yuan, K., & Wang, Y. (2018). miR-486-5p functions as an oncogene by targeting PTEN in non-small cell lung cancer. *Pathology - Research and Practice*, *214*(5), 700–705. <http://doi.org/10.1016/j.prp.2018.03.013>

Glinge, C., Clauss, S., Boddum, K., Jabbari, R., Jabbari, J., Risgaard, B., et al. (2017). Stability of Circulating Blood-Based MicroRNAs - Pre-Analytic Methodological Considerations. *PloS One*, *12*(2), e0167969. <http://doi.org/10.1371/journal.pone.0167969>

Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., et al. (2011). Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. *PloS One*, *6*(11), e27156. <http://doi.org/10.1371/journal.pone.0027156>

Grasso, M., Piscopo, P., Talarico, G., Ricci, L., Crestini, A., Tosto, G., et al. (2019). Plasma microRNA profiling distinguishes patients with frontotemporal dementia from healthy subjects. *Neurobiology of Aging*, 1–12. <http://doi.org/10.1016/j.neurobiolaging.2019.01.024>

Hachulla, É., & Launay, D. (2011). Diagnosis and classification of systemic sclerosis. *Clinical Reviews in Allergy & Immunology*, *40*(2), 78–83. <http://doi.org/10.1007/s12016-010-8198-y>

Hachulla, É., Gressin, V., Guillevin, L., Carpentier, P., Diot, E., Sibilia, J., et al. (2005). Early detection of pulmonary arterial hypertension in systemic sclerosis: a French nationwide prospective multicenter study. *Arthritis and Rheumatism*, *52*(12), 3792–3800. <http://doi.org/10.1002/art.21433>

Hanley, J. A., Radiology, B. M., 1982. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Pubs.Rsna.org*

, *143*(1), 29–36. <http://doi.org/10.1148/radiology.143.1.7063747>

Hannafon, B. N., Trigoso, Y. D., Calloway, C. L., Zhao, Y. D., Lum, D. H., Welm, A. L., et al. (2016). Plasma exosome microRNAs are indicative of breast cancer. *Breast Cancer Research : BCR*, *18*(1), 90. <http://doi.org/10.1186/s13058-016-0753-x>

Hemnes, A. R., Trammell, A. W., Archer, S. L., Rich, S., Yu, C., Nian, H., et al. (2015). Peripheral blood signature of vasodilator-responsive pulmonary arterial hypertension. *Circulation*, *131*(4), 401–9– discussion 409. <http://doi.org/10.1161/CIRCULATIONAHA.114.013317>

Humbert, M., Morrell, N. W., Archer, S. L., Stenmark, K. R., MacLean, M. R., Lang, I. M., et al. (2004). Cellular and molecular pathobiology of pulmonary arterial hypertension. *Journal of the American College of Cardiology*, *43*(12 Suppl S), 13S–24S. <http://doi.org/10.1016/j.jacc.2004.02.029>

Humbert, M., Yaici, A., de Groote, P., Montani, D., Sitbon, O., Launay, D., et al. (2011). Screening for pulmonary arterial hypertension in patients with systemic sclerosis: clinical characteristics at diagnosis and long-term survival. *Arthritis and Rheumatism*, *63*(11), 3522–3530. <http://doi.org/10.1002/art.30541>

Hurdman, J., Condliffe, R., Elliot, C. A., Davies, C., Hill, C., Wild, J. M., et al. (2012). ASPIRE registry: Assessing the Spectrum of Pulmonary hypertension Identified at a REferral centre. *European Respiratory Journal*, *39*(4), 945–955. <http://doi.org/10.1183/09031936.00078411>

International PPH Consortium, Lane, K. B., Machado, R. D., Pauciulo, M. W., Thomson, J. R., Phillips, J. A., et al. (2000). Heterozygous germline mutations in BMPR2, encoding a TGF-beta receptor, cause familial primary pulmonary hypertension. *Nature Genetics*, *26*(1), 81–84. <http://doi.org/10.1038/79226>

