University of Nottingham
UK | CHINA | MALAYSIA

NEPHRONECTIN AND THE GENETICS OF LUNG FUNCTION VARIABILITY

By

Guoqing Qian BM, MD

Student ID Number: 14342027

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

May 2022

COVID-19 Thesis Impact Statement

I am a Respiratory Physician from China. Because my PhD studies involved a mixture of laboratory based and clinical studies which started before the COVID-19 pandemic, but which continued through the pandemic, there was significant potential for disruption of my research. To help mitigate this, I was fortunate that the Ningbo First Hospital kept my clinical position open for me during the time I was doing my PhD studies in the UK. However, when I came back to China to collect some COPD patients' samples for my PhD research at the end of 2019. I was assigned to participate in the front line of fighting against COVID-19 on the isolation ward in Ningbo city, Zhejiang Province, China. Subsequent travel restrictions prevented me from returning to the UK to undertake some planned functional experiments.

I therefore had to reduce some laboratory work because of a lack of access to laboratory facilities or delays in getting cells/tissue. In my planned research protocol, I was going to study the NPNT contribution to biological mechanisms/ function in the human bronchial epithelial cell by single cell, CRISPR/Cas9 and other methods but this work had to be reduced in scope. However, I was able to undertake some alternative studies by accessing additional clinical samples in China. To ensure I could complete my research I have received a 6-month COVID-19 extension to registered study to 31 March 2022. Despite the pandemic and the inevitable disruption it caused I successfully finished my PhD

thesis under the support and guidance of my supervisor Professor Hall and Professor Sayers.

Abstract

Genetic variation at the chromosome 4 locus containing the gene nephronectin (NPNT) is strongly associated with both lung function variability and the risk of developing chronic obstructive pulmonary disease (COPD) in genome wide association studies (GWAS). NPNT is a conserved extracellular matrix protein, which appears to play a crucial role in regulating cell proliferation, differentiation and adhesion. The main aims of the thesis were i) to investigate the profile of NPNT in different airway structural cell lines and lung tissues; ii) to identify the contribution of targeting NPNT by two approaches (RNA interference, or CRISPR/Cas9) to determine differentially expressed genes and potential pathways involved in human bronchial epithelial cell homeostasis; and iii) to assess the potential for serum NPNT protein levels to be used as a biomarker in either stable COPD or patients with exacerbations.

I found that SNP rs17331332 and rs34712979 are eQTLs for NPNT, and the G allele of variant rs34712979 was negatively associated with COPD risk in a phenome wide association study (PheWAS). NPNT mRNA was identified in foetal and adult lung tissues, as well as human bronchial epithelial cells (HBECs), human airway smooth muscle cells (HASMs) and BEAS-2B cells. NPNT protein was detected in foetal, adult healthy controls and COPD lung tissues by immunohistochemistry. NPNT protein is highly expressed in the

alveolar and fibro-elastic layers, especially in pneumocytes and fibroblasts. NPNT expression gradually increased with increasing foetal lung development, but was decreased in smokers with COPD.

Two approaches to analyse the differentially expressed genes after knockdown of NPNT by siRNA were used Microarrays and RNA sequencing. There were 24 and 442 differentially expressed genes identified by Microarray and RNA sequencing, respectively, but poor agreement between the two platforms. NPNT expression itself was only seen to be significantly reduced using RNA sequencing analysis. Gene set enrichment analysis revealed inflammation pathways were upregulated, and cell proliferation signalling were downregulated by NPNT knockdown. To extend these studies, I successfully established a protocol for fast, robust, cloning-free and high efficiency electroporation-based CRISPR/Cas9 genome editing system for primary HBECs. This will be of value in the future to further assess the downstream genes and pathways regulated by NPNT.

Serum NPNT protein levels in COPD patients were measured by ELISA from a UK (Nottingham) cohort and a China (Ningbo) cohort. Serum NPNT expression levels showed weak positive correlations for age and creatinine blood, but were not associated with lung function. Serum NPNT levels were decreased in patients with COPD exacerbations, and in this group NPNT levels were associated with FEV_1 and FEV_1/FVC.

In summary, the work described in this thesis has identified the potential casual variant associated with lung function decline and

COPD risk, and has defined the expression profiles of NPNT in foetal and adult lung tissue and airway structural cells. NPNT appears to be implicated in HBEC homeostasis. An electroporation based CRISPR/Cas9 genome editing system was established and will be of value for future work. Measuring serum NPNT levels in patients is feasible and may throw further light on the role of NPNT in lung disease.

Publications

Peer reviewed publications

Part I

1. K Fawcett, R Scott, **GQ Qian**, Alina Farmaki, C John, N Shrine, L Wain, M Tobin, IP Hall. Pleiotropic effects of heterozygosity for the SERPINA1 Z allele in the UK Biobank. ERJ Open Research, 2020, 7(2):00049-02021.
2. **GQ Qian**, Y Zhang, JJ Shi. Recent Advances in Nephronectin. Sheng Li Xue Bao. 2019, 71(5):799-805. (Article in Chinese)
3. **GQ Qian**. Advances in genome-wide association study of chronic obstructive pulmonary disease. Hereditas (Beijing). 2020, 42: 832-846. (Article in Chinese)
4. JJ Shi, **GQ Qian (Corresponding author)**, FY Yin, GX Li. Research progress in risk factors in female patients with chronic obstructive pulmonary disease. Journal of Clinical and Pathological Research, 2018,38(6): 1333-1336. (Article in Chinese)

Part II

5. **GQ Qian**, NB Yang, AHY Ma, LP Wang, GX Li, XQ Chen, XM Chen. A COVID-19 Transmission within a family cluster by pre-symptomatic infectors in China. Clinical Infectious Diseases, 2020, 113(7):521-522.
6. **GQ Qian**, NB Yang, F Ding, AHY Ma, ZY Wang, YF Shen, CW Shi, X Lian, JG Chu, L Chen, ZY Wang, DW Ren, GX Li, XQ Chen, HJ Shen, XM Chen. Epidemiologic and Clinical Characteristics of 91

Hospitalized Patients with COVID-19 in Zhejiang, China: A retrospective, multi-centre case series. QJM: An International Journal of Medicine, 2020, 113(7):474-481.

7. NB Yang, YF Shen, CW Shi, AHY Ma, X Zhang, XM Jian, LP Wang, JJ Shi, CY Wu, GX Li, MQ Lu, **GQ Qian (Corresponding author)**. In-flight Transmission Cluster of COVID-19: A Retrospective Case Series. *Infectious Diseases (Lond),* 2020; 52: 891-901.
8. **GQ Qian**, AHY Ma, NB Yang, LM Ruan. Response letter to COIVD-19 Disease: Tackling a pandemic in 21st Century. QJM: An International Journal of Medicine, 2020, 113(7):521-522.
9. **GQ Qian**, X Zhang, AHY Ma, NB Yang. Response letter to Eosinophil count in severe coronavirus disease 2019 (COVID-19). QJM: An International Journal of Medicine*,* 2020, 113(7):513-514.
10. **GQ Qian**, XQ Chen, DF Lv, AHY Ma, LP Wang, NB Yang, XM Chen. Duration of SARS-CoV-2 viral shedding during COVID-19 infection. *Infectious Diseases (Lond),* 2020, 52(7):511-512.
11. CY Wu, XP Yu, AHY Ma, LP Wang, NB Yang, GX Li, JJ Shi, **GQ Qian (Corresponding author)**. COVID-19 with gastrointestinal symptoms as initial manifestations: a case report. Journal of International Medical Research. 2020, 48(9):300060520952256.
12. **Qian GQ**, YW Lin, ADH Ma, ZH Zhang, GX Li, XZ Ruan, XQ Chen, LM Ruan. Early clinical and CT features of COVID-19 and community acquired pneumonia from fever observation ward in

Ningbo, China. Singapore Medical Journal. 2021. doi: 10.11622/smedj.2021004.

13. **GQ Qian**, Y Zhang, Y Xu, WH Hu, IP Hall, J Yue, HY Lu, LM Ruan, MQ Ye, J Mei. Reduced inflammatory responses to SARS-CoV-2 infection in children presenting to hospital with COVID19 in China. EClinicalMedicine. 2021, 34:100831.
14. DF Lv, QM Ying, YW He, J Liang, BB Lu, **GQ Qian**, JG Chu, XB Weng, XQ Chen, QT Mu. Differential diagnosis of coronavirus disease 2019 pneumonia or influenza A pneumonia by clinical characteristics and laboratory findings. *Journal of Clinical Laboratory Analysis,* 2021, 35(2):e23685.
15. C Cao, L He, JP Ma, MP Chen, YT Li, QW Jiang, SY Wu, LL Yu, WN Huang, **GQ Qian**, CB Zhu, JG Chu, XM Chen. Clinical features and predictors for patients with severe SARS-CoV-2 pneumonia at the start of the pandemic: a retrospective multicenter cohort study. BMC Infect Dis. 2021, 21: 666.
16. JJ Shi, NB Yang, **GQ Qian (Corresponding author)**. Case Report: Metagenomic Next-Generation Sequencing in Diagnosis of Talaromycosis of an Immunocompetent Patient. Frontier in Medicine (Lausanne). 2021, 8:656194.
17. YT Ye, Y Yang, JY Zhou, **GQ Qian (Corresponding author)**, JG Chu (Corresponding author). Case Report: Metagenomic Next-Generation Sequencing in Diagnosis of Disseminated Tuberculosis of an Immunocompetent Patient. Frontier in Medicine. 2021, 8:687984.

18. JJ Shi, L Zhou, AHY Ma, NB Yang, L Chen, **GQ Qian (Corresponding author)**. Cryptococcosis as a complication of therapy for sarcoidosis: A case report. SAGE Open Med Case Rep. 2021, 9:2050313X211054268.

Posters

1. **Guoqing Qian**, Amanda Henry, Charlotte Billington, Bo Liu, Binaya KC, Ian Sayers, Ian P Hall. NPNT knock down leads to reveal function in human bronchial epithelial cells. Poster 917. 2019 American Thoracic Society International Conference. Dallas, USA, 2019.05.17-05.23.
2. Katherine A Fawcett, **Guoqing Qian**, Alina Farmaki, Catherine John, Nick Shrine, Louise V Wain, Martin D Tobin, Ian P Hall. Phenotypic consequences of the SERPINA1 Z allele in a population-based setting: A UK Biobank study. Poster, 2019 Meeting of Joint Respiratory Research Day. 29th May 2019. Nottingham, UK.

Oral presentation

1. **GQ Qian**. The expression and regulation of Nephronectin in HBECs. Respiratory Medicine Scientific Meeting, 20th September 2019. Nottingham, UK.
2. **GQ Qian**, I Sayer, I Hall. The expression profile and functional properties of Nephronectin: a novel biomarker for COPD? Sue Watson Presentation Event, 28th May 2021. Nottingham, UK. (Online)

Acknowledgements

On 27th September 2013, a hot and beautiful day, I suddenly received a called from Dr. Fengying Yin, and then I quickly rode my second-hand bike from home to hospital in order to listen to a lecture by Professor Ian Hall, Dean of the Faculty of Medicine University of Nottingham. After that, we communicated through e-mail. In 2014, under the support of Professor Liemin Ruan, Director of Ningbo First Hospital. I went to the Queen's Medical Centre from January to December 2015 as an Academic Visitor to broaden my horizons. I successfully applied for a full-time Ph.D. in 2018 and won the Scholarship of Vice-Chancellor's International Research Excellence from the University of Nottingham. Subsequently, I encountered the COVID-19 pandemic which has impacted on my studies. Fortunately, I finally overcame the difficulties with the help of Professor Hall, my family, and colleagues.

Firstly, I would like to thank my supervisor, Professor Ian Hall for giving these great opportunities for me to study as an Academic Visitor and Ph.D. student in the UK and for his professional support, guidance and for making time for my supervisor meeting through face-to-face meetings or online every month. The supervisor meetings have always filled me with energy, encouragement and passion, and have positively impacted on my life. I would also like to thank my second supervisor, Professor Ian Sayers for his professional support and guidance.

I would like to thank all the staff in Respiratory Medicine for all their help. In particular, I would like to thank Helen Naylor for her administrative support, and thanks to Robert Hall, Amanda Henry, Charlotte Wood, Suzanne Swann, Michael Portelli, Bo Liu, Dhruma Thakker (Dee), Arundhati Dongre-Patel, and Binaya Kc.

I would also like to thank Sally Zhou from Asia business centre, and Michael Yin from the Ningbo campus.

I would like to thank all my colleagues in Ningbo First Hospital, including Lieming Ruan, Guanjing Zhao, Fengying Yin, Guoxiang Li, and Haina Xu.

Finally, I would like to extend my gratitude to my wife Jin Cao, my parents, parents-in-law, and my sons for their great support for my studies in the UK.

Abbreviation

3'UTR	Three Prime Untranslated Region
5'UTR	Five Prime Untranslated Region
AATD	Alpha-1-antitrypsin deficiency
ATS	American Thoracic Society
BEGM	Bronchial Epithelial Growth Media
bp	Base Pairs
CB	Chronic bronchitis
cDNA	Complimentary DNA
Ct	Cycle Threshold
COPD	Chronic Obstructive Pulmonary Disease
COPDGene	COPD Genetic Epidemiology
CSE	Cigarette Smoke Extract
DLCO	Lung Carbon Monoxide Diffusion Function
DMEM	Dulbecco's Modified Eagles's Medium
DMSO	Dimethyl Sulphoxide
DNA	Deoxy Ribose Nucleic Acid
ECM	Extracellular Matrix
EDTA	Ethylenediaminetetraacetic Acid
eGFR	Estimated Glomerular Filtration Rate
ELISA	Enzyme Linked Immunosorbent Assay
eQTL	Expression Quantitative Trait Locus
ERS	European Respiratory Society

ERV	Expiratory Reserve Volume
ECLIPSE	Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points
FEV1	Forced Expiratory Volume in 1 Second
FEV1pp	Forced Expiratory Volume in 1 Second (percentage predicted)
FPKM	Fragments Per Kilobase Million
FVC	Forced Vital Capacity
GenKOLS	Genetics of Chronic Obstructive Lung Disease Study
GEO	Gene Expression Omnibus
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GSEA	Gene Set Enrichment Analysis
GWAS	Genome-Wide Association Studies
HASM	Human Airway Smooth Muscle
HBEC	Human Bronchial Epithelium Cell
IC	Inspiratory Capacity
IRV	Inspiratory Reserve Volume
IVC	Inspiratory Vital Capacity
lncRNA	long non-coding RNA transcripts
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
miRNA	micro RNA
mRNA	message RNA
MMP	Matrix Metalloproteinase

mMRC	Modified British Medical Research Council
MTT	3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide
NETT	National Emphysema Treatment Trial
NGS	Next Generation Sequencing
NPNT	Nephronectin
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDAR	Pre Developed Assay Reagent
PVDF	Polyvinylidene Difluoride
qPCR	Quantitative PCR
RIN	RNA Integrity Number
RISC	RNA induced Silencing Complex
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
SDS	Sodium Dodecyl Sulphate
SNP	Single Nucleotide Polymorphism
STAR	Spliced Transcripts Alignment to a Reference
TLC	Total Lung Capacity
VC	Vital Capacity
VN	Vitronectin
VT	Tidal Volume
UACR	Urinary Albumin Creatinine Ratio
WHO	World Health Organisation

Table of Contents

1. General Introduction

1.1 Introduction

1.1.1 Overview

Chronic obstructive pulmonary disease (COPD) is a global public health challenge and is predicted to be the third leading cause of death in 2022, accounting for 6% of death globally (1, 2). The molecular mechanisms underlying COPD pathogenesis depend upon environmental and genetic risk factors (3). In general, cigarette smoking is the main environmental risk factor for COPD in developed countries. However, only 20.7% and 3.6% of smokers developed moderate and severe COPD (respectively) during a 25 year observation period of the general population (4). It is reported that the risk of COPD is about two to three times higher in smokers in family studies (5). Lung function measures showed heritability estimates of 0.91and 0.77 for force vital capacity (FVC) and forced expiratory volume in 1 s (FEV_1), respectively (6). These observations suggest that environmental risk factors and genetic factors both contribute to the development of COPD together.

In the past 15 years, genetic studies have been helped by the completion of the human genome project and progress in next-generation sequencing, which has resulted in much greater knowledge about variability in the human genome. This in turn has helped Genome-Wide Association Studies (GWAS), which aim to identify variants that are associated with complex traits, and in particular, at identifying associations between single-nucleotide polymorphisms (SNPs) and disease. In 2007, the first GWAS of quantitative lung

function measures was conducted in the Framingham Heart Study (FHS) (7). They reported a SNP in the interleukin 6 receptor (IL6R) on chromosome 1 which was associated with percent predicted forced expiratory flow 25~75% ($FEF_{25\sim75\%}$), and a nonsynonymous coding SNP of glutathione S-transferase omega 2 (GSTO2) on chromosome 10q2.12 (7). In 2009, the first GWAS for COPD identified two SNPs tagging susceptibility loci at the CHRNA3/5 locus and the HHIP locus on chromosome 4 (8). At the same time, an intergenic locus at 4q31 (near HHIP) associated with FEV_1/FVC and COPD was identified (9). The application of GWAS meta-analyses furthered the development of GWAS, by moving from small samples to large scale consortia or biobank approaches using resources such as UK Biobank, and has led to the identification of large number of risk SNPs (10-12).

In 2010, meta-analyses in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium consisting of 20,890 individuals identified eight loci associated with FEV_1/FVC, and one new locus in 4q24 (INTS12-GSTCD-NPNT) associated with FEV_1 (10). This paper identified 46 SNPs located in chromosome 4q24, five SNPs harboured in Nephronectin (NPNT), including the variant with the smallest P value (rs17331332, P=4.00E-10), the others were rs17036341, rs10516529, rs11730660, and rs7660534. A search of on line bio-informatic resources revealed that SNP rs17331332 is associated with NPNT expression of lung, eosphagus gastroesophageal junction, and esophagus muscularis (Open targets Genetics platform, https://genetics.opentargets.org/). In a recent study,

it was reported that NPNT is the only putative causal gene identified both as an expression quantitative trait loci (eQTL) and a protein quantitative trait loci (pQTL) in a meta-analysis of UK Biobank and SpiroMeta datasets (12).

Overall, NPNT is therefore a plausible causative gene in this region requiring investigation. This PhD project set out to investigate the role and regulation of NPNT in human bronchial epithelial cells (HBECs) and COPD.

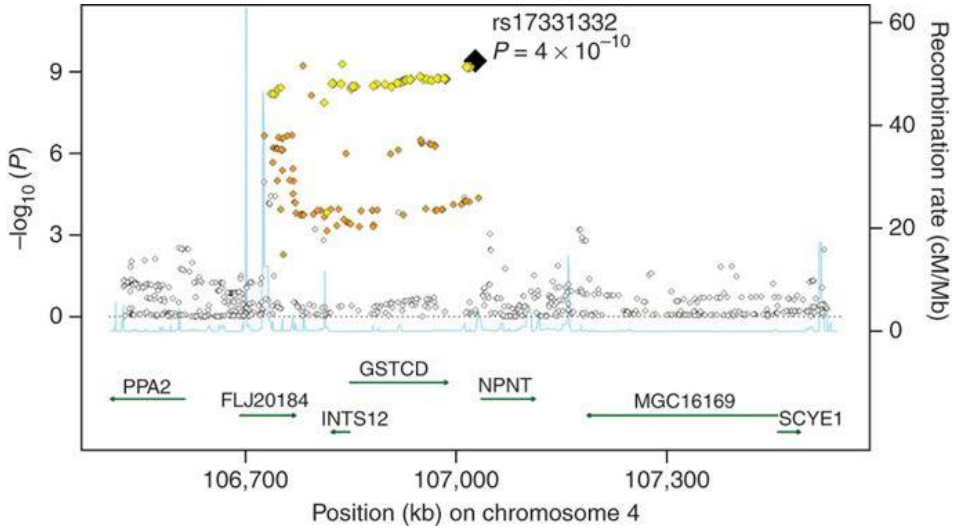


Figure 1-1 Regional association plot of the chromosome 4q24 locus associated with FEV_1 in the CHARGE consortium at genome-wide significance, which include FLJ20184, INTS12, GSTCD and NPNT. Adapted from Hancock (Hancock, 2010)

1.2 Spirometry

Lung function is usually measured by spirometry, which is a well-established physiological test (13). There is a wide range of applications for spirometry including diagnostic, monitoring, disability/impairment evaluations, and public health screening. The vital capacity (VC) is the maximum amount of full inspiration and complete expiration, equal to the sum of inspiratory reserve volume (IRV), tidal volume (V_T) and expiratory reserve volume (ERV). Inspiratory capacity (IC) is the volume consisting of taking a slow full inspiration with no hesitation and the complete expiration, and it equals IRV and V_T (Figure 1-2). FVC is the maximal volume of air exhaled with maximally forced effort from a maximal inspiration. FEV_1 is the maximal volume of air exhaled in the first second of a forced expiration from a position of full inspiration (13).

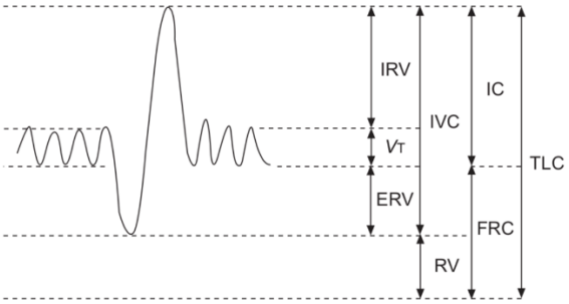


Figure 1-2. Lung volumes and capacities based on a volume-time spirogram of an inspiratory vital capacity (IVC). IRV: inspiratory reserve volume; VT: tidal volume (V_T); ERV: expiratory reserve volume; RV: residual volume; IC: inspiratory capacity; FRC: functional residual capacity; TLC: total lung capacity. Adapted from Miller, 2005.

There are three main types of ventilatory defects, including obstructive abnormalities, restrictive abnormalities, and mixed. The most common diseases of airway obstruction are asthma and COPD, which demonstrate airway narrowing during exhalation. In the early stages of airflow obstruction, the small airways are affected but this only produces small changes on the initial part of the spirogram and the expiratory flow-volume curve shows a concave pattern (Figure 1-3a, b) (14). As more and larger airways are involved the FEV_1 is reduced out of proportion to the reduction in VC (14). The expiratory flow of restrictive ventilatory defect usually shows a convex flow-volume curve (Figure 1-3c), and is typical of diseases such as interstitial lung disease (ILD). The mixed ventilatory defect is defined as coexisting obstruction and restriction (Figure 1-3d).(14).

In the following sections, I will discuss COPD.

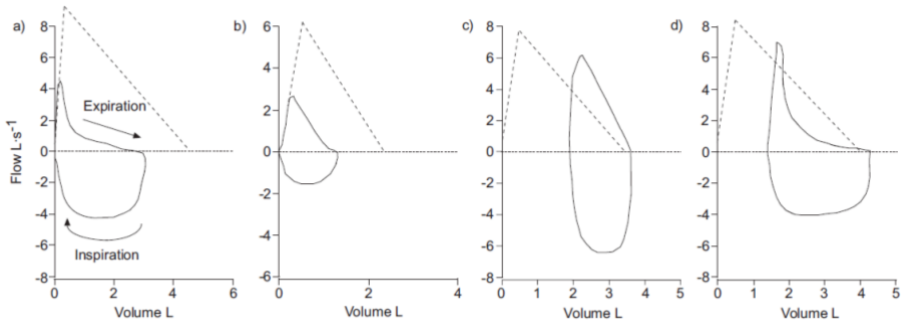


Figure 1-3. Spirometry patterns in respiratory diseases. Example of obstructive ventilatory defect with a low (a) or normal (b) ratio of FEV_1/VC. c) Example of restrictive ventilatory defect. d) Example of mixed defect. Adapted from Miller, 2005.

1.3 Chronic obstructive pulmonary disease (COPD)

1.3.1 Definition, symptoms and assessment

1.3.1.1 Definition

COPD is characterized by the presence of persistent respiratory symptoms and chronic airway inflammation. The gold standard diagnostic criteria for COPD is defined as irreversible airflow limitation with post-bronchodilator $FEV_1/FVC < 0.7$ (15). The diagnosis of COPD however depends on which guidelines are used; examples including the ERS from 1995, the National Institute for Clinical Excellence (NICE) guidelines and the National Heart, Lung, and Blood Institute (NHLBI)/World Health Organisation (WHO) Global Initiative for COPD (GOLD). Currently, the criteria from the various GOLD reports are the most popular in clinical practice worldwide (16).

1.3.1.2 Symptoms

The most common symptom of COPD is breathlessness, which is present in 72.5% patients with moderate and severe COPD (17). Other common symptoms of COPD patients include chronic cough or sputum production, wheezing and chest tightness (15). The symptoms may vary day-to-day during disease progress, morning being the worst time of day (17). Interestingly, a study suggested that the night time symptoms and symptomatic sleep disturbance may occur in over 75% of COPD patients, and are usually unnoticed by physicians or patients themselves (18).

1.3.1.3 Assessment

To assess the level of airflow limitation of COPD, the degree of spirometric abnormality, symptoms, exacerbations, future risk and comorbidities should be considered. Specific spirometric cut-offs are used for classification of severity of airflow limitation, which are divided into 4 stages from GOLD 1 to 4 (Table 1-1) (15). Symptom severity can also be evaluated by the Modified British Medical Research Council (mMRC), COPD Assessment Test (CAT) and St. George's Respiratory Questionnaires (SGRQ). Multiple methods are used to assess COPD, however, the degree of airflow obstruction defined by spirometry correlates partly with symptoms and also correlates partly with exacerbation risk (19). Genetic studies use lung function measurements (FEV_1, FVC, FEV_1/FVC etc.) as quantitative traits to define COPD but this does not consider other COPD phenotypes such as exacerbation frequency.

In patients based on post-bronchodilator $FEV_1/FVC < 0.7$:		
GOLD 1	Mild	$FEV_1 \geq 80\%$ predicted
GOLD 2	Moderate	$50\% \leq FEV_1 < 80\%$ predicted
GOLD 3	Severe	$30\% \leq FEV_1 < 50\%$ predicted
GOLD 4	Very Severe	$FEV_1 < 30\%$ predicted

Table 1-1 Classification of airflow limitation severity in COPD.

1.3.2 Epidemiology and burden of COPD

The prevalence of COPD can be estimated based on diverse definitions, including symptoms, diagnosis made by doctors or spirometry. The prevalence of COPD in five Latin American regions ranged from 7.8% to 19.7% as measured using post-bronchodilator spirometry (20). In China, the overall prevalence of lung function based COPD is 8.6% but in people aged 40 or older, the prevalence is 13.7%, and it is estimated that approximately 99.9 million people have COPD in those aged 20 years or older in 2015 (21). However, it has been suggested that only using post-bronchodilator FEV_1/FVC ratio <0.7 to define COPD risks overdiagnosis, particularly in elder subjects (22, 23).

The economic and social burden of COPD are a major public health problem. The direct costs of COPD every year have been estimated at 38.6 billion Euros and 32 billion dollars in the European Union and United States, respectively (15, 24). In developing countries, the burden of COPD is increasing due to high rates of smoking and air pollution, particularly in Asia. Furthermore, the burden of acute exacerbations of chronic obstructive pulmonary disease (AECOPD) is significantly increasing (25). The Global Burden of Disease (GBD) reported 3.2 million people died from COPD worldwide in 2015, an increase of 11.6% compared with 1990 (26), and in a recent study COPD accounted for 6% of deaths globally (2).

1.3.3 Pathology, pathogenesis and pathophysiology of COPD

The pathologic alterations in COPD consist of lung parenchyma changes, airway alterations, and pulmonary vascular changes (27). Gas trapping and progressive airflow limitation results from parenchymal tissue destruction and small airway fibrosis (15) (Figure **1**-**4**). Emphysema is defined as permanent parenchymal destruction, or destruction of the gas-exchange surfaces of the lung (alveoli) and these often contribute to COPD (15). There are two different types of emphysema, including “centrilobular” and “panlobular”, which can be distinguished on gross examination (28). The pathology of chronic bronchitis demonstrates luminal narrowing resulting from thickened bronchial walls, and mucous blocking the airways. Histologic features of chronic bronchitis are goblet cell hyperplasia, subepithelial basement membrane thickening, bronchial wall fibrosis, and glands enlarging (29).

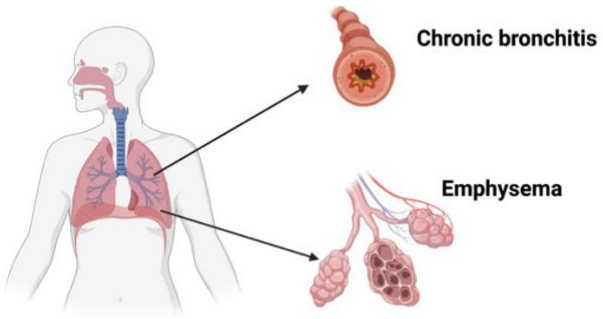


Figure 1-4. Chronic bronchitis and emphysema are the two main types of COPD pathology. This figure was sourced from BioRender (https://app.biorender.com/biorender-templates).

COPD is a chronic inflammatory disease, and hence research has focused heavily on the inflammatory response. Many factors, such as cigarette smoking and air pollution, are associated with persistent airway inflammatory responses and continuing oxidative stress, which has been recognised as a major factor of pathogenesis of COPD (30) (Figure 1-5). Neutrophils, macrophages, T lymphocytes, eosinophils and mast cells have all been shown to be involved in the inflammatory response in chronic bronchitis and COPD (31). Furthermore, biomarkers including cytokines and nitric oxide, which are released from inflammatory cells and reflect the level of systemic inflammation, are often elevated in COPD (32).

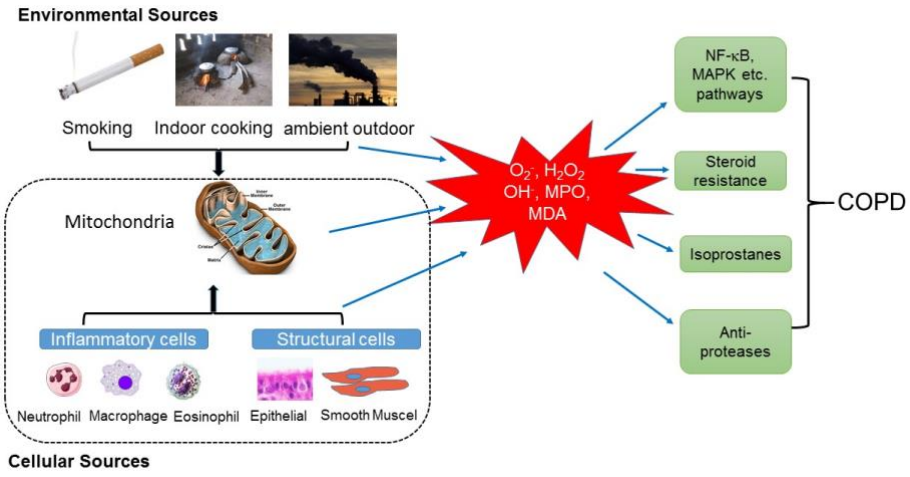


Figure 1-5 Oxidative stress in COPD. A mechanism for the development of COPD driven by oxidative stress from environmental and cellular sources which result in several damaging effects in COPD, including activation of nuclear factor-κB (NF-κB), mitogen-activated protein kinase (MAPK), increased steroid resistance, increased

production of isoprostanes and decreased anti-protease defenses, such as α-1-antitrypsin (AT).

1.3.4 Risk factors

1.3.4.1 Environmental risk factors

Tobacco smoking is usually considered the major environmental risk factor for COPD. However, indoor and outdoor air pollution is also a major risk factor for COPD, particularly in developing countries or areas with poor environmental controls in place (33). Smoking induces an inflammatory response and oxidative stress, which promotes development of COPD. Indoor air pollution in poorly ventilated homes comes mainly from cooking and heating by biomass fuels (33, 34) (Figure 1-5). For example, a recent study reported that the prevalence of COPD increased two-fold in a $PM_{2.5}$ heavy exposure population compared to the non-smoking general population in China. In addition, $PM_{2.5}$ might have greater harmful impact to young adults and the development of lung (21).

1.3.4.2 Genetic factors

Genetic factors play a key role in determining or altering an individual's lung function and the development of COPD. Whilst alpha-1-antitrypsin deficiency (AATD) is a rare but known genetic risk factor for emphysema (35), on the other hand, COPD in general appears to be a complex polygenic disease (36). In the general population, the family studies of early-onset COPD have shown that the first-degree relatives of COPD patients had significantly lower FEV_1 and FEV_1/FVC ratio (5).

Studies of lung function in twins have found more evidence that variability of lung function is determined by different genetic (6). Furthermore, GWASs have been published and identified hundreds of signals associated with lung function, COPD and related phenotypes (37).

1.4 Genetics of lung function and COPD

1.4.1 Alpha-1-antitrypsin deficiency (AATD)

AATD is the only identified monogenic genetic cause of emphysema and is found in 1~2% of all patients with COPD, 3.2% of patients coming to lung transplantation, and accounts for 10% of lung transplants performed for emphysema (38). Alpha-1-antitrypsin is an important anti-protease produced by liver, lung, alveolar macrophages, neutrophils, and gut epithelial cells. The common gene variants are called M, S and Z, and the frequencies are 0.93, 0.05, and 0.02, respectively (39). The MM genotype is considered normal. MZ individuals have reduced α1-AT of about 60% of MM genotype and also show increased risk of COPD. Homozygous ZZ patients have less than 15% of normal levels, and if they smoke often show a rapid decline in FEV1 (40). In addition, PI-MZ ever-smokers have been demonstrated to have a higher risk of COPD compared to never-smokers. The prevalence of PI-ZZ is approximately 1 in 1,600 newborns in Sweden, and 1 in 5,097, 1 in 2,857, and 1 in 3,694 from Oregon, St. Louis, and New York, respectively (41). Furthermore, in our recent study from UK Biobank amongst 303,612 unrelated individuals, 1.135% (N = 107) were homozygous for the SERPINA1 Z allele (PI-ZZ) and 3.91% (N = 11,886) were heterozygous for the Z allele (PI-MZ) (42).

1.4.2 Genome-wide association studies (GWAS)

Lung function and COPD are both heritable traits. Candidate gene and GWAS have identified many genetic variants associated with COPD,

and even more associated with lung function, including FEV_1, FVC, FEV_1/FVC, $FEF_{25\sim75\%}$, and peak expiratory flow (PEF). The use of lung function measures as quantitative traits underlying COPD allows which larger sample sizes to be used as well as providing increased statistical power. In 2007, the first GWAS of quantitative lung function was published from the FHS, and then several additional GWASs investigating lung function and COPD were published during the next 15 years. One GWAS demonstrated an intergenic locus at 4q31, near Hedgehog Interacting Protein (HHIP), which showed association with FEV_1/FVC and COPD (9). The next GWAS were small GWAS utilizing high throughput SNP genotyping with COPD (8, 43). These studies identified several genetic loci association with lung function and COPD, including HHIP, cholinergic receptor nicotinic acetylcholine 3/5 (CHRNA3/5), and family with sequence similarity 13 member A (FAM13A). However, the results of these smaller GWASs identified that greater statistical power was required to identify genes with confidence demonstrating the need for larger population sizes, which led to the application of meta-analyses to analyse the results of independent GWAS together to increase the statistical power for novel genetic variant discovery.

1.4.3 GWAS meta-analysis

GWAS meta-analyses has the potential to identify novel lung function and COPD loci. In 2010, the SpiroMeta and CHARGE consortium studies were the first two meta-analyses studies to examine FEV_1 and FEV_1/FVC and both identified novel loci associated with lung function

(10, 11). More recent studies identified 43 novel signals associated with one or more FEV_1, FVC or FEV_1/FVC, increased from 54 to 97 signals in UK Biobank and UK BiLEVE. The genes thought to underlie these signals were enriched in development, elastic fibres and epigenetic regulation pathways and include some potential drug targets (44). In 2019, Shrine and colleagues reported that they have defined 279 lung function signals, including 139 new signals, which reached genome-wide significance, providing further novel insights into the understanding and potential for utilising therapeutics (12). As Hall et al reported, many variants have shown overlap in association with lung function and COPD (37) (Figure **1**-**6**).

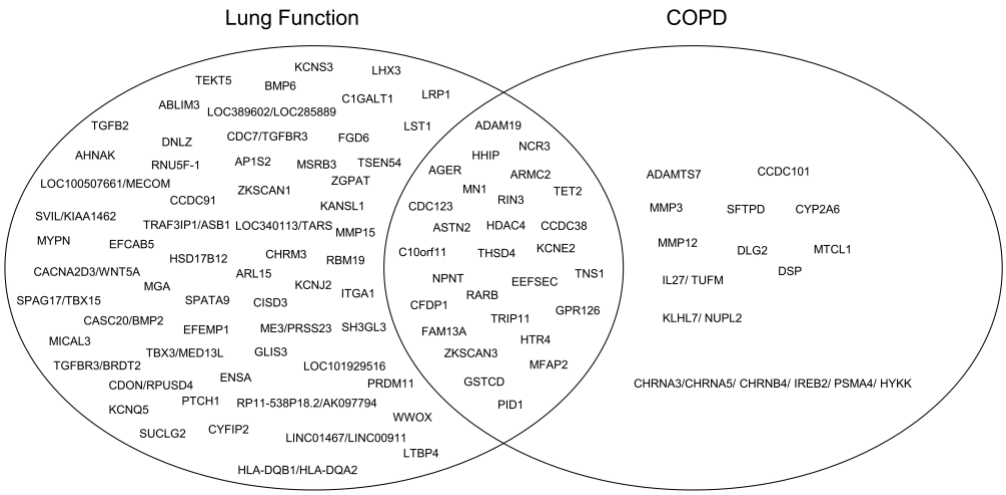


Figure 1-6. Susceptibility genes which have been reported to show association with lung function and/or COPD. The Venn diagram shows the genes which have association signals for lung function and COPD and the genes which have signals for both. The genetic risk score based on 95 lung function variants is highly predictive of COPD risk. Figure produced from Hall et al (37). Note that only those genes showing genome wide association signals are included in the overlap;

many of the lung function genes also show association with COPD at lower levels of significance.

1.4.4 Genetics association with COPD phenotypes

1.4.4.1 Emphysema

The genetic impacts on emphysema distribution patterns are another area which has attracted attention. The first GWAS of emphysema was performed using high-resolution chest computed tomography in 2,380 COPD patients and identified a SNP in BICD cargo adaptor 1 (BICD1) (12q11.2) associated with emphysema (45). In addition, Manichaikul et al reported the first GWAS genetic analyses for percent of emphysema by cardiac computed tomography in a large population-based cohort (46). Furthermore, Cho et al identified 5 genetic loci associated with emphysema phenotypes and concluded that previously identified GWAS loci not only associated with lung function and COPD, but also associated with emphysema or other airway phenotypes (47). Besides that, Boueiz and colleagues confirmed two loci previously associated with COPD susceptibility (HHIP and CHRNA5) and three novel loci on SOWAHB (4q13), TRAPPC9 (8q24), and KIAA1462 (10p12) (48). In the previous studies, the mRNA and protein levels of HHIP (as a COPD susceptibility gene in 4q31) decreased in COPD lung tissues and it was suggested that increased binding of Sp3 which is a transcriptional suppressor may be the underlying mechanism (49). On the other hand, CHRNA5 is also associated with COPD and smoking behaviour (50). Recently, one study reported a further putative causal variant for

emphysema, rs7962469, related to ACVR1B expression in lung, and showed it to be expressed in human bronchial epithelium (51) (Table **1**-**2**).

1.4.4.2 Chronic bronchitis

Chronic bronchitis (CB) is defined as being present in individuals if they have a cough productive of sputum for at least 3 months in each of two consecutive years (15). Symptoms in CB subjects without COPD have been reported to be approximately 9.1%, being more common in smokers and males (52). CB patients more frequently had dyspnoea, worse lung function, poor quality of life, and increased mortality (52, 53). In 2014, Lee et al reported the first GWAS of CB in a COPD population compared to smokers with normal lung function from three cohorts: COPDGene Study, GenKOLS (Bergen, Norway), and ECLIPSE (54). GWAS of CB COPD relative to smokers with normal spirometry, identified a previously reported COPD susceptibility region in FAM13A, and identified a novel SNP rs34391416 on chromosome 11p15.5 (EFCAB4A, CHID1 and AP2A2). They also found SNPs at loci on 2p25.1 (CYS1) associated with CB in two cohorts but not in ECLIPSE. A new study reported a novel SNP rs4150275 located in ERCC5 which showed association with CB (55).Thus, these studies provide evidence that genetic variants associate with the CB phenotype indicating there may be heterogeneity in genetic susceptibility to COPD (54).

1.4.4.3 Smoking behaviour

Smoking plays a vital role in affecting lung function and COPD development, but indirect evidence has been improved genetic factors in smoking behaviour. GWASs have identified up to eight associated loci associated with smoking behaviour, the strongest association locus at 15q25 (CHRNA3/6) (56). In a meta-analysis, they identified 8 SNPs near BDNF associated with smoking initiation, one SNP located near DBH was significantly related to smoking cessation (57). SNPs located on CHRNA5 (rs17486278) and near CYP2A6 (rs11667314) were associated with cigarettes per day. However, once the contribution of smoking itself is corrected for, there is no association between variants predicting smoking behaviour and lung function(12).

Gene	Phenotype	Study design and sample size	Year	Ref
BICD1	emphysema	using high-resolution chest computed tomography in 2,380 COPD patients (Bergen, ECLIPSE study, and NETT)	2011	45
SNRPF, PPT2	emphysema	determined percent emphysema and upper-lower lobe ratio in emphysema, non-Hispanic white (n = 2,587), African American (n = 2,510), Hispanic (n = 2,113), and Chinese (n = 704)	2014	46
HHIP, 15q25, AGER, SERPINA10, DLC1	emphysema	to identy genetic determinanats of quantitative imaging phenotypes. total sample size 12,031	2015	47

HHIP, CHRNA5, SOWAHB, TRAPPC9, and KIAA1462	emphysema	to identy the gentic influences of emphysema distribution in non-alpha-1 antitrypsin-deficient smokers. A total of 11,532 subjects.	2016	48
ACVR1B	emphysema	to characterize the functions of emphysema apico-basal distribution associiated variants.	2018	51
FAM13A, EFCAB4A, CHID1 and AP2A2, CYS1	chronic bronchitis	to investigate genetic variants associated with COPD subjects with CB relative to smokers with normal spirometry. COPD with CB n=1662; smokers with normal spirometry n=3520; COPD without CB n=3777.	2014	54
CHRNA3/6	smoking behaviour	GWAS meta analyses for the number of ciigarette smoked per day in smokers (n=31,266) and smoking initiation (n=46,481).	2010	56
CHRNA3, EGLN2, BDNF, DBH	smoking behaviour	meta-analyses of several smoking phenotypes within cohorts of the Tobacco and Genetics Consortium (n = 74,053).	2010	57

Table 1-2 Summary of published GWAS for emphysema, chronic bronchitis, and smoking behaviour

1.4.5 Summary

Overall, COPD is a major cause of morbidity and mortality worldwide. GWAS has identified hundreds of genetic variants associated with lung function and COPD. There are a number of genes associated with all 3 traits, ie lung function, COPD and emphysema (**Figure 1-7**, **Table 1-3**). 74 genes (including TET2, GSTCD, INTS12, and NPNT) were associated with lung function and COPD. According to GWAS meta-

analyses, the region 4q24 (including FLJ20184, TET2, INTS12, GSTCD and NPNT) seems to be a highly reproducible locus showing association with FEV_1, FEV_1/FVC and COPD. In the present study, I aimed to explore the expression and function of NPNT in human lung cells and tissues as well as the potential biological mechanisms whereby it contributes to COPD.