Izumiya, Y., Jinnn, M., Kimura, Y., Wang, Z., Onoue, Y., Hanatani, S., et al. (2015). Expression of Let-7 family microRNAs in skin correlates negatively with severity of pulmonary hypertension in patients with systemic scleroderma. *International Journal of Cardiology. Heart & Vasculature*, *8*, 98–102. <http://doi.org/10.1016/j.ijcha.2015.06.006>

J, M., F, P., M, V., A, B., I, J., L, D., et al. (2015). Bromodomain-Containing Protein 4: The Epigenetic Origin of Pulmonary Arterial Hypertension. *Circulation Research*, *117*(6), 525–535. <http://doi.org/10.1161/CIRCRESAHA.115.307004>

JE, I., X, Z., Rangaraj, V, S, M., & H, L. (2019). Microarray profiling identifies extracellular circulating miRNAs dysregulated in cystic fibrosis. *Scientific Reports*, *9*(1), 1585. <http://doi.org/10.1038/s41598-019-51890-7>

Ji, X., Wu, B., Fan, J., Han, R., Luo, C., Wang, T., et al. (2015). The Anti-fibrotic Effects and Mechanisms of MicroRNA-486-5p in Pulmonary Fibrosis. *Scientific Reports*, *5*(1), 394. http://doi.org/10.1038/srep14131

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., et al. (1989). Identification of the cystic fibrosis gene: genetic analysis. *Science (New York, N.Y.)*, *245*(4922), 1073–1080. <http://doi.org/10.1126/science.2570460>

Khalyfa, A., Kheirandish-Gozal, L., Khalyfa, A. A., Philby, M. F., Alonso-Álvarez, M. L., Mohammadi, M., et al. (2016). Circulating Plasma Extracellular Microvesicle MicroRNA Cargo and Endothelial Dysfunction in Children with Obstructive Sleep Apnea. *American Journal of Respiratory and Critical Care Medicine*, *194*(9), 1116–1126. <http://doi.org/10.1164/rccm.201602-0323OC>

Kozomara, A., Birgaoanu, M., & Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Research*, *47*(D1), D155–D162. <http://doi.org/10.1093/nar/gky1141>

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the BorutaPackage. *Journal of Statistical Software*, *36*(11). <http://doi.org/10.18637/jss.v036.i11>

L, C., O, B., M, H., N, D., AM, H., R, G., et al. (2011). A novel murine model of severe pulmonary arterial hypertension. *American Journal of Respiratory and Critical Care Medicine*, *184*(10), 1171–1182. <http://doi.org/10.1164/rccm.201103-0412OC>

L, S., LH, R., WD, J., R, S., JA, W., SC, B., et al. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, *24*(9), 1151–1161. <http://doi.org/10.1038/nbt1239>

Lai, Y.-C., Potoka, K. C., Champion, H. C., Mora, A. L., & Gladwin, M. T. (2014). Pulmonary arterial hypertension: the clinical syndrome. *Circulation Research*, *115*(1), 115–130. <http://doi.org/10.1161/CIRCRESAHA.115.301146>

Lawrie, A. (2014). The role of the osteoprotegerin/tumor necrosis factor related apoptosis-inducing ligand axis in the pathogenesis of pulmonary arterial hypertension. *Vascular Pharmacology*, *63*(3), 114–117. <http://doi.org/10.1016/j.vph.2014.10.002>

Lee, R., Feinbaum, R., & Ambros, V. (2004). A short history of a short RNA. *Cell*, *116*(2 Suppl), S89–S92. <http://doi.org/10.1016/S0092-8674(04)00035-2>

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*, *28*(6), 882–883. <http://doi.org/10.1093/bioinformatics/bts034>

Li, Yanli, Wang, D., Li, X., Shao, Y., He, Y., Yu, H., & Ma, Z. (n.d.). MiR-199a-5p suppresses non-small cell lung cancer via targeting MAP3K11. *Journal of Cancer*, *10*(11), 2472–2479. <http://doi.org/10.7150/jca.29426>