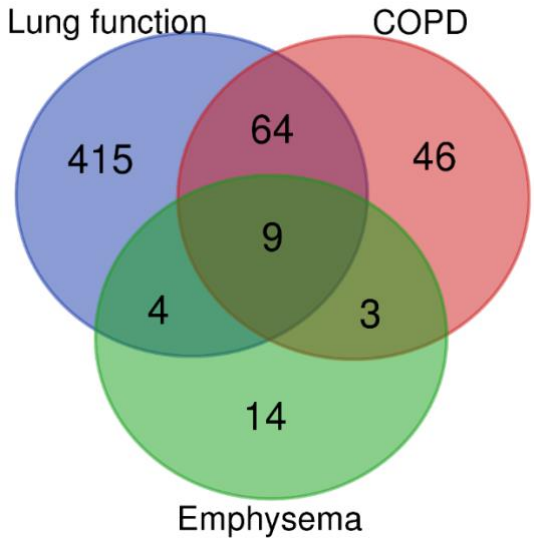


Figure 1-7. Susceptibility genes overlap among lung function, COPD and emphysema. Lung function consists of FEV_1, FVC, FEV_1/FVC, PEF, TLC, and DL_{CO}. The Venn diagram shows the gene which associated with lung function, COPD and emphysema reviewed and screened from GWAS (7-12, 43-48, 51, 54, 58-86).

Table 1-3. Literature review of susceptibility genes associated with the 3 main traits: lung function, COPD and emphysema.

Gene	Gene full name	Location	Biological function
TGFB2	transforming growth factor beta 2	1q41	Decreased in animal model after 12 weeks of tobacco smoking; Contributes to the production of ECM, epithelial to mesenchymal transition (87).
HHIP	hedgehog interacting protein	4q31.21	As COPD susceptibility gene, binding to transcript suppressor Sp3, DEGs enriched for ECM and cell proliferation pathways after knocking down HHIP (88).
PPT2	palmitoyl-protein thioesterase 2	6p21.32	Associated with loss of lung function(89)
AGER	advanced glycosylation end-product specific receptor	6p21.32	Potential biomarker to identify exacerbation COPD phenotype (90).
MMP12	matrix metallopeptidase 12	11q22.2	Variants at the MMP12 locus are protective against the development of COPD and emphysema induced by cigarette smoking (91).
CHRNA3	cholinergic receptor nicotinic alpha 3 subunit	15q25.1	Associated with smoking behaviours (50).
CHRNA5	cholinergic receptor nicotinic alpha 5 subunit	15q25.1	
IREB2	iron responsive element binding protein 2	15q25.1	
CYP2A6	cytochrome P450 family 2 subfamily A member 6	19q13.2	Associated with smoking behaviour (92).

1.5 Nephronectin (NPNT)

1.5.1 NPNT introduction

As previously described, eQTL evidence suggest that NPNT is a viable candidate gene associated with lung disease. NPNT was firstly named because it is found in the kidney extracellular matrix (ECM) and it is a novel ligand mediating α8β1 function in the developing mouse kidney (93). It is also called POEM (preosteobalst epidermal growth factor-like repeat protein with meprin, A5 protein, and receptor protein-tyrosine phosphatase μ domain) (94). Recent studies revealed that NPNT might play an important role in many pathological conditions, including acute kidney injury (95, 96), bone development and osteoporosis (97, 98), lung function (79), and silicosis (99). Interestingly, NPNT is implicated in the regeneration and proliferation of renal tubular epithelial cells, and it might play an important role in maintenance of epithelial integrity (95) (**Figure 1-8**).

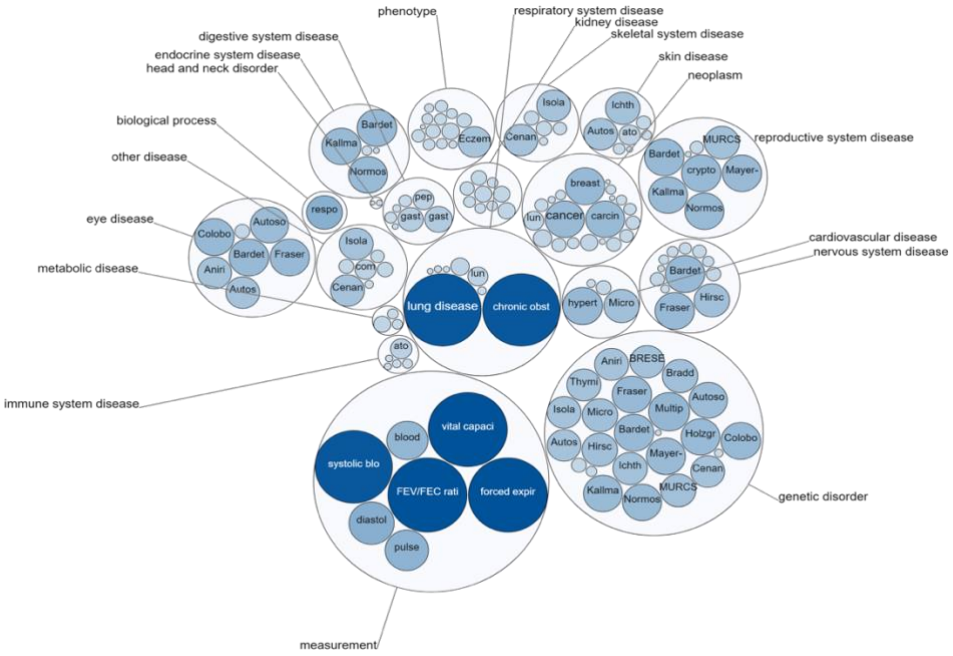


Figure 1-8. NPNT could contribute to 102 diseases. Data from the open targets (https://www.targetvalidation.org/). NPNT plays a critical role in the respiratory diseases (e.g. chronic obstructive pulmonary disease), and lung function (FEV/FEV ratio, FEV_1 and vital capacity). The circle size associated with effect size of diseases.

1.5.2 NPNT Gene and protein structure

NPNT mRNA has been detected in lung, kidney, ear and eye during foetal development; and is widely found in lung, kidney, brain, thyroid gland, uterus and blood vessels in adults (100).The human NPNT gene is located on chromosome 4; it contains 13 exons and encodes a 61 kDa proteins. There are currently 12 transcripts described in PubMed Gene (Figure **1**-**9**). Human NPNT protein contains five Epidermal Growth Factor (EGF)-like repeats (amino acids 57-250), an Arg-Gly-

Asp (RGD) motif (amino acids 382-384), N-glycosylation site repeats (amino acids 273), and a COOH terminal meprin-A5 protein-receptor protein tyrosine phosphatase μ (MAM) domain (amino acids 417-561) (93, 94) (Figure **1**-**10**). The EGF-like repeat and MAM domain are common features in ECM proteins, and the EGF-like proteins often act as a transmembrane protein and control cell differentiation (101). The RGD motif of NPNT promotes cell attachment and can be recognised by several integrin heterodimers, including α8β1 (94) and independently mediated cell adhesion (102).

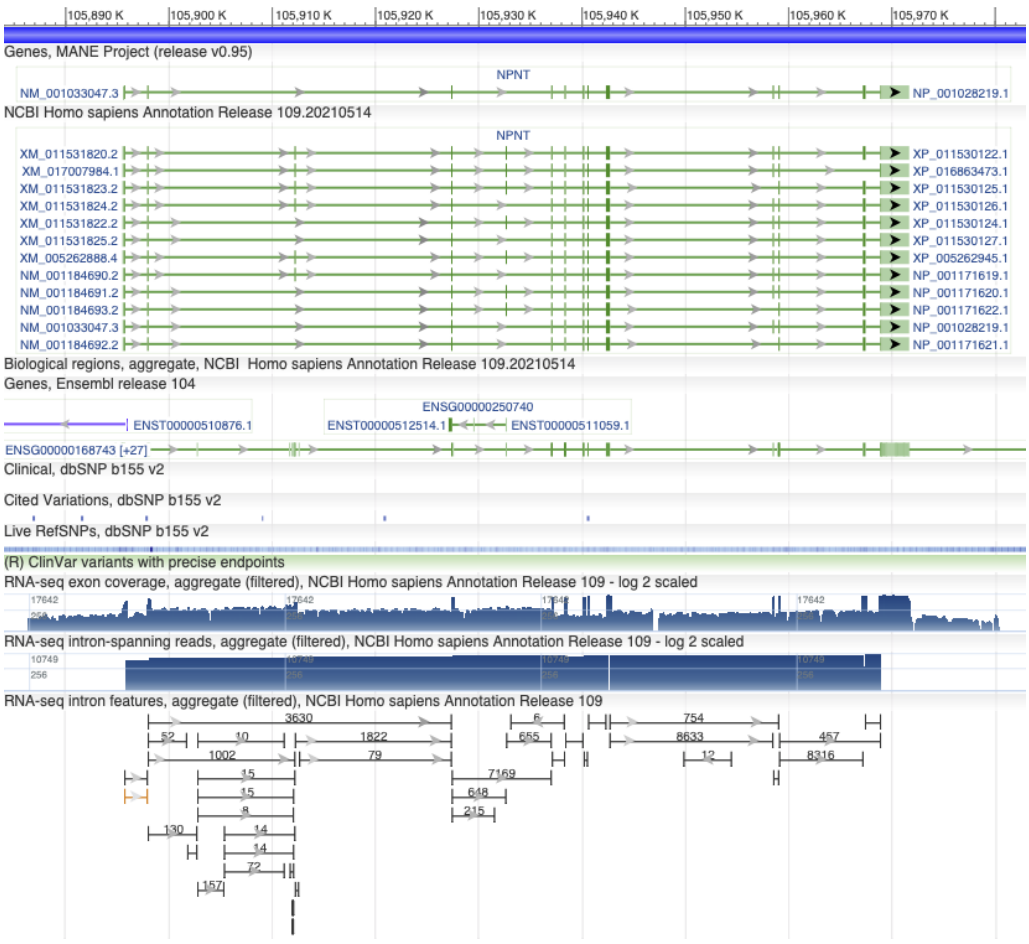


Figure 1-9. Genomic regions, transcripts, and products from Pubmed Gene (https://www.ncbi.nlm.nih.gov/gene/255743) (14th September 2021). 12 transcripts were described for NPNT in NCBI Homo sapiens

Annotation Release 109.20210514. Accession prefix: NM and XM, which means protein-coding transcript (usually curated) and predicted model protein-coding transcript, respectively. The next block shows Genes from Ensembl. The last block is RNA-seq exon coverage, intron-spanning reads, and intron features.

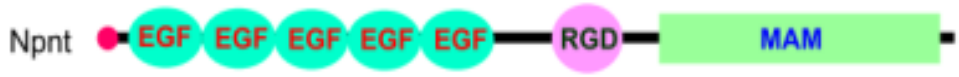


Figure 1-10. Structure of human nephronectin. NPNT protein contains EGF-like repeats, an RGD, N-glycosylation site repeats, and a MAM domain.

1.5.3 NPNT function in non-respiratory diseases

In NPNT gene-edited mice, NPNT is involved in the process of kidney development via integrin α8β1 and promotes glial cell line-derived neurotrophic factor (Gdnf) expression (103). NPNT has also been suggested as a potential biomarker during acute kidney injury (96). In addition, NPNT is present in the mesangial matrix expansion of diabetic nephropathy, and in one study it was suggested that it distinguishes diabetic nephropathy from IgA glomerulonephritis, membrane-proliferative glomerulonephritis and lupus nephritis in renal biopsy specimens (104). Furthermore, in one recent study reported that the expression of NPNT protein in serum from patients with nephrotic syndrome can be used as a biomarker of steroid resistance (105).

NPNT is implicated in injury and bone homeostasis (106). Kahai and colleagues revealed that NPNT regulates osteoblast differentiation, and the EGF-like repeats were essential for this function (107). Interestingly, NPNT in osteoblasts was itself regulated by many factors, such as TGF-β1, oncostatin M (OSM), tumour necrosis factor (TNF)-α, interleukin (IL)-1β and fibroblast growth factors (FGF)-2 (108-113) (Table **1**-**4**). NPNT plays important roles in other biological processes. Activation of Wnt/β-catenin signal increases NPNT expression in the hair follicle bulge germ, but suppresses NPNT expression in the interfollicular epidermis (114). Patra and colleagues demonstrated that NPNT is involved in cardiac development of zebrafish by regulating endothelial cell differentiation as an upstream regulator of the BMP4-Has2 signal pathway (115). In addition, Wu et al suggested that postnatal cardiac ECM slit guidance ligand 2 (SLIT2) and NPNT promote postnatal cardiomyocyte cytokinesis (116, 117). Furthermore, NPNT may have a role as a matrix effector in cancer. Recent studies reported that NPNT was downregulated in breast cancer and involved in promoting metastasis (118, 119). NPNT promoted gastric cancer cell proliferation, migration and invasion and could be a potential prognostic biomarker in gastric cancer (120).

Function	Cytokine	Pathway	Cells type	Comments	Author-year
suppress NPNT (down-regulated)	IL-1β	via ERK1/2 and JNK	mouse osteoblastic cell line MC3T3-T4	IL-1β strongly inhibited NPNT expression in both time- and dose-dependent manners via activation ERK1/2 and JNK signal pathways.	Iezumi, 2017
	TGF-β1	N/A	mouse osteoblastic cell line MC3T3-T1	NPNT was down-regulated (-5.2 fold) in the MC3T3-E1 osteoblasts that had been treated with TGF-β1 for 6 days by cDNA microarray.	Kahai, 2004
	FGF-2	via activation JNK and PI3K pathways	mouse osteoblastic cell line MC3T3-T1	FGF-2 can down-regulate NPNT expression in a dose- and time-dependent manners via JNK/PI3K pathway.	Kato, 2018
	oncostatin M (OSM)	via activation JAK/STAT and MAPK (including MEK, p38, JNK)	mouse osteoblastic cell line MC3T3-T2	OSM, is a cytokine and member of the IL-6 subfamily, strongly inhibited Npnt mRNA expression in MC3T3-E1 cells from a mouse osteoblastic cell line, and inhibited the cell differentiation.	Kurosawa, 2015
	TGF-β1	via activation TGF-β RI, ERK1/2 and JNK pathways	mouse osteoblastic cell line MC3T3-T2	TGF-β1 inhibited NPNT expression in a time- and dose-dependent manner via activation TGF-β receptor I and extracullular signal-regulated knase/c-Jun N-terminal kinase pathways.	Miyazono, 2007
	TNF-α	stimulate NF-kB signal pathway	mouse osteoblastic cell line MC3T3-T3	TNF-α, a pleiotropic pro-inflammatory cytokine, can down-regulate the expression of POEM in both time- and dose-dependent manners via the nuclear factor kappa B pathway.	Tsukasaki, 2011

promote NPNT (up-regulated)	inhibitor of TGF-β type I receptor(Alk5) inhibitor	Alk5-SMAD signaling pathway in osteobalsts	mouse osteoblastic cell line MC3T3-T5	Alk5 inhibitor, strongly induced Npnt expression in osteoblast-like MC3T3-E1 cells.	Tsukasaki, 2012
	BMP signaling anatagonist Noggin	BMP pathways	hair follicle bulge	Noggin upregulated NPNT expression	Fujiwara, 2011
	BMP2	via the BMP-SMAD signaling pathway	murine skeletal muscle cell line (C2C12)	BMP-2, an osteogenesis inducing cytokine, strongly 29pregulated the expression of NPNT via the BMP-SMAD pathway in murine skeletal muscle cell line (C2C12).	Kurosawa, 2016
	1α,25-dihydroxyvitamin D 3	via activation the VDR	mouse osteoblastic cell line MC3T3-T5	vitamin D3 in both a time- and dose-dependent way via vitamin Dreceptor (VDR) activation in MC3T3 cells	Hiranuma, 2016
	Wnt3a	Wnt/β-catnin signal pathway	mouse osteoblastic cell line MC3T3-T2	Wnt3a strongly enhanced NPNT mRNA expression in both time- and dose-dependent manners via the Wnt/β-catenin pathway.	Ikehata, 2017
	Wnt/β-catenin	Wnt/β-catenin signal pathway	interfollicular epidermis	Aactivation of Wnt/β-catenin signal increased NPNT expression in the hair follicle bulge germ, but suppressed NPNT expression in the interfollicular epidermis	Fujiwara, 2011

miRNAs	miR-378	enhanced differentations of osteoblasts	mouse osteoblastic cell line MC3T3-T2	there is competition between NPNT and GaINT7 for miR-378 binding, resulting in increased GaINT7 activity, which in turn lead to increased NPNT glycosylation and product secretion, thereby resulting in a higher rate of osteoblast differentiation.	Kahai, 2009
	miR-23a, miR101a, miR-296-5p, miR-328, miR-340-3p, and miR-425	N/A	mouse osteoblastic cell line MC3T3-T3	N/A	Lee, 2011

Table 1-4 Summary of current knowledge on potential functional roles for NPNT and regulation of NPNT expression.

1.5.4 The role and regulation of NPNT in lung

As discussed above, NPNT variants are not only associated with lung function, but NPNT is also potentially implicated in foetal lung development. In addition, NPNT is highly expressed in the foetal and adult lung and expressed in epithelia cells and deposited in the ECM which between epithelial cells and mesenchymic cells (93, 100). Miller and colleagues reported NPNT expression is elevated with increasing foetal lung age by microarray (121). Thus, NPNT appears to be involved both in the development process and control of function of the lung. Interestingly, Lee and colleagues reported the serum NPNT was higher in silicosis patients than in healthy volunteers. They suggest NPNT could be utilised as a suitable biomarker to follow up the progress of silicosis patients (99). Importantly, the sentinel SNP rs34712979, located in NPNT, contributes to COPD risk (122).

1.5.5 Summary

According to GWASs, the locus containing NPNT is strongly associated with lung function and COPD. NPNT mRNA is highly expressed in lung, kidney, and thyroid gland of foetal and adult (100). NPNT mRNA expression is increased during foetal lung development (121). Also, NPNT protein is detected in pneumocytes in lung tissue from healthy donors (https://www.proteinatlas.org/). NPNT protein is a novel ECM which lies between epithelium and mesenchyme, and mediates α8β1 function in the developing mouse kidney. NPNT implicates in epithelial-

mesenchymal interactions, cell proliferation, migration, and differentiation (123, 124). Therefore, all the current evidence suggests that NPNT may contribute to the process of COPD.

1.6 Aims

From the above, it is clear that genetic variation at the chromosome 4 (NPNT) locus is associated with both lung function variability and the risk of developing COPD. The signals at this locus strongly variant implicates NPNT itself as one the likely causal gene underlying this association. Therefore, further work to define the functional role of NPNT in lung homeostasis is important. Although NPNT has been reported to have a range of functions in the kidney, bone, tooth, muscle and tumour, a more in depth analyses of NPNT expression and function in lung is clearly required. Therefore, the aims of the work described in this report were:

a) To investigate the profile of NPNT mRNA and protein in different airway structural cell lines and lung tissue from individuals with COPD and healthy controls.

b) To identify the contribution of targeting NPNT (knockdown by siRNA or knockout by CRISPR/Cas9) to downstream gene expression and potential pathways involved in human bronchial epithelial cell homeostasis.

c) To compare the Microarray and RNA-seq platforms in transcriptome profiling of human bronchial epithelial cells following NPNT knock-down.

d) To detect the expression of NPNT protein in serum of stable or acute exacerbation of COPD and healthy controls.

2. General Materials and Methods

2.1 Mammalian cell culture Methods

2.1.1 Primary human bronchial epithelial cells (HBECs)

Primary HBECs were obtained from Lonza (Berkshire, UK). The majority of the work described in this thesis used cells from 3 donors (195307, 494599, 328324) which were cultured in PneumaCult-Ex Basal Medium (490mL PneumaCult-Ex Basal Medium plus 10mL PneumaCult-Ex 50X Supplement and 500μl 200X hydrocortisone stock solution) (StemCell, UK) (125, 126).

Donor		1	2	3
Lot number		195307	494599	328324
Donor Information	Age	19	19	18
	Sex	M	F	F
	Ethnicity	E	E	E
	Smoker	N	N	N
Cell properties	Viability (%)	91	83	85
	Seeding efficiency (%)	66	61	99
	Doubling time (h)	16	28	31
Virus tests	HIV	N	N	N
	HBV	N	N	N
	HCV	N	N	N
Safety tests	Sterility test	N	N	N
	Mycoplasma test	N	N	N

Table 2-1. Donor information and HBECs properties of cells obtained from Lonza. E=European populations, N=negative. (Adapted from the thesis of Robert Hall, 2020)

2.1.2 Culturing conditions for Mammalian primary cells

The full PneumaCult-Ex Basal Medium (10mL) was added to PneumaCult™-Ex 50X Supplement, 0.5mL of Hydrocortisone Stock Solution and 490mL of PneumaCult™-Ex Basal Medium. Cells were incubated at 37℃ and 5% CO_2 in a T75 cm^2 tissue culture flask, and full medium changes performed every 2 days until cells were approximately 80% confluent and ready to be passaged. To obtain cells from the culture flask through trypsinisation, cells were harvested using 5 mL warmed trypsin (5mg/ml)/EDTA (2mg/ml, Sigma, USA), this process was stopped by adding the same volume of trypsin inhibitor, and then cells were pelleted in a 20ml universal tube. HBECs were centrifuged at 200g and resuspended in 5 mL of medium, and seeded to the required culture vessel. HBECs were utilised at passage 3 when approximately 70~90% confluency.

2.1.3 Cell counting by Haemocytometer

The cell number was counted using the Neubauer haemocytometer. The haemocytometer and coverslip were washed to remove any cell debris and dust. The coverslip covered the central ruled area and gentle pressure was applied until Newton's rings were seen. A 10μL cell suspension was added under the coverslip of the haemocytometer and the cell count determined using an inverted light microscope. The cell number of four corner squares of the haemocytometer was recorded and the average was calculated.

A. Cells/ml = (average count / square) × 2, (dilution factor) × 10^4, (chamber conversion factor); when 10 squares are counted.

For example: average count per square is 45 cells × 2 ×10^4 = 9×10^5 cells per ml.

b. Total cells = cells per mL × the original volume of fluid from which the cell sample was taken.

For example: 9 × 10^5 (cells/mL) × 10mL (original volume) = 9×10^6 total cells.

2.1.4 Freezing Mammalian cells

Some cells were frozen down to ensure the cells were available for further experiments. For this, cells were added to 1mL normal culture media with 10% DMSO in a sterile cryogenic vial, the media was mixed and then transfer to the freezer at -80℃ for one night, then the cryogenic vial was stored in liquid nitrogen (-180℃).

2.2 Molecular Biology Methods

2.2.1 RNA extraction from HBECs

Total RNA was extracted from HBECs, and used for Quantitative Real-time Polymerase Chain Reaction (qRT-PCR) (127) and Microarray or RNA-sequencing experiments (126, 128). The GenElute™ Mammanlian Total RNA Miniprep Kit was purchased from Sigma Aldrich (RTN70-1KT, USA) and was used to isolate mammalian total RNA from cells.

Protocol

Cells were grown in a 6-well plate. The culture media was removed and cells washed with 2mL of cold PBS. 300µl lysis buffer (10μL 2-mercaptoethanol to 1mL lysis solution, both reagents are provided in the kit) was added to each well, cells scraped and then cells from 2 wells together. The 600 µl cell lysate was centrifuged at maximum speed (13,000rpm) for 2 minutes to homogenize the cells. An equal volume of 70% ethanol solution was added, and the lysate/ethanol mixture was transferred to the GenElute Binding Column. The column was centrifuged at maximum speed for 30s. The flow through was discarded. 10µl Dnase I stock solution was added to 70µl Dnase buffer (SLBZ0400, Sigma, USA) for each reaction. 80µl of this mixture was added to each column and incubated at Room Temperature for 15 minutes. 250µl of Wash Solution 1 was pipetted into the column and centrifuged at 13,000rpm for 15 seconds. 500µl of Wash Solution 2 was added into the Column and centrifuged at maximum speed for 15

seconds. The flow through was discarded. A second 500µl volume of Wash Solution 2 was added into the column and centrifuged at maximum speed for 2 minutes. Finally, 50µl of Elution Solution was added into the binding column and centrifuged at maximum speed for 1 minute. The resulting RNA samples were stored in a -80℃ freezer.

2.2.2 Quality control of generated total RNA

2.2.2.1 Nanodrop Quantification of Nucleic Acid

The first quality control method used the Nanodrop ND-2000 spectrophotometer (Thermo Fisher Scientific, USA). 1.5µL of every sample was measured and also assessed for RNA purity. The ratio of absorbance at 260nm and 280nm using the full ultraviolet (UV) and visible spectrum (220~750nm) was used to assess the purity of DNA and RNA. A ratio of -1.8 is generally accepted as “pure” for DNA; a ratio of -2.0 is generally accepted as ‘pure’ for RNA. 260/230 ratio is a secondary measure of nucleic acid purity. 260/230 values for a “pure” nucleic acid are often in the range of 1.8-2.2. If the ratio is significantly lower, this may show the presence of reagent contamination, such as ethanol carryover.

2.2.2.2 RNA Integrity Analysis

Nanodrop is an alternative method to provide a rough quantification of RNA samples. Samples can be run on a RNA bioanalyzer chip to determine the quality. RNA Integrity Number (RIN) values should be 8 or above for Microarray or RNA sequencing (129).

RIN was measured by 2100 Bioanalyzer and RNA LabChip kits (Agilent Technologies, USA). RNA quality was visualised as gel electrophoresis, which produces two bands for the 28S and 18S rRNA species.

Agilent 4200 TapeStation was used to measure RIN for RNA sequencing. The total RNA was prepared, and the RNA concentration was determined using UV-Vis spectrophotometry. The RNA samples were diluted to the described nominal concentration in water based on the quantification. RNA analysis was performing using total RNA according to instruction using a 4200 TapeStation. TapeStation software revisions A.02.01 were used for data analysis. The RIN value was automatically detected and displayed under the individual lane of the gel image. RIN was calculated at a scale from 1 to 10.

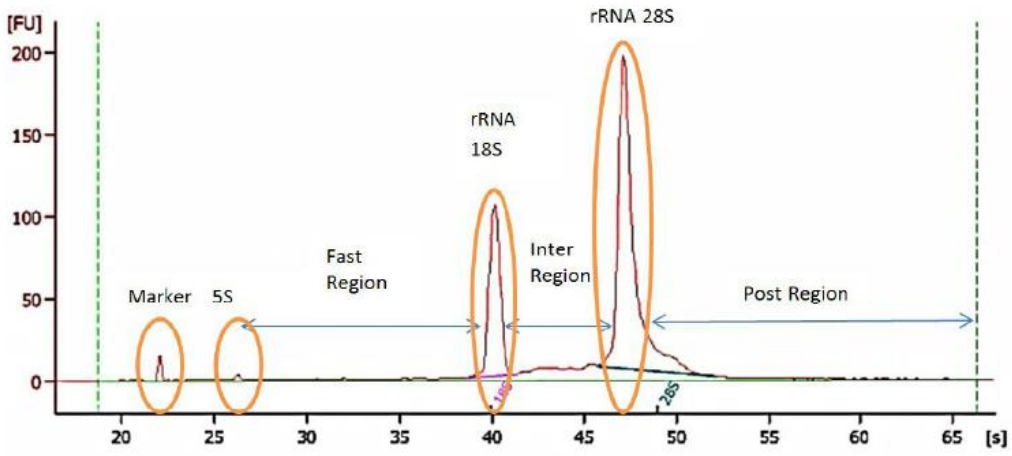


Figure 2-1 Example of a RNA profile showing different regions (pre-, 5S-, fast-, inter-, precursor-, post-region) and peaks (marker, 18S, 28S). These Data were used to analyse profiles, providing for example, RNA area, RNA concentration, and rRNA ratios.

2.2.3 Synthesis of complementary DNA

Complementary DNA (cDNA) is DNA synthesised from a single-stranded RNA template in a reaction catalysed by the enzyme reverse transcriptase (130). The resultant cDNA / RNA complex is degraded by Rnase H, leaving single stranded cDNA for downstream assays. Reactions without Reverse Transcription (RT) enzyme were carried out for each sample was used as controls for genomic DNA contamination.

Protocol

1. Dnase I (Cat. No. 18068015, Invitrogen, Life Technologies) was purchased for RT-PCR. RNA samples were transferred into a new tube with DEPC water to a total volume of 8µl (the total RNA is 1000ng). Then 10× Rnase I Buffer 1 µl and Dnase I Amplifition Grade 1 µl (supplied from the kit) was added at Room Temperature for 15 min. Next step was to add 1µl EDTA and samples were incubated at 65℃ for 15 min then placed on ice for 1 minute.
2. RNA samples were divided into 2 tubes of each 5.5μl. 5.5μl of each sample was then used to make the cDNA for either the RT+ve or RT-ve reaction.
3. The kit of Superscript First strand cDNA synthesis (Cat. No. 11904-018, Invitrogen, Life Technologies, USA) was purchased. The follow reaction was prepared (for every sample):

Substance	Volume (µl)
Dnase treated RNA	5.5
dNTP	1
Random hexamer	1
DEPC water	2.5
Total volume	10

Table 2-2 cDNA synthesis reaction 1.

4. The reaction mixtures were incubated at 65℃ for 5 minutes and placed on ice for one minute.
5. The follow reaction mix was made:

Component	***Volume (µl)***
10X RT reaction buffer	2
25mM $MgCl_2$	4
0.1M DTT	2
40 units/μL RNaseOUT™ Recombinant Ribonuclease Inhibitor	1
total volume	9

Table 2-3 cDNA synthesis reaction 2.

6. This 9ul of mastermix was added to each tube and incubated at 25℃ for 2 minutes. *Superscript II reverse transcriptase enzyme* (1ul) was added to the RT positive (+) samples **ONLY**. Samples were incubated on a Biometra PCR block at: 42℃ for 50 minutes, 70℃ for

15 minutes, and then placed on ice. 1ul *RnaseH/tube (2U)* was added to the RT+ve samples **ONLY**, and samples were incubated at 37℃ for 20 minutes, and then placed on ice. Finally, samples were diluted with 80µl DEPC water and put in a -20℃ freezer (total volume is 100 µl). These samples were prepared for TaqMan PCR.

2.2.4 Quantitative Real-time Polymerase Chain Reaction

2.2.4.1 Principle

qRT-PCR is a widely used method to amplify a particular segment of a DNA strand. This basic PCR process has three main steps; these are separating the target DNA (denaturation), two primers binding to the DNA sequence (annealing), and making a copy (extension) (131).

2.2.4.2 Primer design

We purchased TaqMan Gene Expression Assays for the gene NPNT (Cat No. 4351368, Hs01568131_m,1 covers 4-6, amplicon length 70bp, Thermo Fisher Scientific, USA) kit as primers for qRT-PCR.

Assay Detail:

Interrogated Sequence		**Translated Protein**	**Exon Boundary**	**Assay Location**	**Amplicon Length**
RefSeq	NM_001033047.2	NP_001028219.1	4 – 5	598	70
	NM_001184690.1	NP_001171619.1	5 – 6	649	70
	NM_001184691.1	NP_001171620.1	5 – 6	688	70
	NM_001184692.1	NP_001171621.1	4 – 5	598	70
	NM_001184693.1	NP_001171622.1	5 – 6	688	70
	XM_005262888.4	XP_005262945.1	6 – 7	726	70
	XM_005262890.1	XP_005262947.1	4 – 5	373	70
	XM_011531820.2	XP_011530122.1	6 – 7	729	70
	XM_011531822.2	XP_011530124.1	5 – 6	678	70
	XM_011531823.2	XP_011530125.1	6 – 7	729	70
	XM_011531824.2	XP_011530126.1	5 – 6	639	70
	XM_011531825.2	XP_011530127.1	4 – 5	588	70
	XM_017007984.1	XP_016863473.1	6 – 7	729	70
GenBank mRNA	AK290029.1	-	4 – 5	545	70
	AK295772.1	-	5 – 6	649	70
	AK302477.1	-	5 – 6	687	70
	AK304279.1	-	4 – 5	597	70
	AK304721.1	-	5 – 6	635	70
	AL832465.1	-	4 – 5	512	70
	AY358336.1	-	4 – 5	587	70
	BC131587.1	-	4 – 5	445	70

2.2.4.3 qRT-PCR conditions

1) 21µl of samples (negative or positive cDNA) was added in each new tube (0.5ml Eppendorf).
2) The mix solution as follows was prepared for each sample, and added in each tube (total 70µl), and **centrifuged** to pool the liquid to the button of the tube.

Reagent	Per well of 96 well plate	Put samples in 3 wells of 96 well plate	
		3.5×	
NPNT	1 µl	3.5 µl	49 µl
18s	1 µl	3.5 µl	
H_2O	2 µl	7 µl	
2× Taqman(MM)	10 µl	35 µl	

Table 2-4 qRT-PCR reaction mixture. There was 14 µl total volume in every well of 96-well plate. I made 3.5 times of total volume in each reaction when I prepared the mix solution. Thus, in each reaction the triplicate wells would prepare 49µl.

3) Took 20µl master mix to each well (triplicate of wells) of 96-well plate, and centrifuged 350g for 5min in 4℃ condition.
4) The 96-well plate was put in a Techne PrimeG thermal Cycler and run using the following parameters.

Segment	**Time**	**Temperature**	**Cycle**
Segment 1	2 minutes	50℃	1 cycle
Segment 2	10 minutes	95℃	1 cycle
Segment 3	15 seconds	95℃	40 cycles
	1 minute	60℃	

Table 2-5 qRT-PCR cycling parameters.

2.3 Protein detection methods

2.3.1 Western-Blot Methods

2.3.1.1 Principle of Western Blot

Western blotting is a method of visualization and quantification of proteins using specific antibodies based on the methodology outlined by Towbin (132, 133) (Figure **2-2**). For the work described in this thesis, proteins were separated, transferred to a polyvinylidene difluoride (PVDF) membrane, and assessed for the expression of NPNT.

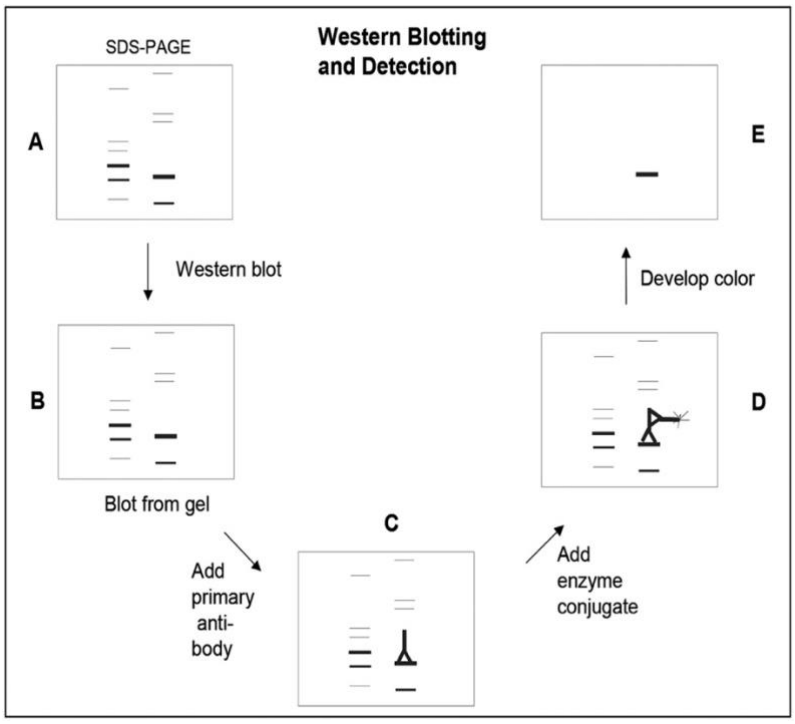


Figure 2-2 Schematic representation of Western blotting and detection procedure. (A) Unstained SDS PAGE gel prior to Western blot. The bands shown are hypothetical (B) Exact replica of SDS PAGE gel obtained as a blot following Western transfer. (C) Primary antibody binding to a specific band on the blot (D) Secondary antibody conjugated to an enzyme (alkaline phosphatase or horse radish

peroxidase) binding to primary antibody I colour development of specific band (Kurien et al, 2015)

2.3.1.2 Cell lysate

Total intracellular protein samples were collected from monolayer cells cultured in 6-well plates. The cells were washed twice with 2mL of cold PBS and 100μL of cytobuster mix solution was added (with 10μL Benzone Nuclease, 1μL DTT and 2μL Pep A in the 1mL cytobuster mix). The cell monolayer was scraped and the solution collected for western blotting.

2.3.1.3 SDS-polyacrylamide gel electrophoresis

Sodium Dodecyl Sulphate polyacrylamide gel electrophoresis (SDS-PAGE) is a method used to monitor protein purifications, and to separate proteins according to their molecular weights or its charge. The molecular weight of Nephronectin is 70-90 kD. A 10% resolving gel (constituted as showed in Table 2-6 Reaction mix to create resolving gels.) was therefore chosen and the Bio-Rad Mini-protean III gel electrophoresis apparatus was used.

Solution components	Volume for 7.5% gel	Volume for 10% gel	Volume for 12.5% gel
Distilled water	4.8 ml	4 ml	3.1 ml
30% (w/v) acrylamide	2.5 ml	3.3 ml	4.2 ml
10% SDS	100 ul	100 ul	100 ul

1.5M Tris-HCl (pH8.8)	2.5 ml	2.5 ml	2.5 ml
10% APS	100 ul	100 ul	100 ul
TEMED	4 ul	4 ul	4 ul

Table 2-6 Reaction mix to create resolving gels.

Once the resolving gel had set, the stacking gel was poured on top of the resolving gel until the glass plates were filled (constituted as showed in Table **2**-**7**). A 10-well comb/1.5 mm space plate was quickly inserted into the stacking gel, taking care not to introduce any air bubbles. The gel was left to set and the comb removed. The position of the wells was marked and a48ormalized48edsed acrylamide washed out with 1x running buffer. The shape of each well, if necessary, was restored by straightening the teeth of the stacking gel with a micro-spatula.

Solution components	Volume
Distilled water	3.4 ml
30% (w/v) acrylamide	830 μl
10% SDS	50 μl
1.0M Tris-HCl (pH6.8)	630 μl
10% APS	50 μl
TEMED	5 μl

Table 2-7 Reaction mix details to prepare 5% stacking gel.

The concentration of protein was measured by Bradford protein assay (Bio-Rad, USA). Loading buffer and protein samples were loaded into wells (i.e. 7.5μl of each for 0.75mm gel, 15μl of each for 1.5mm gel). The gel tank was filled with sufficient 1x Tris-glycine/SDS running buffer, 1x Tris-glycine/SDS running buffer to cover samples. Samples were run through the stacking gel at 100V and through the resolving gel at 200V for about 1 hours. Just before the *"blue" front* ran off the bottom of the gel, the power supply was switched off and the lid removed. The stacking gel was immediately cut from the resolving gel and the resolving gel soaked in 1x *transfer* buffer.

2.3.1.4 Western Blotting

Proteins were transferred to a PVDF membrane after separation by SDS-PAGE. Non-specific protein binding sites were blocked by blocking solution after protein was transferred. The primary antibody binds against a specific protein and a labelled polyclonal secondary antibody then binds to the primary antibody, and also amplifies the signal. Finally, the protein was visualised using photographic film.

Briefly, the resolving gel was equilibrated in 1X transfer buffer [25mM Tris base; 192mM glycine; 20% v/v methanol (optional for PVDF), pH8.3] for 15 min, and the PVDF membrane was activated by dipping in 100% methanol, and the filter paper to the dimensions of the resolving gel (approximately 7cm x 9cm) also equilibrated in 1X transfer buffer. The gel sandwich was prepared as follows:

a) Place the cassette, with the black side down, on a plastic tray.
b) Place one pre-wetted fibre pad on the black side of the cassette.
c) Place a sheet of pre-wetted filter paper on the fibre pad.
d) Place the equilibrated gel on the filter paper.
e) Place the pre-wetted (and activated, in the case of PVDF) membrane on the gel.
f) Complete the sandwich by placing a piece of filter paper on the membrane.
g) Add the last fibre pad.

The Cassette was firmly closed and placed in a tank. On completion the power supply was turned off, the blotting sandwich disassembled and the membrane removed for blocking. The membrane was blocked with 5ml of blocking buffer (goat serum, non-fat dry milk, or BSA in PBST/TBST) for 2 hours at room temperature. The primary antibody for NPNT was diluted to 1:250 (20μl antibody in 5000μl milk blocking buffer), β-acin diluted to 1:1000 and then the membrane was incubated with the primary antibody (Table **2**-**8**) solution overnight at 4℃ fridge. The membrane was then washed with TBST 3 times for 5-10 minutes each wash, followed by incubation with a horseradish peroxidase (HRP)-conjugated secondary antibody (Table **2**-**8**) diluted in the blocking solution for 45-60 min at room temperature.

Blocking Agent	Primary Antibody	Secondary Antibody
10% non-fat dry Milk	NPNT polyclonal (1:250).	polyclonal goat anti-mouse HRP (1:1000)

Table 2-8 The antibodies applied in Western blotting for NPNT.

2.3.1.5 ECL method of Western Blotting protein detection

The Enhanced Chemiluminescence (ECL) method of protein detection was used to determine labelled protein levels. The membrane was washed 3X with PBST, and then was incubated for one minute at room temperature in mix solution (1ml of ECL reagent 1 and 1ml of ECL reagent 2 for each 7cm x 10cm blot). In the dark room photographic film was loaded (Hyperfilm, GE Healthcare) into the cassette. The exposed film was visualised using Kodak Developer, washed by tap water, and then turned transparent by Kodak Fixer.

The density of bands in exposed films was calculated by software ImageJ. The result was normalised to the concentration of β-actin staining using the same cell lysate.

2.3.2 Tissue Immunohistochemistry

Immunohistochemical techniques detect antigens in tissue sections through immunological and chemical reactions (134). The antibody-antigen interaction can then be visualised by microscopy.

2.3.2.1 Foetal and adult lung tissues

Slides of foetal and adults' lung tissues were provided by the Joint MRC/Wellcome Trust (grant # 099175/Z/12/Z) Human Developmental Biology Resource (www.hdbr.org). Anti-NPNT antibody produced in rabbit was purchased from Sigma (HPA003711, USA) and Peroxidase/DAB+ Rabbit/Mouse from Agilent (Lot: 20012355, REF: K5007, USA), which contains HRP Rabbit/Mouse, Substrate buffer, DAB+CHROMOGEN.