Li, Yu, & Kowdley, K. V. (2012). MicroRNAs in Common Human Diseases. *Genomics, Proteomics & Bioinformatics*, *10*(5), 246–253. <http://doi.org/10.1016/j.gpb.2012.07.005>

López-Romero, P. (2011). Pre-processing and differential expression analysis of Agilent micr... - PubMed - NCBI. *BMC Genomics*, *12*(1), 64. <http://doi.org/10.1186/1471-2164-12-64>

M, L.-Q., R, R., A, Y., J, M., W, L., & T, T. (2002). Identification of tissue-specific microRNAs from mouse. *Current Biology : CB*, *12*(9), 735–739. <http://doi.org/10.1016/s0960-9822(02)00809-6>

Ma, L., Roman-Campos, D., Austin, E. D., Eyries, M., Sampson, K. S., Soubrier, F., et al. (2013). A Novel Channelopathy in Pulmonary Arterial Hypertension. *The New England Journal of Medicine*, *369*(4), 351–361. <http://doi.org/10.1056/NEJMoa1211097>

MacIntyre, N., Crapo, R. O., Viegi, G., Johnson, D. C., van der Grinten, C. P. M., Brusasco, V., et al. (2005). Standardisation of the single-breath determination of carbon monoxide uptake in the lung. *European Respiratory Journal*, *26*(4), 720–735. <http://doi.org/10.1183/09031936.05.00034905>

Malhotra, R., Hess, D., Lewis, G. D., Bloch, K. D., Waxman, A. B., & Semigran, M. J. (2011). Vasoreactivity to Inhaled Nitric Oxide with Oxygen Predicts Long-Term Survival in Pulmonary Arterial Hypertension. *Pulmonary Circulation*, *1*(2), 250–258. <http://doi.org/10.4103/2045-8932.83449>

Manimaran, S., Selby, H. M., Okrah, K., Ruberman, C., Leek, J. T., Quackenbush, J., et al. (2016). BatchQC: interactive software for evaluating sample and batch effects in genomic data. *Bioinformatics (Oxford, England)*, *32*(24), 3836–3838. <http://doi.org/10.1093/bioinformatics/btw538>

Matin, F., Jeet, V., Moya, L., Selth, L. A., Chambers, S., Clements, J. A., & Batra, J. (2018). A Plasma Biomarker Panel of Four MicroRNAs for the Diagnosis of Prostate Cancer. *Scientific Reports*, *8*(1), 1–15. http://doi.org/10.1038/s41598-018-24424-w

Maués, J. H. D. S., Moreira-Nunes, C. de F. A., & Burbano, R. M. R. (2020). Computational Identification and Characterization of New microRNAs in Human Platelets Stored in a Blood Bank. *Biomolecules*, *10*(8), 1173. <http://doi.org/10.3390/biom10081173>

McHugh, D. R., Steele, M. S., Valerio, D. M., Miron, A., Mann, R. J., LePage, D. F., et al. (2018). A G542X cystic fibrosis mouse model for examining nonsense mutation directed therapies. *PloS One*, *13*(6), e0199573. <http://doi.org/10.1371/journal.pone.0199573>

Mestdagh, P., Hartmann, N., Baeriswyl, L., Andreasen, D., Bernard, N., Chen, C., et al. (2014). Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nature Methods*, *11*(8), 809–815. <http://doi.org/10.1038/nmeth.3014>

Mitchell, A. J., Gray, W. D., Hayek, S. S., Ko, Y.-A., Thomas, S., Rooney, K., et al. (2016). Platelets confound the measurement of extracellular miRNA in archived plasma. *Scientific Reports*, *6*(1), 2202. <http://doi.org/10.1038/srep32651>

Moll, M., Christmann, R. B., Zhang, Y., Whitfield, M. L., Wang, Y. M., Rice, L., et al. (2018). Patients with systemic sclerosis-associated pulmonary arterial hypertension express a genomic signature distinct from patients with interstitial lung disease. *Journal of Scleroderma and Related Disorders*, *3*(3), 242–248. <http://doi.org/10.1177/2397198318764780>