2.3.2.2 Immunohistochemistry (IHC)

The following procedure was used:

Step 1-Removal of paraffin wax: Slides were put in Histoclear for 10 minutes, removed and drained briefly, then placed in absolute ethanol for 2× 5 minutes. Then slides were transferred to 95% ethanol for 2× 5 minutes. Slides were removed and washed in running tap water.

Step 2-Antigen retrieval-Steamer: The steamer was heated for 20 minutes. The slides were put in the steamer and covered by Sodium Clitrate buffer in a tip box. Slides were treated for 10 minutes then cooled for 20 minutes.

Step 3-Immunohistochemistry: Slides were incubated with 2mL PBS and agitated for 2 minutes. 100µl peroxidase blocking solution was

added to cover the specimen for 5 minutes, and then washed with PBS. 100µl primary antibody was then added (anti-NPNT produced in rabbit, Sigma Aldrich, HPA003711), which was diluted 1:25, 1:50, 1:100 and 1:200, and incubated for 1 hour at room temperature. The slides were then washed with PBS again. Secondary antibody (Bottle A) was applied to cover specimen (3 drops) for 30 minutes, then washed with 10mL PBS for 5 minutes. 100µl of the substrate-chromogen solution (DAB) was applied to cover the specimen and incubated for 5 minutes. Step 4-To set and view under microscope. Slides were counterstained in Mayers haematoxylin for 10 seconds, and then were washed under tap water. Slides were washed in 10mL 95% ethanol, 100% enthanol and histoclear each for 2 times for 5minutes. Slides were mounted using Vectamount and left to set. Slides were examined by microscopy.

2.3.3 Enzyme Linked Immunosorbent Assays (ELISA)

Enzyme linked immunosorbent assays (ELISA) can be used to detect specific antibodies, soluble antigens, or cell-surface antigens (135). In this thesis, ELISAs were used to quantify protein levels of NPNT in serum of COPD patients, and healthy controls, using similar methods to those previously reported by Watany and colleagues (105).

2.3.3.1 Reagent preparation

NPNT was measured using enzyme-linked immunosorbent assay (ELISA) (SEH522Hu, Cloud-Clone Corp, UK) following the instructions provided by the manufacturer and as used in a previous study (105). All kit components and samples were brought to room temperature (18-25℃) before use. The Standard sample was reconstituted with 1.0mL of Standard Diluent, shaken gently, and kept for 10 minutes at room temperature. The concentration of the standard sample in the stock solution is 20ng/mL. 8 tubes containing 0.5mL Standard Diluent were prepared to dilute the stock solution and series concentrations according to the Figure **2**-**3**. 0.5mL of the stock solution was taken from the first tube and was firstly diluted to 10ng/mL as the highest standard as shown in Figure **2**-**3**. The solutions were mixed in each tube thoroughly before transferring to the next tube. The diluted standards were 10ng/mL, 5 ng/mL, 2.5 ng/mL, 1.25 ng/mL, 0.625 ng/mL, 0.312 ng/mL, 0.156 ng/mL, and the last EP tubes with Standard Diluent is a blank at 0 ng/mL.

The stock Detection A and Detection B was spun or centrifuged before use. These were diluted to a working concentration of 100-fold with

Assay Diluent A and B, respectively. The Wash Solution was prepared as 20mL of Wash Solution concentration (30×) with 580mL of deionized or distilled water to obtain a final Wash Solution of 600mL for use (1×).

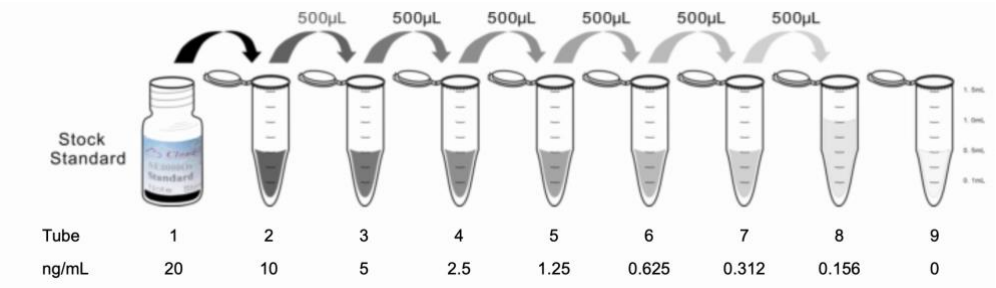


Figure 2-3. The preparation standard samples of ELISA. Adapted from the instruction manual of Cloud-Clone Corp (SEH522Hu).

2.3.3.2 Samples preparation and assay procedure

In brief, 100μL each of diluted standards or serum samples were assayed in 96 well ELISA plates. After incubation for 1h at 37℃ the liquid from each well was removed and 100μL of detection reagent A working solution added and incubated for 1h at 37℃. Each well was then washed with wash solution 3 times. Then 100μl of detection reagent B working solution was added to each well, the plate covered and incubated for 30 min at 37℃. The wash process was repeated 5 times and then 90μL of substrate solution was added to each well, and incubated for 10-20 min at 37℃. Finally, the stop solution (50μL) was added to each well, and NPNT was measured at 450nm immediately using a plate reader. NPNT concentration of serum from AECOPD individuals and healthy controls was compared.

2.4 Molecular Genetic technologies

2.4.1 RNA interference

2.4.1.1 Small interfering RNAs (siRNAs) design

Small interfering RNAs (siRNAs) is a useful tool which can potentially be applied as therapeutic agents against various diseases. In cells, siRNAs are produced by enzymatic cleavage of long dsRNAs by the RNase-III class endoribonuclease Dicer. The siRNAs associate with the RNA induced Silencing Complex (RISC) in a process that is facilitated by Dicer (136).

Traditional 21-mer siRNAs are chemically synthesized RNA duplexes that mimic Dicer products and bypass the need for Dicer processing. Dicer-Substrate RNAs are chemically synthesized 27-mer RNA duplexes that are optimized for Dicer processing and show increased potency when compared with 21-mer duplexes. For NPNT, I used three target-specific duplexes. The siRNA kit also contains a scrambled universal negative control RNA duplex (DS Scrambled Neg) that is absent in human, mouse, and rat genomes.

Product	NPNT (Human)- unique 27mer siRNA duplexes – 2 nmol each	
Catalog No	SR316838	
Duplex Sequences:	SR316838A	CCAUAUGCACUAAGAAUAGAACAAG
	SR316838B	GCACUAAACUGAUGGAAGAAGUUAT
	SR316838C	AGACAAUUAUGUAGAGUGGGCACGT

Table 2-9 NPNT unique 27mer siRNA duplexes, including sequence information.

2.4.1.2 Reconstitution of siRNA duplexes in duplex buffer

Briefly, the tubes were centrifuged to ensure that all material was in the bottom of the tube and not in the cap before opening for the first time. 100µL of RNase-free duplex buffer was used for resuspending 2 nM of each duplex, including the control duplex, vortexed thoroughly to result in 20μM final concentration. The tubes were heated to 94℃ for 2min to ensure that the contents were fully duplexed, then removed from heat and allowed to cool to room temperature, and aliquoted (5μL/tube). Tubes were labelled and placed in a freezer box for further use.

2.4.2 CRISPR-Cas9

Clustered regularly interspaced short palindrome repeats associated Cas9 nuclease (CRISPR-Cas9) technology, which is derived from an adaptive immune system in bacteria, have been modified for genome engineering and successfully used to generate gene knockouts both in human cells and induced pluripotent stem cells (iPS) (137). Rapiteanu et al has developed a new method which has successfully been applied to primary human bronchial epithelial cells (138).

2.4.2.1 gRNA design

Guide RNA (gRNA) containing a mature CRISPR RNA (crRNA) and transactivating crRNA (tracrRNA). DESKGEN libraries (https://www.deskgen.com/landing/#/) are AI-designed and help provide robust guide sequences for use in CRISPR-Cas9 experiments.
For NPNT, I selected the knockout option, and searched by the ENSEMBL gene name. (i.e. NPNT, ENSG00000168743), and then screened for gRNAs.

2.4.2.2 Making gRNA

In order to make the gRNA for use in experiments I combined crRNA and tracrRNA. These were hybridized to 50μM final concentration. By resuspending tracrRNA and crRNA in Duplex Buffer (IDT, 11-01-03-01) to 100μM and then hybridizing the crRNA and tracrRNA to 50μM final concentration.)

Reagent	1 well (NT)	1 well (Custom KO)
100μM crRNA (μL)	5.5	5.5
100μM tracrRNA (μL)	5.5	5.5
Final volume of 50μM hybridised gRNA (μL)	11.0	11.0

The hybridized solution was heated to 95℃ for 5min and left to cool for ~5 min at room temperature.

2.4.2.3 Ribonucleoprotein assembly

The transfection of Cas9 protein pre-complexed with guide RNA (Cas9 ribonucleic protein, Cas9 RNP) was assembled by combining 90pmoles of Cas9 with 450pmoles crRNA-tracrRNA hybrid (molar ratio of gRNA:Cas9 = 5:1) resulting in a 10.5μl volume solution as shown below, and incubated for 5 min at room temperature.

Reagent	1 well (NT)	1 well (Custom KO)
Cas9 10mg/ul (ul)	1.5	1.5
gRNA (ul)	9.0	9.0
Final volume RNP (ul)	10.5	10.5

2.4.2.4 Electroporation enhancer

The electroporation included by mixing RNP with ssDNA oligo (ssODN) then adding 0.9µL of electroporation enhancer (100µM; IDT, 1075916). This complex was stable and could be left at room temperature while cells were prepared.

Reagent	1 well NT	1 well Custom KO
RNP (μL)	10.5	10.5
Enhancer 100uM (μL)	0.9	0.9
Total (μL)	11.4	11.4

2.4.2.5 Preparing primary human bronchial epithelial cells

HBECs cells were used in T75 cm^2 flask when reaching 70% confluency. The cells were washed in HBSS and trypsinised, spun down and then resuspended with 200,000 cells in 8.6μL P3 Primary cell nucleofector solution (Lonza, V4XP-3032) (always keeping one non-electroporated control).

2.4.2.6 RNP solution mixed HBECs

The cell suspension was mixed with the RNP solution and transferred to a well of a 16-well *Nucleocuvette Strip (Lonza, V4XP-3032)*

Component	1 well NT	1 well Custom KO
Cells in P3 (μl)	8.6	8.6
RNP + enhancer (μl)	11.4	11.4
Total (μl)	20	20

2.4.2.7 Electroporate cells

Passage 3 Cells were electroporated using the Amaxa 4D Nucleofector (Lonza) as follows:

The CM-113 program on the Amaxa 4D Nucleofector (Lonza) was used to Electroporate HBECs, and then electroporated cells were diluted by 2mL pre-warmed median in each well and were transferred to 6-well-plates and incubated at 37℃ (Passage 4).

2.5 Gene expression analyses

2.5.1 Microarray

2.5.1.1 Principle and preparations

The total RNA sample was extracted as described above and the RNA integrity number (RIN) measured to assess RNA quality. Microarray procedures were followed as described in the GeneChip™ WT PLUS Reagent Kit manual and GeneChip™ WT Terminal Labelling and Controls Kit protocols (Thermo Fisher Scientific, USA). Microarrays were run using Affymetrix GeneArray Scanner 3000 7G Plus (Affymetrix, USA) (139). The human gene 2.1 array (Catalog number: 902114, Thermo Fisher Scientific, USA) was used: this is a whole-transcript array which include probes to measure both mRNA and long non-coding RNA transcripts (lncRNA). The array covers >33,500 coding transcripts and >11,000 long non-coding transcripts, with a median of 21 unique probes per transcript. Each unique probe is 25 bases in length in the expression array (Figure **2**-**4**).

2.5.1.2 Array pre-processing

Bioconductor contains a number of packages, such as *limma* and *oligo*. The *oligo* package helps to read in and organise information from Affymetrix Human Gene 2.1 ST Array Strip Microarray CEL files. The package *limma* was built primarily for spotted array data. Quality control (QC) is a first step. The R environment provides packages for visualization of distribution and statistical QC, such as box plots.

This is illustrated by the command as shown in the Chapter 9 appendix, which generate image of the box plots. The next step is normalization in analysing spotted array. The command has showed in Chapter 9 appendix.

Further analysis could be done in the *limma* package, which has advantages for designing matrices for spotted array experiments (140). The function *lmFit* applies design matrix information together and estimates of relative mRNA abundance in each sample. The function *eBayers* uses estimates of probability of differential expression for each gene (141). The command has been showed in chapter 9 appendix.

2.5.1.3 Annotation

Affymetrix Human Gene 2.1 ST Array Strip Microarray CEL files were annotated by R Bioconductor (https://www.bioconductor.org/) package (hugene21sttranscriptcluster.db) and RMA-normalized.

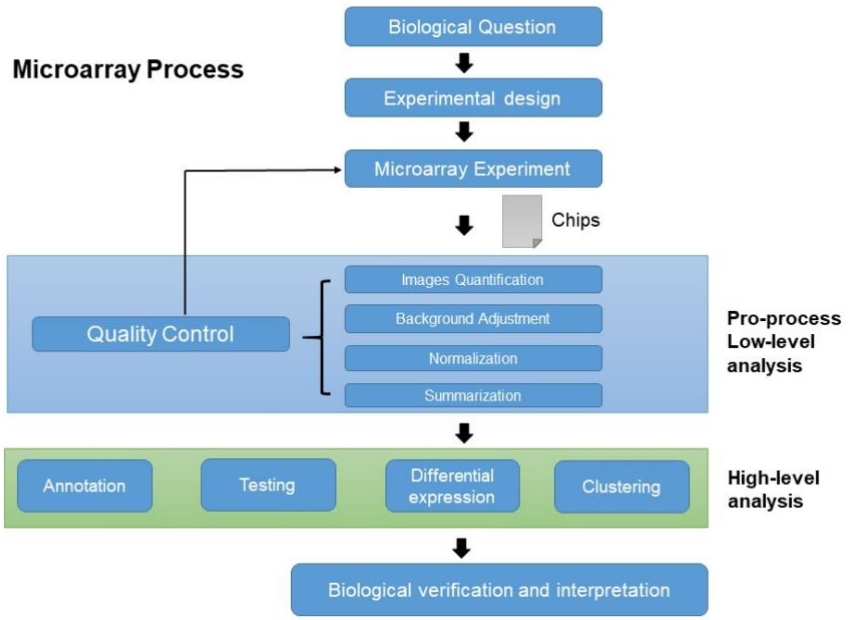


Figure 2-4. Flowchart of microarray process and data analysis. We used the Affymetrix human gene 2.1 array strip to analysis the profiling of genes in HBECs after using NPNT siRNAs. Initial raw data was performed by the R environment to analyse the quality, and output boxplots of quality control. And then, data analysis was performed using R packages to annotate and to generate differentially expressed genes. Gene set enrichment analysis (GSEA) was used to provide insight into the pathways or biological of HBECs after NPNT knockdown.

2.5.2 RNA-sequencing

RNA sequencing (RNA-seq) uses next generation sequencing (NGS) to provide insight into the transcriptome of populations of cells (142). RNA-seq has distinct advantages when compared to previous Sanger sequencing- and microarray-based methods, detecting far higher coverage and greater quantitative view of gene expression, alternative splicing, and allele-specific expression (142). In this thesis, RNA samples from HBECs were sequenced. In brief, the RNA was isolated from HBECs, and converted to complementary DNA (cDNA), sequencing libraries prepared, and sequencing undertaken on an NGS platform (Figure **2**-**5**).

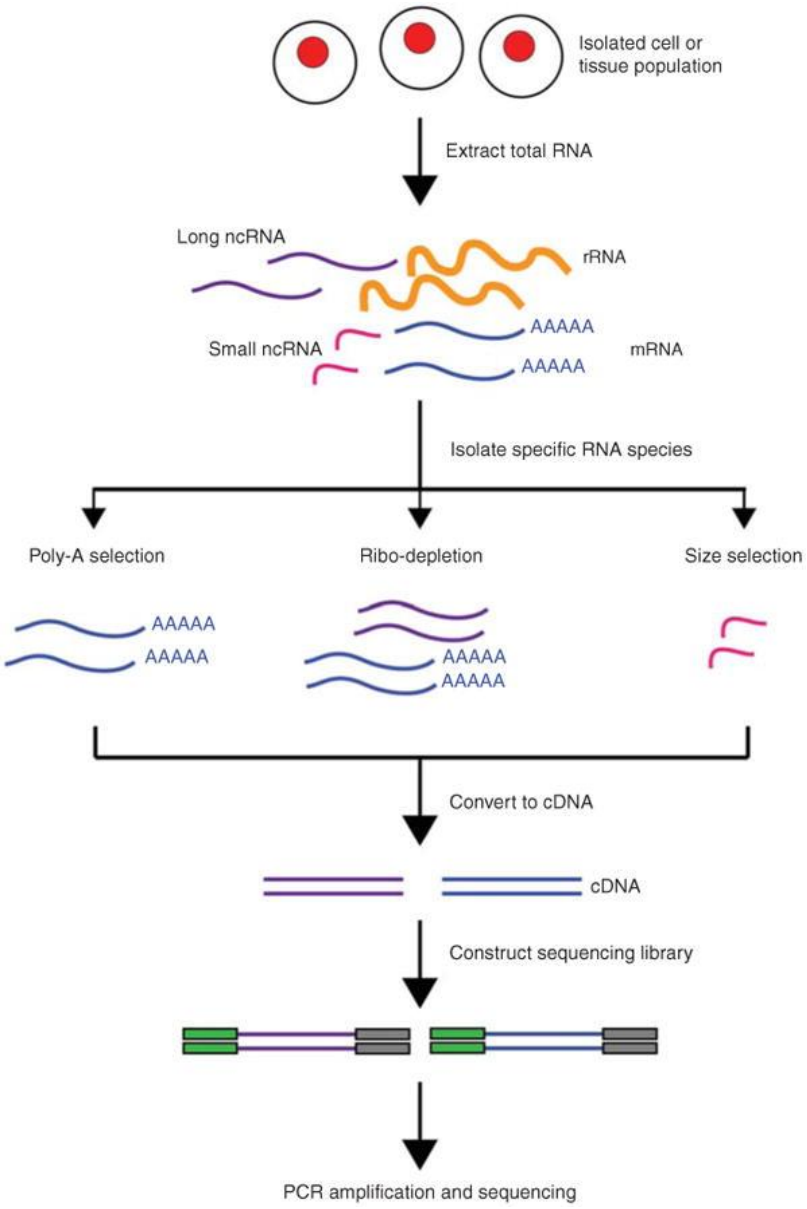


Figure 2-5 Overview of RNA-Seq. First, RNA is extracted from the biological material of choice (e.g., cells, tissues). Second, subsets of RNA molecules are isolated using a specific protocol, such as the poly-A selection protocol to enrich for polyadenylated transcripts or a ribo-depletion protocol to remove ribosomal RNAs. Next, the RNA is converted to complementary DNA (cDNA) by reverse transcription and sequencing adaptors are ligated to the ends of the cDNA fragments.

Following amplification by PCR, the RNA-Seq library is ready for sequencing. (From Cold Spring Harb Protoc, Kukurba et al, 2015)

2.5.2.1 Library Preparation and Sequencing

Library preparation and RNA-sequencing were carried out by Oxford Genomics Centre. RNA was quantified using RiboGreen (Invitrogen, Glasgow, UK) on the FLUOstar OPTIMA plate reader (BMG Labtech, Aylesbury, UK) and the size profile and integrity analysed on the 4200 TapeStation (Agilent). The amount of RNA was adjusted to 100ng prior to library preparation. Polyadenylated transcript enrichment and strand specific library preparation was completed using NEBNext Ultra II mRNA kit (NEB, Hitchin, UK) following the manufacturer's instructions. Libraries were amplified on a Tetrad (Bio-Rad, London, UK) using in-house unique dual indexing primers. Individual libraries were normalizedised using Qubit, and the size profile was analysed on the 4200 TapeStation. Individual libraries were normalizedised and pooled together accordingly. The pooled library was diluted to ~10 nM for storage. The 10 nM library was denatured and further diluted prior to loading on the sequencer.

Paired end (150bp) sequencing was performed using a NovaSeq6000 platform and HiSeq 3000/4000 PE Cluster Kit and 150 cycle SBS Kit (Illumina, Cambridge, UK), generating a raw read count of ~30 million reads per sample.

2.5.3 Processing Sequencing Data

2.5.3.1 Concatenation of FASTQ files

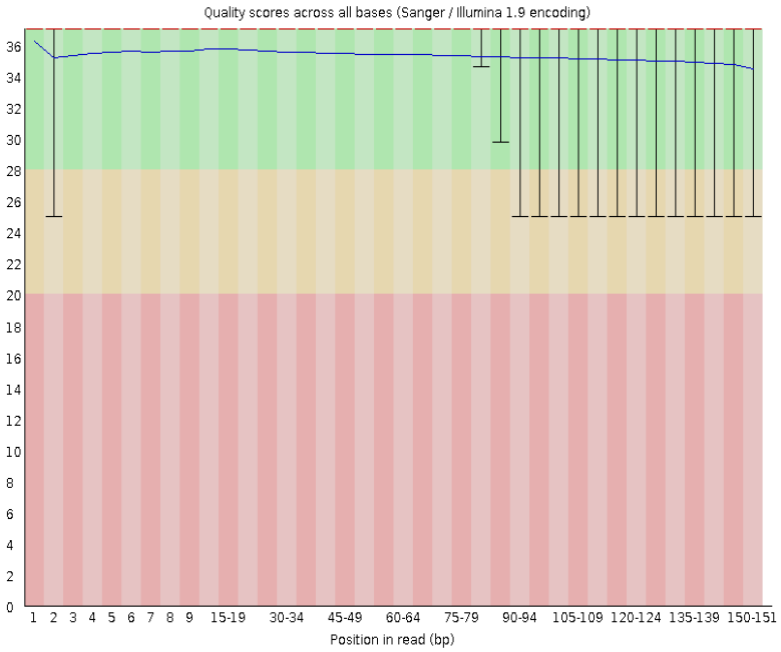
Data were concatenated from the FASTQ files from each of the sequencing lanes. This combines the sequencing data from each of the multiple lanes of sequencing for each sample.