Montani, D., Günther, S., Dorfmüller, P., Perros, F., Girerd, B., Garcia, G., et al. (2013). Pulmonary arterial hypertension. *Orphanet Journal of Rare Diseases*, *8*(1), 97. <http://doi.org/10.1186/1750-1172-8-97>

Mukerjee, D., St George, D., Knight, C., Davar, J., Wells, A. U., Bois, Du, R. M., et al. (2004). Echocardiography and pulmonary function as screening tests for pulmonary arterial hypertension in systemic sclerosis. *Rheumatology (Oxford, England)*, *43*(4), 461–466. <http://doi.org/10.1093/rheumatology/keh067>

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, *12*(5), 453–457. <http://doi.org/10.1038/nmeth.3337>

NW, M., X, Y., PD, U., KB, J., N, M., KK, S., & RC, T. (2001). Altered growth responses of pulmonary artery smooth muscle cells from patients with primary pulmonary hypertension to transforming growth factor-beta(1) and bone morphogenetic proteins. *Circulation*, *104*(7), 790–795. <http://doi.org/10.1161/hc3201.094152>

Oglesby, I. K., & McKiernan, P. J. (2017). MiRNA Expression in Cystic Fibrosis Bronchial Epithelial Cells. *Methods in Molecular Biology (Clifton, N.J.)*, *1509*, 57–69. <http://doi.org/10.1007/978-1-4939-6524-3_7>

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., & Lai, E. C. (2007). The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila. *Cell*, *130*(1), 89–100. <http://doi.org/10.1016/j.cell.2007.06.028>

Parikh, V. N., Jin, R. C., Rabello, S., Gulbahce, N., White, K., Hale, A., et al. (2012). MicroRNA-21 Integrates Pathogenic Signaling to Control Pulmonary Hypertension: Results of a Network Bioinformatics Approach. *Circulation*, *125*(12), 1520–1532. <http://doi.org/10.1161/CIRCULATIONAHA.111.060269>

Pirola, C. J., Fernández Gianotti, T., Castaño, G. O., Mallardi, P., San Martino, J., Mora Gonzalez Lopez Ledesma, M., et al. (2015). Circulating microRNA signature in non-alcoholic fatty liver disease: from serum non-coding RNAs to liver histology and disease pathogenesis. *Gut*, *64*(5), 800–812. <http://doi.org/10.1136/gutjnl-2014-306996>

Qin, Y., Yao, J., Wu, D. C., Nottingham, R. M., Mohr, S., Hunicke-Smith, S., & Lambowitz, A. M. (2015). High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *Rna-a Publication of the Rna Society*, *22*(1), 111–128. <http://doi.org/10.1261/rna.054809.115>

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., et al. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341. <http://doi.org/10.1186/1471-2164-13-341>

Rabinovitch, M. (2012). Molecular pathogenesis of pulmonary arterial hypertension. - PubMed - NCBI. *Journal of Clinical Investigation*, *122*(12), 4306–4313. <http://doi.org/10.1172/JCI60658>

Radchenko, G. D., Zhyvylo, I. O., & Sirenko, Y. M. (2019). Analysis of pulmonary hypertension patient survival after treatment in referral center (data of first Ukrainian register). *Pulmonary Circulation*, *9*(2), 204589401984560. <http://doi.org/10.1177/2045894019845604>

Reyfman, P. A., Walter, J. M., Joshi, N., Anekalla, K. R., McQuattie-Pimentel, A. C., Chiu, S., et al. (2019). Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis. *American Journal of Respiratory and Critical Care Medicine*, *199*(12), 1517–1536. <http://doi.org/10.1164/rccm.201712-2410OC>

Rhodes, C. J., Wharton, J., Boon, R. A., Roexe, T., Tsang, H., Wojciak-Stothard, B., et al. (2013). Reduced microRNA-150 is associated with poor survival in pulmonary arterial hypertension. *American Journal of Respiratory and Critical Care Medicine*, *187*(3), 294–302. <http://doi.org/10.1164/rccm.201205-0839OC>

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. <http://doi.org/10.1093/nar/gkv007>

Roderburg, C., Luedde, M., Vargas Cardenas, D., Vucur, M., Scholten, D., Frey, N., et al. (2013). Circulating microRNA-150 serum levels predict survival in patients with critical illness and sepsis. *PloS One*, *8*(1), e54612. <http://doi.org/10.1371/journal.pone.0054612>

Rothman, A. M. K., Arnold, N. D., Pickworth, J. A., Iremonger, J., Ciuclan, L., Allen, R. M. H., et al. (2016). MicroRNA-140-5p and SMURF1 regulate pulmonary arterial hypertension. *Journal of Clinical Investigation*, *126*(7), 2495–2508. <http://doi.org/10.1172/JCI83361>

Ruby, J. G., Jan, C. H., & Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. *Nature*, *448*(7149), 83–86. <http://doi.org/10.1038/nature05983>

Saliminejad, K., Khorram Khorshid, H. R., & Ghaffari, S. H. (2019). Why have microRNA biomarkers not been translated from bench to clinic? *Future Oncology*, *15*(8), 801–803. <http://doi.org/10.2217/fon-2018-0812>

Sanada, T. J., Tanabe, N., Ishibashi-Ueda, H., Ishida, K., Naito, A., Sakao, S., et al. (2019). Involvement of pulmonary arteriopathy in the development and severity of reperfusion pulmonary edema after pulmonary endarterectomy. *Pulmonary Circulation*, *9*(2), 2045894019846439. <http://doi.org/10.1177/2045894019846439>

Sanders, J. L., Han, Y., Urbina, M. F., Systrom, D. M., & Waxman, A. B. (2019). Metabolomics of exercise pulmonary hypertension are intermediate between controls and patients with pulmonary arterial hypertension. *Pulmonary Circulation*, *9*(4), 2045894019882623. <http://doi.org/10.1177/2045894019882623>

Sarrion, I., Milian, L., Juan, G., Ramon, M., Furest, I., Carda, C., et al. (2015). Role of Circulating miRNAs as Biomarkers in Idiopathic Pulmonary Arterial Hypertension: Possible Relevance of miR-23a. - PubMed - NCBI. *Oxidative Medicine and Cellular Longevity*, *2015*(20), 1–10. <http://doi.org/10.1155/2015/792846>

Sato, F., Tsuchiya, S., Terasawa, K., & Tsujimoto, G. (2009). Intra-Platform Repeatability and Inter-Platform Comparability of MicroRNA Microarray Technology. *PloS One*, *4*(5), e5540. <http://doi.org/10.1371/journal.pone.0005540>

Schlosser, K., McIntyre, L. A., White, R. J., & Stewart, D. J. (2015). Customized Internal Reference Controls for Improved Assessment of Circulating MicroRNAs in Disease. *PloS One*, *10*(5), e0127443–22. <http://doi.org/10.1371/journal.pone.0127443>

Sehgal, S., Chowdhury, A., Rabih, F., Gadre, A., Park, M. M., Li, M., et al. (2019). Counting Steps: A New Way to Monitor Patients with Pulmonary Arterial Hypertension. *Lung*, *197*(4), 501–508. <http://doi.org/10.1007/s00408-019-00239-y>

Shanmugam, E., Jena, A., & George, M. (2015). Riociguat: Something new in pulmonary hypertension therapeutics? *Journal of Pharmacology & Pharmacotherapeutics*, *6*(1), 3–6. <http://doi.org/10.4103/0976-500X.149132>

Simonneau, G., Gatzoulis, M. A., Adatia, I., Celermajer, D., Denton, C., Ghofrani, A., et al. (2013). Updated clinical classification of pulmonary hypertension. *Journal of the American College of Cardiology*, *62*(25 Suppl), D34–41. <http://doi.org/10.1016/j.jacc.2013.10.029>