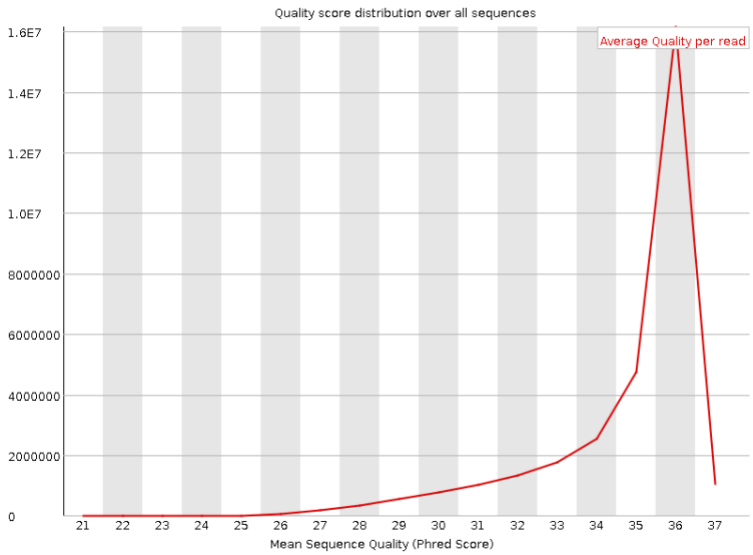
2.5.3.2 Adapter and Quality-Trimming

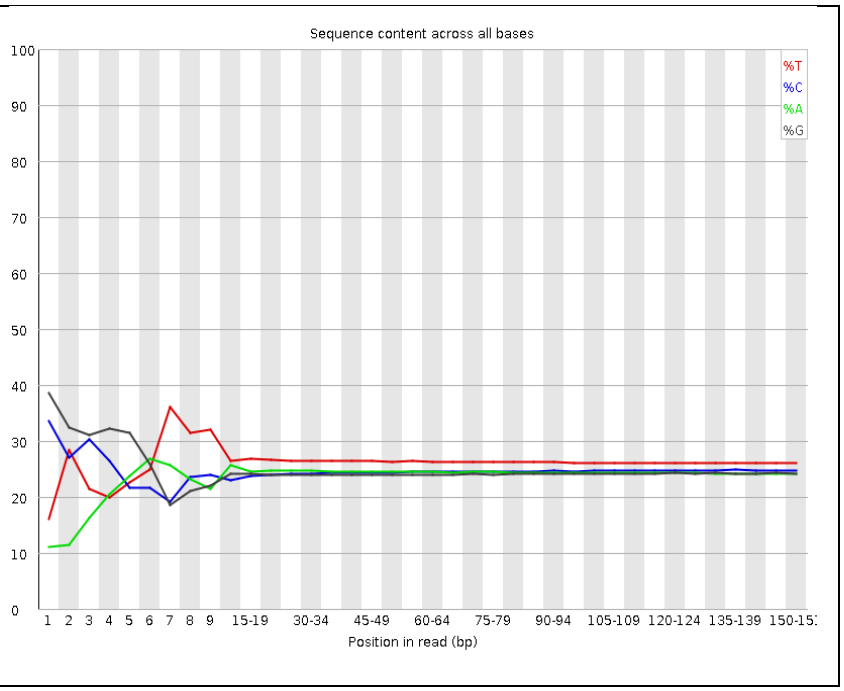
Oxford Genomics Centre provided original files for each forward and reverse read per sample. The adapter sequences were trimmed from the concatenated files using the Scythe programme. Further quality trimming was achieved using the Sickle programme which trims read with a Q score < 20.

2.5.3.3 RNA-seq Quality Control

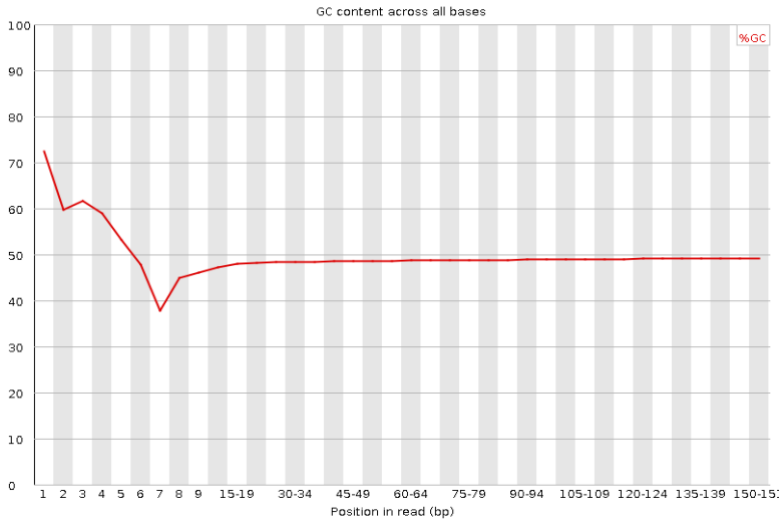
The final step used in quality control was the FASTQ analysis which generates graphical outputs to show the quality of sequence data. An example of the visual output from this is shown below.

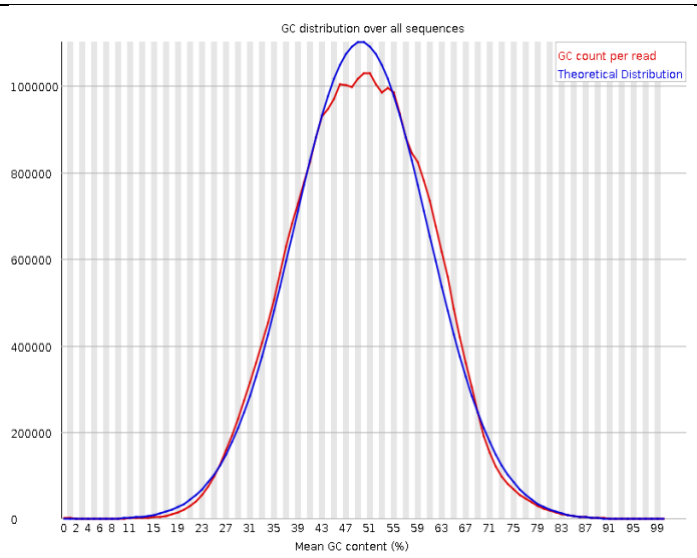
Basic Statistics	Simple composition statistics for the analysed file. For example, filename, file type, total sequences, sequence length and % GC content.
Per Base Sequence Quality	
	An overview of the range of quality values across all bases at each position in the FastQ file. Very good quality calls are represented in the green section, reasonable calls are orange and calls of poor quality are red. A warning will be issued if the lower quartile for any base is < 10 or if the median for any base is < 25. A failure will be issued if the lower quartile for any base is <5 or if the medium for any base is <20.

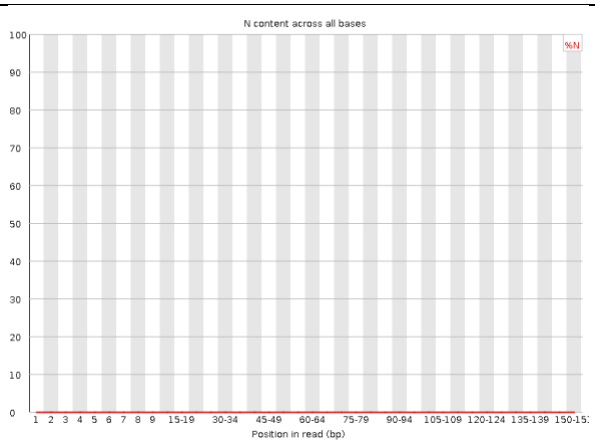
Per Sequence Quality Scores	
	This allows the visualisation of a subset of sequences which have universally low quality scores. A warning is issued if the most frequently observed mean quality is <27 (0.2% error rate) and a failure if the mean quality is <20 (1% error rate).

Per Base Sequence Content	

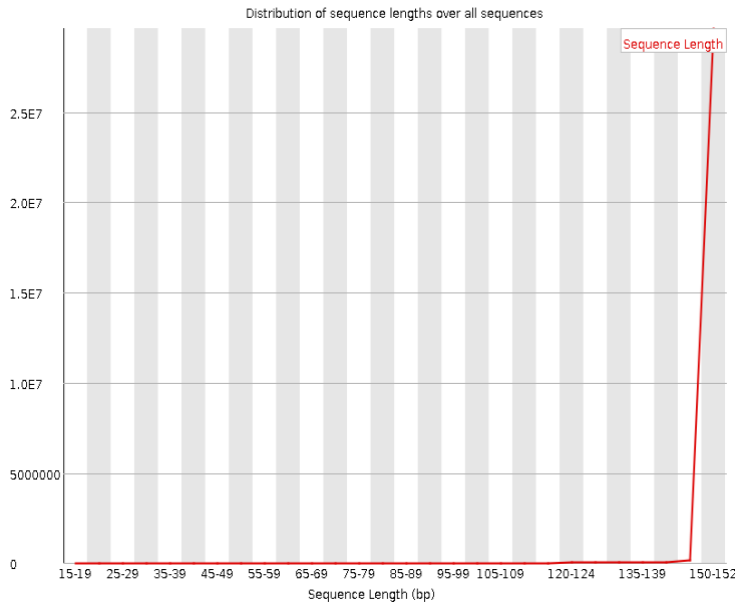
	This plot shows the proportion of each base position in a file for which each of the four normal DNA bases has been called. In a random library, there should be little to no difference between the different bases of a sequence run so the lines should not deviate from each other. A warning is issued if the difference between A and T or G and G is >10% and a failure is issued if this is >20%.

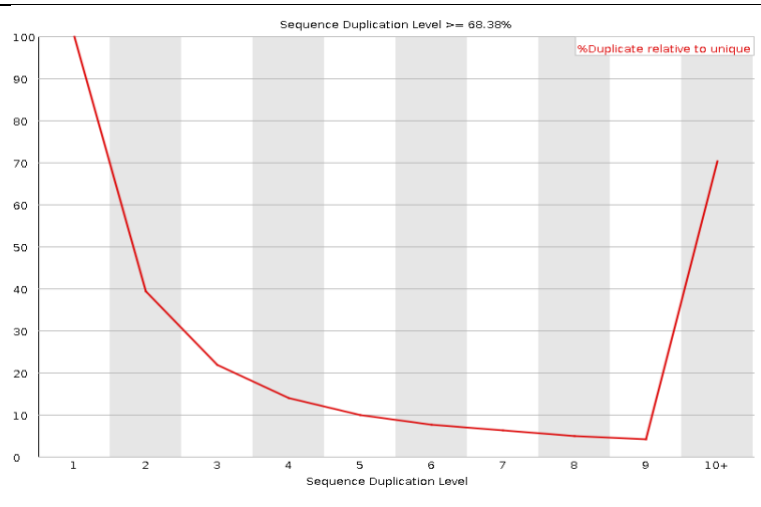
Per Base GC Content	
	This plots out the GC content of each base position. There should be little to no difference between the different bases of a sequence run. The overall GC content reflect the GC content of the genome. A warning is issued if the GC content of any base deviates > 5% from the mean GC content. A failure is issued if this deviates >10%.

Per sequence GC content	
	This measured GC content across the whole length of each sequence and compares it to a normal distribution of GC content. In a normal library you would expect to see a normal distribution of the GC content. A warning equates to a scenario in which the sum of the deviation from the normal distribution represents > 15% of the reads and a failure is >30%.

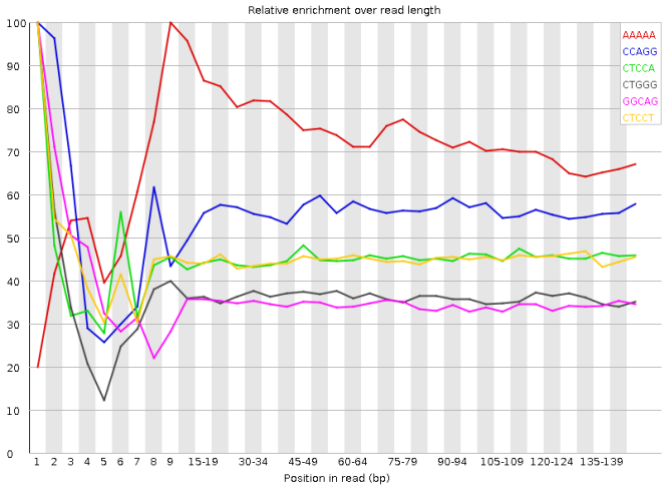
Per Base N content	
	If the sequencer is unable to make a base call with sufficient confidence it places an N rather than a

	conventional base. N content >5% is a warning and >20% is a failure.

Sequence Length Distribution	
	This shows the distribution of fragment sizes. A warning is raised if all the sequences are not the same length and a fail is if any of the sequences have zero length.

Duplicate Sequences	

	Most sequences in a diverse library will occur only once. Low level duplication may indicate very high level of coverage of the target sequence but high level duplication is more likely to be due to some kind of enrichment bias. The plot shows the relative number of sequences with different degrees of duplication. A warning is issued if non-unique sequences make up >20% of the total and >50% constitutes a failure.

Over-represented Kmers	
	This counts the enrichment of every 5-mer within the library. It then calculates an expected level at which this k-mer should have been based on the base content of the library as a whole and then uses the actual count to calculate and observed/expected ratio for that k-mer.

	A k-mer showing >3 fold overall enrichment or >5 fold enrichment at any base will be reported. An error is issued if any k-mer is enriched> 10 fold at any base position.

2.5.3.4 Trimming and alignment

The FastQ files were trimmed of adaptor sequences and poor quality sequences by Trimmomatic. Spliced Transcripts Alignment to a Reference (STAR) was used to align files to the genome and produce BAM files and count files.

For the experiments described in this thesis, FastQ files were received as raw data from the Novaseq 6000 in the Oxford Genomic Centre (150bp reads). The FastQ files were trimmed for adapter sequences and then assess for the quality with FASTQC as previously descripted (143). The FastQ files were transferred into the desired directory on the HPC (using drag and drop) using WinSCP together with the Snakefile, config. yaml and sub_trim.sh files. The reference genome and primary assemblies were downloaded to the terminal. A file containing the adaptor sequence was provided by Oxford Genome Centre to remove reads by Trimmomatic. The STAR software package provided a large mapping set of high-throughput sequencing reads to a reference genome for RNA sequencing data analysis (144, 145). STAR generates several output files including BAM files. Furthermore, it also fully generated a count file for each sample which could be analysed in R studio.

2.5.4 Gene Set Enrichment Analysis (GSEA)

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a gene set shows statistically significant enrichment based on prior knowledge from published papers or co-expression data (146, 147). The aim of GSEA is to confirm whether the members of S (the gene set) are randomly distributed through L (the gene list). All the genes are ordered in a ranked list (L), according to their differential expression between the classes. The core members of high scoring gene sets contribute to the enrichment score (ES). The leading-edge subset is the core of a gene set which comprise the enrichment signal. The advantage of GSEA is that it is can be used to interpret a large-scale experiment by identifying pathways and processes rather than the top expressed genes (Figure **2**-**6** and Figure **2**-**7**).

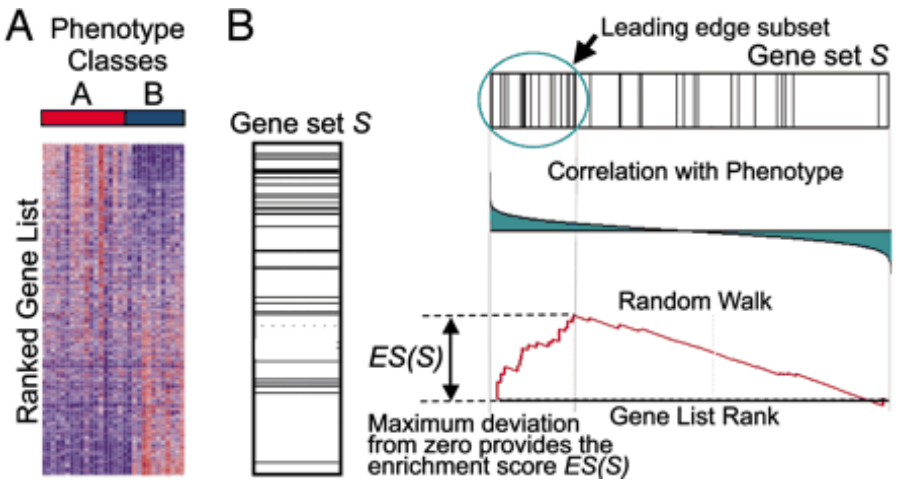


Figure 2-6 A GSEA overview illustrating the method. (A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the “gene tags”, i.e., location of genes from a set S within the

sorted list. (B) Plot of the running sum for S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset.(Adapted from Subramaniian, 2005)*(146)*

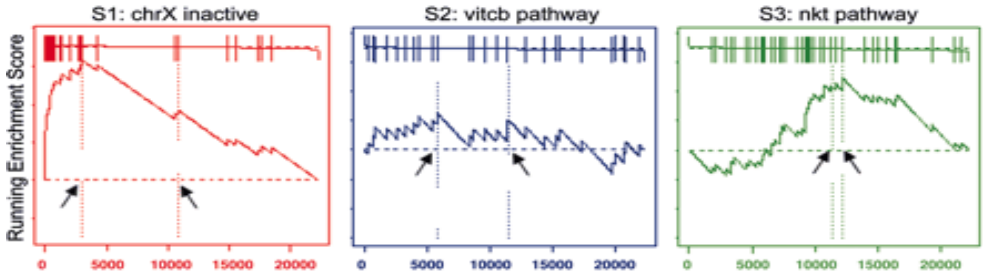


Figure 2-7 Examples of original enrichment score behaviour. The distribution of three gene sets, from the C2 functional collection, in the list of genes in the male/female lymphoblastoid cell line example ranked by their correlation with gender: S1, a set of chromosome X inactivation genes; S2, a pathway describing vitamin C transport into neurons; S3, related to chemokine receptors expressed by T helper cells. Shown are plots of the running sum for the three gene sets: S1 is significantly enriched in females as expected, S2 is randomly distributed and scores poorly, and S3 is not enriched at the top of the list but is nonrandom, so it scores well. Arrows show the location of the maximum enrichment score and the point where the correlation (signal-to-noise ratio) crosses zero. (Adapted from Subramaniian, 2005)*(146)*

GSEA (http://software.broadinstitute.org/gsea/downloads.jsp) was applied in this thesis to analysis the gene sets (pathways) altered after siRNA induced knockdown of NPNT in HBECs. The GSEA metrics are described in Table **2**-**10**. The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets for use with GSEA software. I

used two gene set databases within this to investigate potential pathways altered after NPNT knockdown by siRNA.

The Curated datasets were firstly used in this thesis, these are curated from various sources, including online pathway databases and the biomedical literature. The C2 collection is divided into the following two sub-collections: Chemical and genetic perturbations (CGP) and Canonical pathways (CP).

Hallmark gene sets were used as the second dataset, these summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression.

GS	Gene set name. Click the gene set name for a detailed description of the gene set. For MSigDB gene sets, the description is the gene set page on the GSEA web site. For other gene sets, the description is provided by the author of the gene set.
GS DETAILS	For the top 20 gene sets, click the Details link to display the Gene Set Details Report. To generate the Details link for a different number of gene sets, use the Plot graphs for the top sets of each phenotype parameter on the Run GSEA Page.
SIZE	Number of genes in the gene set after filtering out those genes not in the expression dataset
ES	Enrichment score for the gene set; that is, the degree to which this gene set is overrepresented at the top or bottom of the ranked list of genes in the expression dataset.
NES	Normalized enrichment score; that is, the enrichment score for the gene set after it has been normalized across analyzed gene sets.

NOM p-value	Nominal p value; that is, the statistical significance of the enrichment score. The nominal p value is not adjusted for gene set size or multiple hypothesis testing; therefore, it is of limited use in comparing gene sets.
FDR q-value	False discovery rate; that is, the estimated probability that the normalized enrichment score represents a false positive finding.
FWER p-value	Familywise-error rate; that is, a more conservatively estimated probability that the normalized enrichment score represents a false positive finding. Because the goal of GSEA is to generate hypotheses, the GSEA team recommends focusing on the FDR statistic.
RANK AT MAX	The position in the ranked list at which the maximum enrichment score occurred. The more interesting gene sets achieve the maximum enrichment score near the top or bottom of the ranked list; that is, the rank at max is either very small or very large.
LEADING EDGE	Displays the three statistics used to define the leading edge subset: ●Tags. The percentage of gene hits before (for positive ES) or after (for negative ES) the peak in the running enrichment score. This gives an indication of the percentage of genes contributing to the enrichment score. ●List. The percentage of genes in the ranked gene list before (for positive ES) or after (for negative ES) the peak in the running enrichment score. This gives an indication of where in the list the enrichment score is attained. ●Signal. The enrichment signal strength that combines the two previous statistics: $(\text{Tag }\%)(1-\text{Gene }\%)\left(\frac{N}{N-Nh}\right)$ Where N is the number of genes in the list and Nh is the number of genes in the gene set. If the gene set is entirely within the first Nh positions in the list, then the signal strength is maximal or 100%. If the gene set is spread throughout the list, then the signal strength decreases towards 0%. These statistics describe the leading-edge subset of a single gene set. Use the Leading Edge analysisto analyze the overlap between multiple leading-edge subsets.

Table 2-10 Definitions used in GSEA results interpretation. Adapted from GSEA website. (http://software.broadinstitute.org/gsea/downloads.jsp).