Simonneau, G., Montani, D., Celermajer, D. S., Denton, C. P., Gatzoulis, M. A., Krowka, M., et al. (2019). Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *European Respiratory Journal*, *53*(1), 1801913. <http://doi.org/10.1183/13993003.01913-2018>

Simonneau, G., Robbins, I. M., Beghetti, M., Channick, R. N., Delcroix, M., Denton, C. P., et al. (2009). Updated Clinical Classification of Pulmonary Hypertension. *Journal of the American College of Cardiology*, *54*(1s1), S43–S54. <http://doi.org/10.1016/j.jacc.2009.04.012>

Srinivasan, S., Treacy, R., Herrero, T., Olsen, R., Leonardo, T. R., Zhang, X., et al. (2020a). Discovery and Verification of Extracellular miRNA Biomarkers for Non-invasive Prediction of Pre-eclampsia in Asymptomatic Women. *Cell Reports. Medicine*, *1*(2), 100013–100013. <http://doi.org/10.1016/j.xcrm.2020.100013>

Srinivasan, S., Treacy, R., Herrero, T., Reports, R. O. C., 2020. (2020b). Discovery and Verification of Extracellular miRNA Biomarkers for Non-invasive Prediction of Pre-eclampsia in Asymptomatic Women. *Elsevier*

, *1*(2), 100013. <http://doi.org/10.1016/j.xcrm.2020.100013>

T, A., M, N., Y, O., K, N., G, H., Y, G., et al. (2010). Plasma microRNA 499 as a biomarker of acute myocardial infarction. *Clinical Chemistry*, *56*(7), 1183–1185. <http://doi.org/10.1373/clinchem.2010.144121>

Tang, P. (2018). Clinical diagnostic value of circulating serum miR-509-3p in pulmonary arterial hypertension with congenital heart disease. *Hellenic Journal of Cardiology : HJC = Hellenike Kardiologike Epitheorese*. <http://doi.org/10.1016/j.hjc.2018.06.004>

Thakral, S., & Ghoshal, K. (2015). miR-122 is a unique molecule with great potential in diagnosis, prognosis of liver disease, and therapy both as miRNA mimic and antimir. *Current Gene Therapy*, *15*(2), 142–150. <http://doi.org/10.2174/1566523214666141224095610>

Thygesen, K., Mair, J., Katus, H., Plebani, M., Venge, P., Collinson, P., et al. (2010). Recommendations for the use of cardiac troponin measurement in acute cardiac care. *European Heart Journal*, *31*(18), 2197–2204. <http://doi.org/10.1093/eurheartj/ehq251>

Tian, F., Wang, J., Ouyang, T., Lu, N., Lu, J., Shen, Y., et al. (2019). MiR-486-5p Serves as a Good Biomarker in Nonsmall Cell Lung Cancer and Suppresses Cell Growth With the Involvement of a Target PIK3R1. *Frontiers in Genetics*, *10*, 673. <http://doi.org/10.3389/fgene.2019.00688>

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*  <http://doi.org/10.2307/2346178>

Trip, P., Nossent, E. J., de Man, F. S., van den Berk, I. A. H., Boonstra, A., Groepenhoff, H., et al. (2013). Severely reduced diffusion capacity in idiopathic pulmonary arterial hypertension: patient characteristics and treatment responses. *The European Respiratory Journal*, *42*(6), 1575–1585. <http://doi.org/10.1183/09031936.00184412>

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064. <http://doi.org/10.1080/10629360600810434>

Vasilescu, C., Rossi, S., Shimizu, M., Tudor, S., Veronese, A., Ferracin, M., et al. (2009). MicroRNA Fingerprints Identify miR-150 as a Plasma Prognostic Marker in Patients with Sepsis. *PloS One*, *4*(10), e7405. <http://doi.org/10.1371/journal.pone.0007405>

Voelkel, N. F., Cool, C., Lee, S. D., Wright, L., Geraci, M. W., & Tuder, R. M. (1998). Primary Pulmonary Hypertension Between Inflammation and Cancer. *Chest*, *114*(3), 225S–230S. <http://doi.org/10.1378/chest.114.3_Supplement.225S>

Wang, H., Ach, R. A., & Curry, B. (2007). Direct and sensitive miRNA profiling from low-input total RNA. *Rna-a Publication of the Rna Society*, *13*(1), 151–159. <http://doi.org/10.1261/rna.234507>

Warwick, G., Thomas, P. S., & Yates, D. H. (2008). Biomarkers in pulmonary hypertension. *The European Respiratory Journal*, *32*(2), 503–512. <http://doi.org/10.1183/09031936.00160307>

Weber, J. A., Baxter, D. H., Zhang, S., Huang, D. Y., How Huang, K., Jen Lee, M., et al. (2010). The microRNA spectrum in 12 body fluids. - PubMed - NCBI. *Clinical Chemistry*, *56*(11), 1733–1741. <http://doi.org/10.1373/clinchem.2010.147405>

Wei, Y., Schober, A., & Weber, C. (2013). Pathogenic arterial remodeling: the good and bad of microRNAs. *American Journal of Physiology - Heart and Circulatory Physiology*, *304*(8), H1050–H1059. <http://doi.org/10.1152/ajpheart.00267.2012>

Williams, M. H., Handler, C. E., Akram, R., Smith, C. J., Das, C., Smee, J., et al. (2006). Role of N-terminal brain natriuretic peptide (N-TproBNP) in scleroderma-associated pulmonary arterial hypertension. *European Heart Journal*, *27*(12), 1485–1494. <http://doi.org/10.1093/eurheartj/ehi891>

Winter, J., Jung, S., Keller, S., Gregory, R. I., & Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nature Cell Biology*, *11*(3), 228–234. <http://doi.org/10.1038/ncb0309-228>

Wu, D., Talbot, C. C., Jr, Liu, Q., Jing, Z.-C., Damico, R. L., Tuder, R., et al. (2016). Identifying microRNAs targeting Wnt/β-catenin pathway in end-stage idiopathic pulmonary arterial hypertension. *Journal of Molecular Medicine*, *94*(8), 1–11. <http://doi.org/10.1007/s00109-016-1426-z>

X, L., Z, F., T, Z., W, C., L, Z., H, L., et al. (2015). Plasma miR-1, miR-208, miR-499 as potential predictive biomarkers for acute myocardial infarction: An independent study of Han population. *Experimental Gerontology*, *72*, 230–238. <http://doi.org/10.1016/j.exger.2015.10.011>

Yang, S., Sui, J., Liu, T., Wu, W., Xu, S., Yin, L., et al. (2019). Expression of miR‐486‐5p and its signiﬁcance in lung squamous cell carcinoma. *Journal of Cellular Biochemistry*, *120*(8), 13912–13923. <http://doi.org/10.1002/jcb.28665>

Yeri, A., Courtright, A., Danielson, K., Hutchins, E., Alsop, E., Carlson, E., et al. (2018). Evaluation of commercially available small RNASeq library preparation kits using low input RNA. *BMC Genomics*, *19*(1), 331. <http://doi.org/10.1186/s12864-018-4726-6>

Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. (2008). Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. *Biometrical Journal. Biometrische Zeitschrift*, *50*(3), 419–430. <http://doi.org/10.1002/bimj.200710415>

Zelniker, T. A., Huscher, D., Vonk Noordegraaf, A., Ewert, R., Lange, T. J., Klose, H., et al. (2018). The 6MWT as a prognostic tool in pulmonary arterial hypertension: results from the COMPERA registry. *Clinical Research in Cardiology : Official Journal of the German Cardiac Society*, *107*(6), 460–470. <http://doi.org/10.1007/s00392-018-1207-5>

Zhou, G., Chen, T., & Raj, J. U. (2015). MicroRNAs in Pulmonary Arterial Hypertension. *American Journal of Respiratory Cell and Molecular Biology*, *52*(2), 139–151. http://doi.org/10.1165/rcmb.2014-0166TR