2.6 Statistical analysis

For cell and molecular biology experiments, standard statistical methods were used. To compare, means (parametric) or medians (non-parametric) were used between two or more groups. The student's t-test (parametric) was used to assess normally distributed data and a Mann-Whitney test (non-parametric) was used for data that was not normally distributed between two groups. For analysis of variance across 3 or more groups a one way ANOVA (parametric) or a Kruskal-Wallis test (non-parametric) was performed. Tukey's test was used to compare all columns of data to each other, whereas the Dunn's test was used to compare groups to a selected control group.

The simple and multiple linear regression analysis, variable was carried out before the normal distribution analysis. Simple linear regression analysis was firstly used to evaluate the relationship between variables and NPNT. The variables in which $p<0.1$ were then subjected to multiple linear regression analysis.

All statistical analysis was carried out in SPSS or GraphPad Prism 9.

3. Analyses of the 4q24 locus association signal and expression profiling of NPNT mRNA and protein in human lung and airway structural cells

3.1 Introduction

As discussed in the first chapter, the development of COPD is influenced by genetic factors and environmental exposures. In 2010, a genome-wide association studies (GWAS) meta-analysis identified a new locus in 4q24 (*INTS12-GSTCD-NPNT*) associated with FEV_1 (10). In all, it identified 46 SNPs located in chromosome 4q24, five SNPs harboured in *NPNT*, including rs17331332, which had the smallest p value of variants (Figure **3**-**1** and Table 3-1).

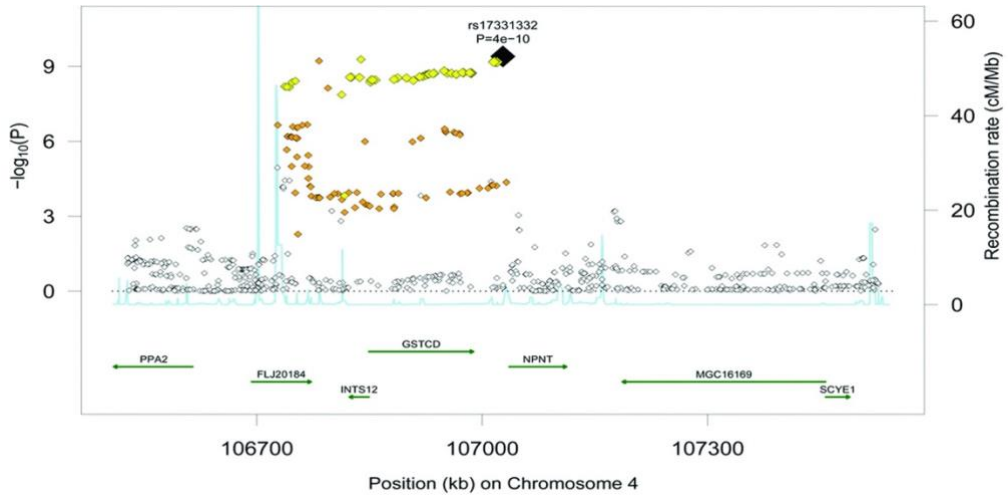


Figure 3-1 Regional association plot for the chromosome 4q24 locus associated with FEV_1 in the CHARGE consortium, which includes *FLJ20184*, *INTS12*, *GSTCD*, and *NPNT*. The lowest *P* value is depicted square black. SNPs are coloured based their r^2 with the labelled SNP. Adapted from Hancock. (10)

Genome-wide significant SNPs ($p<5x10^{-8}$) from meta-analysis of associations with FEV_1 in all participants and in ever smokers from the CHARGE consortium, sorted by GWAS P value. No genome-wide significant associations were observed in never smokers. (Adapted from Hancock, 2010)										
SNP	Chr	Position	SNP type	Gene/ Nearest gene	Reference allele	Allele frequency(a)	CHARGE meta-analysis in all participants		CHARGE meta-analysis in ever smokers	
							β, per-allele change in FEV1 (mL)	P	β, per-allele change in FEV1 (mL)(b)	P©
rs17331332	4	107027556	Intergenic	*NPNT*	A	0.08	60.35	$4.00x10^{-10}$	78.5	$2.24x10^{-9}$
rs17036341	4	107016278	Intergenic	*NPNT*	C	0.93	-58.85	$6.29x10^{-10}$	-76.13	$4.29x10^{-9}$
rs10516529	4	107018765	Intergenic	*NPNT*	G	0.07	58.62	$6.64x10^{-10}$	75.97	$4.24x10^{-9}$
rs11730660	4	107019828	Intergenic	*NPNT*	C	0.07	58.82	$6.68x10^{-10}$	76.2	$4.33x10^{-9}$
rs7660534	4	107015496	Intergenic	*NPNT*	G	0.93	-58.84	$6.70x10^{-10}$	-76.1	$4.58x10^{-9}$

Table 3-1 Five SNPs located in chromosome 4q24 associated with lung function from GWAS meta-analysis by Hancock et al in 2010.

In 2015, Wain et al first reported that rs34712979, which was localised to *NPNT* and was the strongest in never smokers ($p=9.62\times10^{-16}$) (80). Furthermore, further studies have shown a range of variants, including rs34712979 (NC_000004.11:g.106819053 G>A) which is located 5bp upstream of the second exon of *NPNT*, to be associated with extremes of FEV_1 and smoking behaviour (80) and additional fine mapping analysis have also reported rs34712979 (NC_000004.11:G.106819053 G>A) to be implicated in the lung function measures (i.e. FEV_1) and COPD (86) (Figure **3**-**2**, Figure 3-3, and Figure **3**-**4**). Saferali and colleagues reported that the variant rs34712979 is related to a splice junction in *NPNT*, particularly for the junction connecting the 2nd and 4th exons (chr4:105898001-105927336) (122). This study also demonstrated that rs34712979 is highly conserved in different populations and that the promoter or enhancer elements overlap this SNP in many cell types. The minor A allele of rs34712979 was associated with risk of reduced lung function and increased COPD (122).

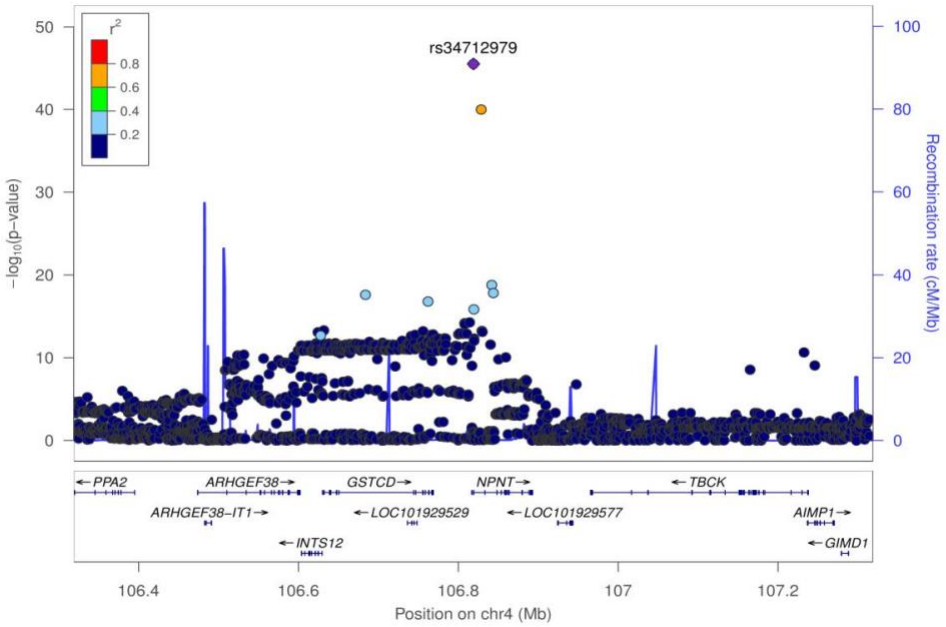


Figure 3-2 Regional association plot for rs34712979 (*NPNT* locus at 4q24) associated with COPD. Adapted from Sakornsakolpat, 2019 (86).

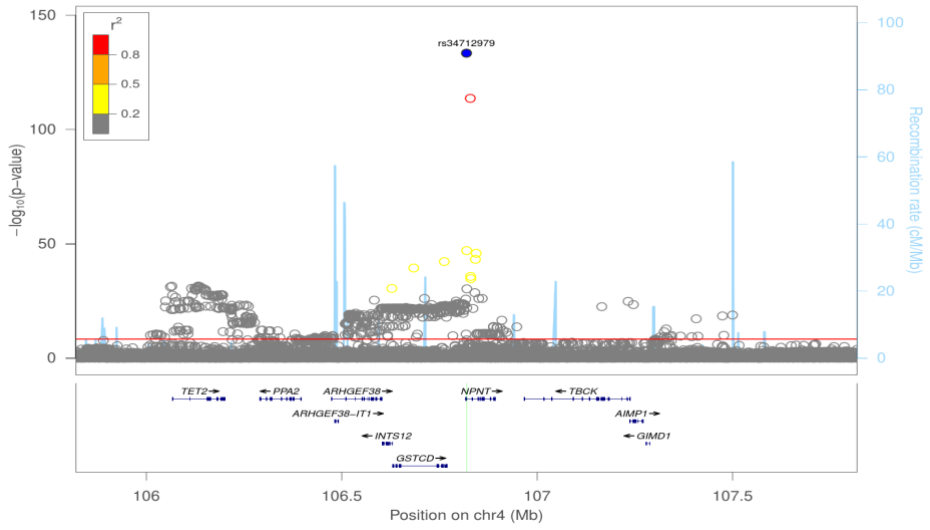


Figure 3-3 *NPNT*, previous signal rs34712979 associated with FEV_1/FVC. Adapted from Shrine, 2019 (12).

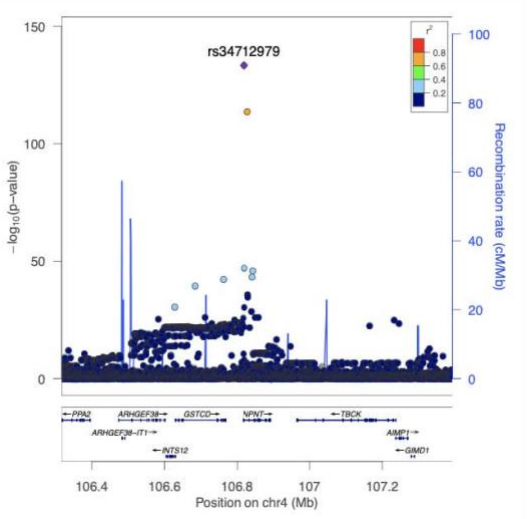


Figure 3-4 Local association plots for genetic association near *NPNT* from GTEx Leafcutter lung sQTL analyses. Adapted from Saferali, 2020 (122).

Linkage disequilibrium (LD) is defined as a non-random association between alleles at two loci and is commonly represented by the r^2 metric (0-1) with the higher the value representing the higher the likelihood that the alleles are inherited together (148). LD can extend beyond the physical chromosomal location and is a sensitive indicator of the population genetic forces that structure the genome (148). The extent of LD can vary due to evolutionary history and population structure, in part due to new variation from insertions, deletions and rearrangement. However, the presence of LD at a locus can make it difficult to identify the gene most likely to be driving functional effects underyling an association signal.

Shrine et al reported in 2019, 107 putative casual genes underlying association signals for lung function across the genome. However, only SNP rs34712979 was both an expression quantitative trait locus (eQTL) and protein quantitative trait locus (pQTL) signal, in both cases as would be expected for *NPNT* (12). An eQTL is one of the fundamental methods to interpretate the effects of genetic variants by using gene expression as a phenotype. There can be both *cis* and *trans* acting variants in the eQTL mapping studies. Here, variants within 1Mb (megabase) up- or down-stream of a gene's transcription start site (TSS) were defined as *cis* (Figure **3**-**5**), whereas the *trans* acting variants were defined as more than 5Mb on both sides from the TSS or on a different chromosome (149).

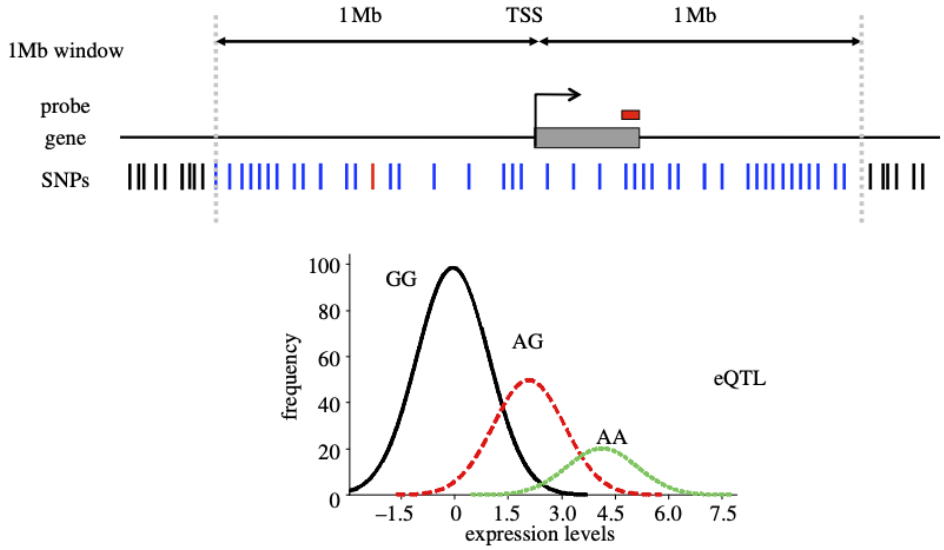


Figure 3-5 A typical eQTL; many SNPs tested against levels of expression measured by a probe or by other means. Adapted from Nica, 2012 (149).

Although GWAS studies have achieved great success, there is a gap between identifying susceptibility genetic variants and understanding the functional basis whereby these contribute to a given phenotype and/or disease. There are a number of approaches which can be used to help address this, including expression studies. As described in the general introduction (chapter 1) there is preliminary evidence that NPNT is expressed in several tissues and cell types, including in lung, airway epithelial cell. NPNT plays a vital role in the acute kidney injury (96), bone homeostasis (106), and cardiac development of zebrafish (115). In addition, it is reported that *NPNT* mRNA is expressed in the foetal and adult lung, and potentially implicated in foetal lung development (100). In recent studies, NPNT has also been measured at the protein level by ELISA in serum, for example in silicosis and nephrotic syndrome (99, 105). Lee and colleagues demonstrated that the serum level of NPNT protein was higher in silicosis patients than healthy controls, and suggested NPNT as a suitable biomarker to monitor the progression of fibrosis in silicosis individuals (99). However, the potential functional or biological role of NPNT in lung diseases, and especially chronic airway diseases, remains largely unknown.

3.2 Aims

The aim of the work described in this chapter were to identify potentially causal variants in the 4q24 locus associated with the lung function and to investigate the expression profiles of *NPNT* mRNA or protein in foetal and adult lung tissue and in airway structural cells.

Specifically, the aims were:

- To identify potentially causal SNPs located in the 4q24 locus
- To investigate the profile of *NPNT* mRNA and protein in foetal lung tissue
- To investigate the expression of mRNA and protein of NPNT in a variety of human cells and tissues, including bronchial epithelial cells, bronchial smooth muscle cells, lung tissues (COPD individuals and healthy controls), kidney and brain.

3.3 Methods

General methods are described in chapter 2. The following additional details are for specific work described in this Chapter.

3.3.1 Bioinformatic analysis of rs34712979 signal

LocusZoom (http://csg.sph.umich.edu/locuszoom) was accessed from a web interface. This is a user-friendly web-based search engine that performs fast visual display of GWAS results in a publication-ready format (150). The LocusZoom plots show an index SNP and a window size, visualizes the gene names and size of flanking region with start and stop positions (150).

In collaboration with colleagues from Leicester University, this tool was used to investigate the region around rs34712979. We attached the rs34712979 500kb up- and down-stream GWAS results for the meta-analysis of FEV_1/FVC from UK Biobank and SpiroMeta using Locuszoom (http://locuszoom.org/genform.php?type=yourdata) to tweak display options. Hg19/100 Genomes 2014 Eur was used as the LD option in the browser.

3.3.2 Identification of NPNT expression in public dataset

3.3.2.1 Determining NPNT mRNA expression

Some high-throughput data are available in public databases. I looked for the expression of *NPNT* mRNA expression by RNA-seq in Protein Atlas (https://www.proteinatlas.org/) RNA sequencing was studied in 95 human individual tissue samples representing 27 different tissues for

the purpose of determining tissue-specificity of all protein-coding genes in the Human Protein Atlas (HPA) RNA-seq normal tissues project. The mRNA expression profiles could then be combined with the analysis with antibody-based profiling of the same tissues.

3.3.2.2 NPNT expression in GEO database

The Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) is an international public database that archives and freely provides outputs from original high-through gene expression studies and other genomics data sets (151). I used the GEO database to analyse the expression of NPNT mRNA in foetal lung tissues, COPD individuals and healthy controls.

The first high-throughput of gene expression data set (GSE14334, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14334) was designed to detect the developing lung transcriptome in RNA samples derived from 38 human foetal lung samples (estimated gestational age 7-22 weeks or 53-154 days post conception) (152). Affymetrix U133 Plus 2 array was used to measure mRNA. The data was downloaded from GEO and I analysed the value of *NPNT* expression during the different stages of lung development.

Furthermore, the second microarray data from GEO dataset GSE38974 was used (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38974). The lung tissue samples from smokers with COPD (n=26) and smokers with no evidence of obstructive lung disease (n=9) confirmed by spirometer

were annotated by Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Feature Number version) (153). I compared the *NPNT* expression level between COPD smoker and smoker with no obstructive lung disease as control group.

3.3.2.3 Expression quantitative trait locus (eQTL)

The Open Targets Genetics Platform (https://genetics.opentargets.org/) provides an open resource which integrates evidence from genetics, genomics, transcriptomics, drugs, animal models and drug-target identification (154). I integrated GWAS results with eQTL from the Open targets Genetics platform (155).

3.3.2.4 NPNT protein expression in the Protein Atlas

The Protein Atlas Portal (https://www.proteinatlas.org/) is designed to explore the human proteome using gene centric and genome-wide antibody-based profiling on tissue microarray (156, 157). In this program, they have analysed 32 tissue types, representing all major tissues or organs in human body, and have detected 20,344 putative protein-coding genes with RNA-seq data (157).

I searched NPNT as gene symbol and listed all the tissues which have been detected the expression of NPNT protein. The expression levels of NPNT in the human body were sorted as low, medium and high level based on the score.

3.3.3 Phenome-Wide Association Studies (PheWAS)

Phenome-wide association studies (PheWAS) can be applied to discover associations between a given variant and a wide range of phenotypes using data from resources, such as UK Biobank (158). Gene Atlas (http://geneatlas.roslin.ed.ac.uk/) is a database which could contains 778 traits and 30 millions of genetic variants and can be used to undertake PheWAS analyses using 452,264 UK Biobank British individuals (159).

In brief, I selected one variant (rs34712979) from the GWAS meta-analysis to study using PheWAS. I then selected the most significant traits.

3.3.4 Determining *NPNT* mRNA expression by TaqMan PCR

RNA was extracted from primary human bronchial epithelial cells at passage 3 using the GenElute™ Mammalian Total RNA Miniprep Kit (Sigma-Aldrich) using the protocol outlined in the Chapter 2. cDNA was synthesised by SuperScript™ First-Strand Synthesis System for RT-PCR (Invitorgen) as in the protocol in Chapter 2. The other cDNA samples were provided by Dr Amanda Henry (Division of Respiratory Medicine, University of Nottingham, UK) as follows: HASMs, differentiated HBECs, BEAS-2B, PBMC, and the whole tissue cDNA from Kidney, Brain, Placenta and Lung. Finally, TaqMan PCR was undertaken as described in the protocol in Chapter 2.

3.3.5 Determining *NPNT* mRNA and isoform expression in HBECs and HASMs by RNA sequencing

RNA-seq data of *NPNT* mRNA and isoform expression in HBECs or HASMs was provided by Robert Hall (Division of Respiratory Medicine, University of Nottingham, UK). The human bronchial epithelial cells from donors (D1, D2, D3, D4, D7, D8, D9 and D11) were cultured (126, 160). RNA was extracted from cell lysates using the RNeasy Mini Kit (Qiagen, USA). Similarly, the human airway smooth muscle cells were cultured from donors (D006, D007, D0023, D0063 and D00529) and then RNA extracted for RNA-seq (126). The extracted RNA was checked to confirm the quantity and quality by nanodrop and Agilent 2100 Bioanalyser, and then sent to Oxford Genomics Centre for RNA sequencing as previously described in Chapter 2.

3.3.6 Determining NPNT protein expression by Immunohistochemistry

Immunohistochemistry was used to detect the expression of NPNT protein expression in foetal lung tissues, as well as the lung tissue from adult COPD and healthy control. The kidney tissue was used as a positive control tissue as previously reported in literature.

In brief, the Immunohistochemistry standard protocol was processed as in Chapter 2 our previous publication (161). The slides were removed of paraffin was in Histoclear, then placed in ethanol and washed. The slides were put in the steamer and covered by Sodium Clitrate buffer. The slides were added primary antibodies one hour before being

blocked by peroxidase blocking solution. The secondary antibodies were incubated for a half-hour at room temperature. The substrate-chromogen solution (DAB) (Agilent Dako) was applied to cover the specimen and incubated for 5 minutes. The slides were counterstained in Mayers haematoxylin, and then were washed by tap water, different concentration of ethanol. Slides were mounted by Vectamount and examined by microscopy.

3.4 Results

3.4.1 Genetic association signal for lung function in NPNT

In order to visualize the functional genetic variant, the web-based tool of LocusZoom was used as described above to provide a graphic display of locus specific association results. The result of LocusZoom plot shows association *P*-value on the -log10 scale on the vertical axis, and the chromosomal position along the horizontal axis (150).

As shown in Figure 3-3, the SNP rs34712979 located in *NPNT* locus as the most significantly genetic variant which associated with the FEV_1/FVC of meta-analysis UK Biobank and SpiroMeta Consortium data. This, together with the observed eQTL and pQTL signals for this variant, highly suggest that *NPNT* may be the causal gene underlying the observed association. The expression and function of *NPNT* was therefore investigated except *GSTCD* or *INTS12* of the locus 4q24 in airway structural cells and lung.

3.4.2 NPNT expression profiling from public database

3.4.2.1 Quantitative transcriptomics analysis: mRNA and protein profile

I used publically available data sets to study *NPNT* mRNA expression in all major human tissues or organs samples using RNA sequencing data. The Reads Per Kilobase Per Million Mapped Reads (RPKM) value of all samples was represented at the gene expression level (162). As shown in Figure 3-6, I found that *NPNT* mRNA expression in lung tissue is the second highest expression in human tissue samples, lower than thyroid, higher than prostate, small intestine and kidney. The RPKM value of skin, liver and bone marrow are low.

NPNT protein expression was investigated in the Protein Atlas Portal (https://www.proteinatlas.org/). As shown in Figure 3-7, in human lung tissue NPNT protein was the highly expressed consistent with the mRNA expression results in these samples. NPNT was highly expressed in pneumonocytes and fibroblasts (Figure 3-8). Furthermore, the kidney and pancreas showed medium level expression of NPNT protein of human tissues or organs. The NPNT protein was not detected in urinary bladder, thyroid gland, spleen, placenta, ovary, lymph node, liver, bronchus, breast, bone marrow, and adrenal gland base on the antibodies-based detection.

Interestingly, the RPKM value of thyroid was the highest in 27 tissues/organs, however, there was not detected of protein. On the

other hand, the RNA-seq data and antibody-based detection of protein were confirmed that NPNT is highly expressed in lung tissue.

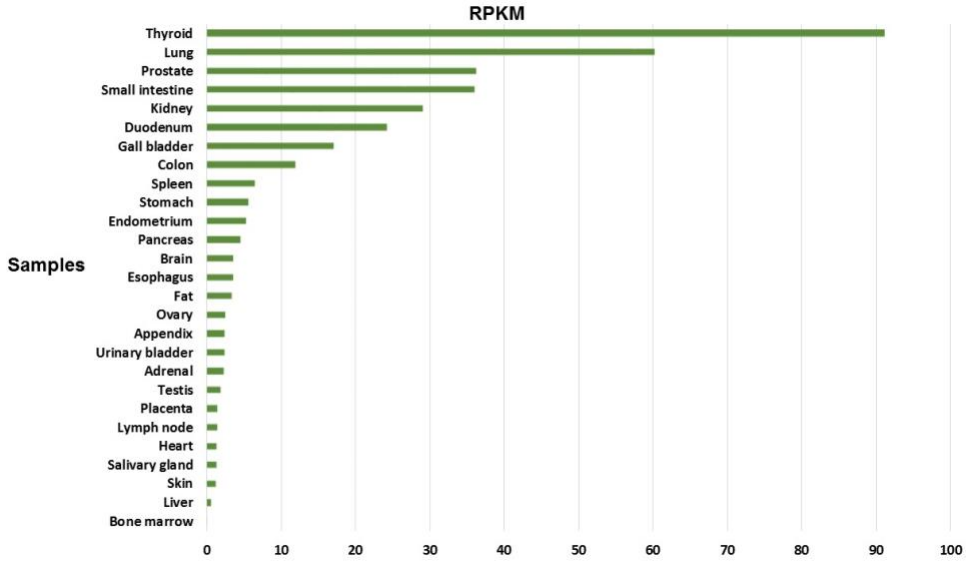


Figure 3-6. RNA-seq data showing RPKM expression values of *NPNT* in 27 types of tissues or organs. Expression was highest in thyroid, followed by lung, prostate, small intestine and kidney. RPKM, Reads Per Kilobase Per Million Mapped Reads. Data from Proteinatlas (https://www.proteinatlas.org/).

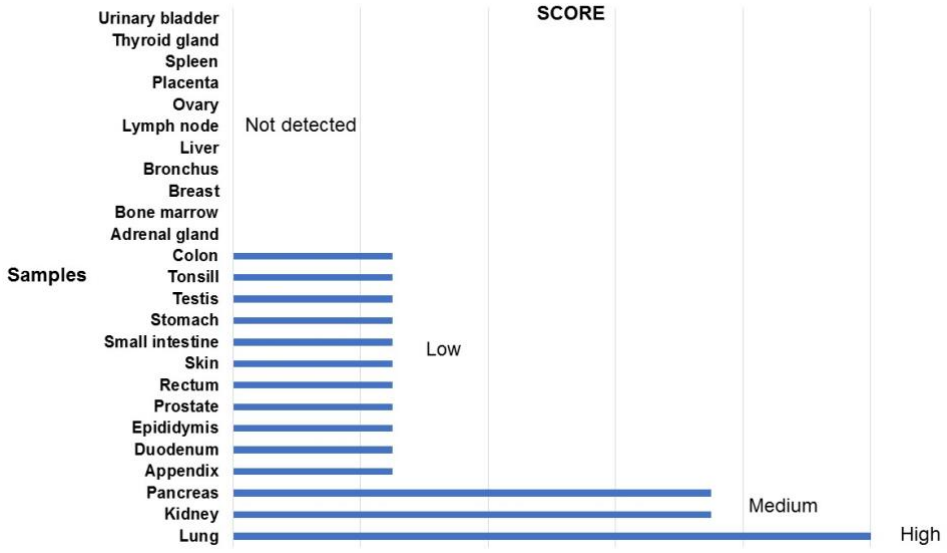


Figure 3-7. NPNT protein expression overview human tissues/organs. The NPNT protein expression is highest in lung tissue, followed by kidney, pancreas. The NPNT protein was not detected in urinary bladder, thyroid gland, spleen, placenta, ovary, lymph node, liver, bronchus, breast, bone marrow, and adrenal gland. The data were accessed on 14th March, 2019. Data from Proteinatlas (https://www.proteinatlas.org/).

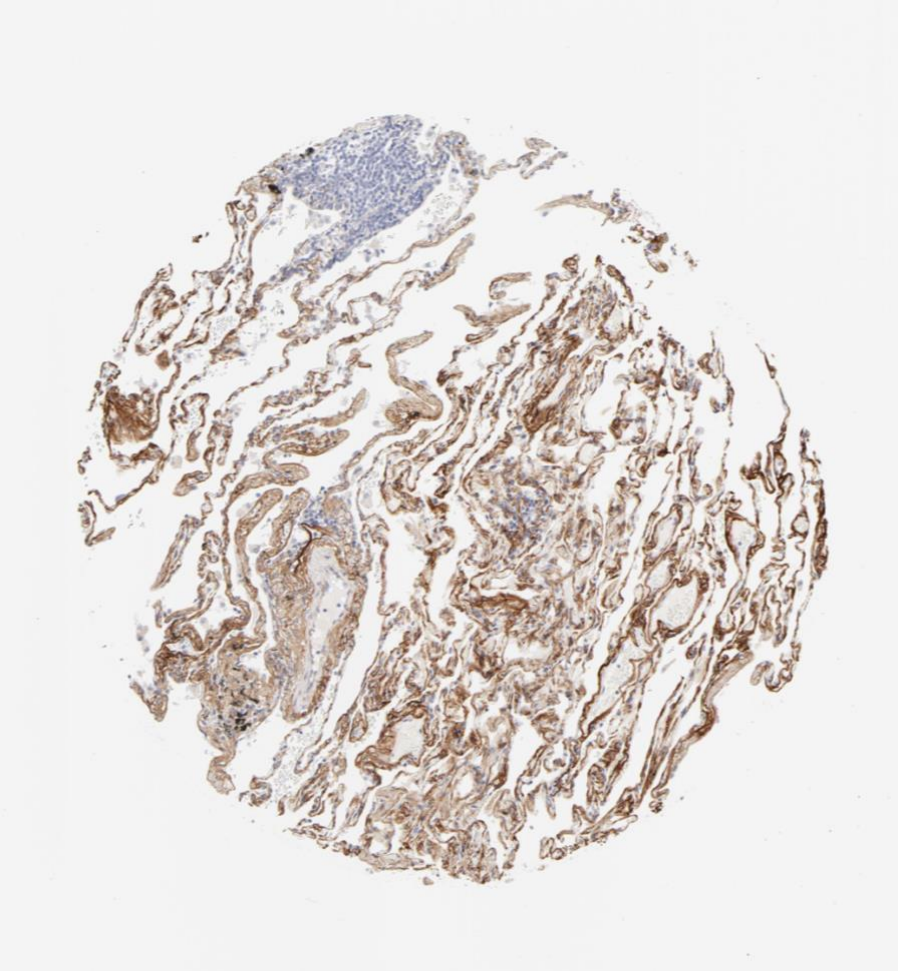


Figure 3-8 NPNT protein was highly expressed in the normal lung tissue from a 69-year-old man (patient id: 2373) in proteinatlas website (https://www.proteinatlas.org/ENSG00000168743-NPNT/tissue/lung#img).

3.4.2.2 NPNT expression in GEO datasets of foetal lung tissues

Be specific, *NPNT* mRNA expression in foetal lung samples was assessed using a publically available database as described above in the methods section. In this dataset, 38 human foetal lung tissues from gestational ages 7-22 weeks were anlaysed. *NPNT* mRNA expression level measured by one probe 225911_at was found to elevate steadily with increasing foetal lung development across the Pseudoglandular and into the Canalicular stages (Figure 3-9). The expression intensity of *NPNT* in probe of 225911_at was higher than for probe of 244747_at.

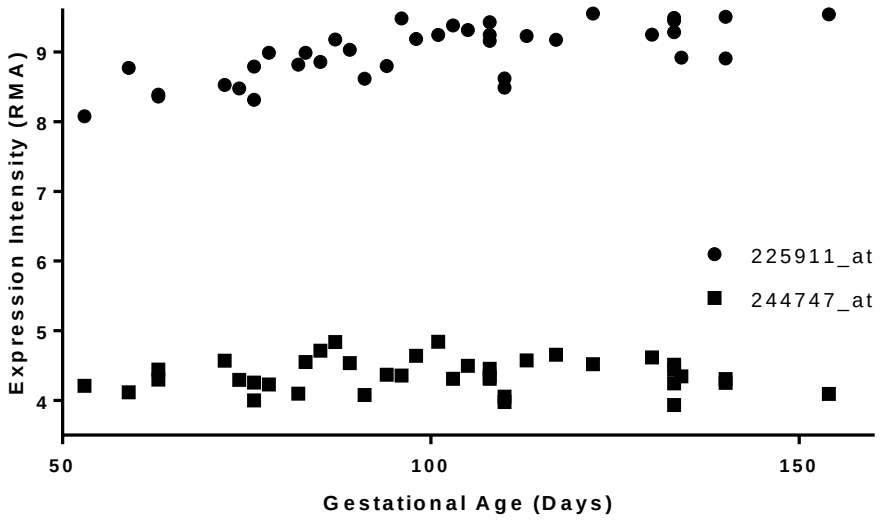


Figure 3-9. *NPNT* probes expression from 38 foetal lungs measured by Microarray. *NPNT* probes 225911_at and 244747_at were plotted of each foetal lung samples with stages and 225911_at showed an increase in expression with foetal lung age.

3.4.3 SNP rs34712979 are eQTL for *NPNT*

To assess whether SNP rs34712979 may regulate *NPNT* gene expression, a further eQTL anlaysis was performed. In brief, I integrated GWAS results with eQTL from the Open targets Genetics platform (https://genetics.opentargets.org/) (155). The study-trait page in the Open Targets Portal for the associated trait COPD of the rs34712979 was selected (163). This confirmed that NPNT is likely to underlie the observed association signal seen at this locus, with there being less evidence that other genes such as GSTCD or INTS12 at this locus were the primary driver of the association signal. (Table **3**-**2**).

Lead Variant	Trait	P value	Odds Ratio	95% Confidence Interval	Author (Year)
4_105897896_G_A	Lung function in never smokers (low FEV1 vs high FEV1)	1.0e-15	1.3	1.2, 1.3	Wain LV (2015)
	Chronic obstructive pulmonary disease x ever smoker interaction (main effect)	1.0e-29	1.2	1.2, 1.3	Kim W (2020)
	Chronic obstructive pulmonary disease in never smokers	1.0e-29	1.2	1.2, 1.3	Kim W (2020)
	Chronic obstructive pulmonary disease in non-current smokers	6.0e-39	1.2	1.2, 1.2	Kim W (2020)
	Chronic obstructive pulmonary disease x current smoker interaction (main effect)	6.0e-39	1.2	1.2, 1.2	Kim W (2020)
	Chronic obstructive pulmonary disease	3.0e-46	1.2	1.2, 1.2	Sakornsakolpat P (2019)
	Lung function in heavy smokers (low FEV1 vs high FEV1)	1.0e-8	1.2	1.1, 1.2	Wain LV (2015)
	Chronic obstructive pulmonary disease in ever smokers	2.0e-16	1.1	1.1, 1.2	Kim W (2020)
	Asthma/COPD (KELA code 203)	3.6e-8	1.1	1.0, 1.1	FINNGEN_R5 (2021)

Table 3-2 GWAS lead variants are linked with variant rs34712979 which found in the Open Target Portal. A allele of 4_105897896_G_A is risk factor for lung function in never smokers or heavy smokers, COPD in never/ever smokers or non-current smokers, Asthma/COPD.

3.4.4 Phenome-Wide Association Studies (PheWAS)

Phenome-wide association study (PheWAS) was used to identify additional genetic relationships and explore novel phenotype association with genetic variants **(164)**.Gene Atlas (http://geneatlas.roslin.ed.ac.uk/) was used to identify variants association with traits. As described above, SNP rs34712979 is localised at *NPNT*, although LD also spans *INTS12* and *GSTCD* **(80)**. As shown in Figure **3**-**10** and Table **3**-**3**, SNP rs34712979 was associated with a number of phenotypes, including those which would be expected based on the known lung function association, such as COPD and asthma, but in addition other traits including arm fat percentage, as well as weight. The figure shows log p values for these associations, however, given that GeneAtlas uses data from >1000 traits (although many are not true independent variables) multiple testing remains a potential issue in these exploratory analyses.

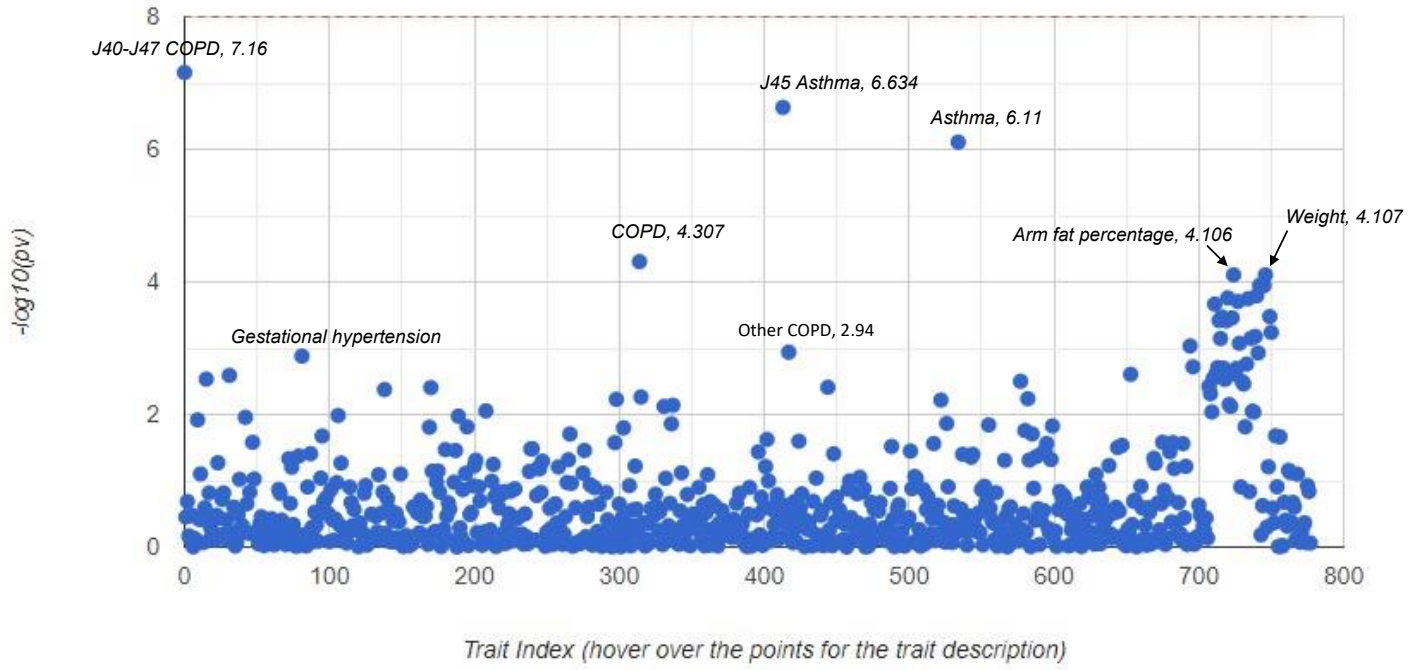


Figure 3-10 Results of PheWAS assessing the genetic variant rs34712979. Y axis represents the significance of the association of the SNP with each phenotype. The phenotypes are represented across the x axis. Phenotypes passing a nominal p value threshold are indicated by descriptors.

Position	Eff. Allele	Trait	Beta	p-value	MAF	HWE	OR beta*
106819053	G	J40-J47 Chronic lower respiratory diseases	-0.0033	6.9202e-08	0.2565	0.7757	0.955
106819053	G	J45 Asthma	-0.0029	2.3226e-07	0.2565	0.7757	0.952
106819053	G	asthma	-0.0035	7.7646e-07	0.2565	0.7757	0.966
106819053	G	chronic obstructive airways disease/copd	-0.0005	4.9357e-05	0.2565	0.7757	0.848
106819053	G	Weight	0.1060	7.8137e-05	0.2565	0.7757	

Table 3-3 Phenotypes associated with a genetic variant (rs34712979). The beta are negatively for the G allele of number of phenotypes, including chronic airway diseases, such as COPD and asthma, except weight where it is positive. Thus, the G allele associated with increased weight is the one associated with reduced risk of COPD, asthma etc. These data were exported from gene atlas (http://geneatlas.roslin.ed.ac.uk/).

3.4.5 *NPNT* mRNA expression in the lung tissues and airway relevant cells

Quantitative PCR (TaqMan) was carried out on cDNA synthesised from RNA from a wide range of human tissue samples and cells. The house-keeping gene was 18S and the TaqMan *NPNT* PDAR spans exons 4-6 and covers all known protein coding variants of *NPNT*, as descripted in Chapter 2. The primer was designed and purchased from Thermo Fisher Scientific.

I found that *NPNT* is highly expressed in kidney and lung tissue, with very low expression in placenta and brain tissues. Compared to the lung tissue which was normalised as standard, the expression levels of *NPNT* in submerged cultures of HBECs, differentiated HEBCs, and BEAS-2B were 2.88±3.74%, 24.18±12.67%, and 2.66±3.47%, respectively (n=3). *NPNT* mRNA expressions were very low in the HASMCs (1.29±0.99%, compared to lung, n=3), and could not be detected in PBMC (Figure **3**-**11**).

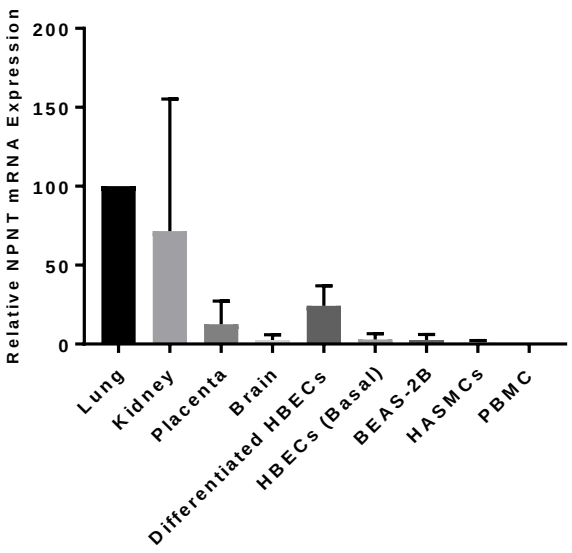


Figure 3-11. *NPNT* mRNA expressed in multiple human tissues and cells. TaqMan PCR was used to investigate the *NPNT* mRNA expression in human tissues (lung, kidney, placenta and brain), a range of human airway cells (basal HBECs, differentiated HBECs, BEAS-2B and HASMs) and the PBMC (mean$\pm\boldsymbol{SD}$, n=3).

3.4.6 *NPNT* mRNA expression in HBECs and HASMs using RNA-seq

RNA was extracted from 5 HASMs donors (passage 3) and 8 HBECs donors from Lonza (passage 2), and sent to the Oxford Genomics Centre for RNA sequencing, and then the sequencing data was analysed by Cuffnorm package to provide FPKM (fragments per kilobase of transcript per million fragments mapped) values in each sample. *NPNT* mRNA expression value values determined by RNA sequencing are shown in Figure **3**-**12**. There was some variability between donors in expression levels with NPNT mRNA expression values of two HBECs donors were higher than 10. Overall, there was no significant different in average FPKM values between HBECs (n=8, mean=4.66) and HASMs (n=5, mean=1.62) donors (Figure **3**-**13**).

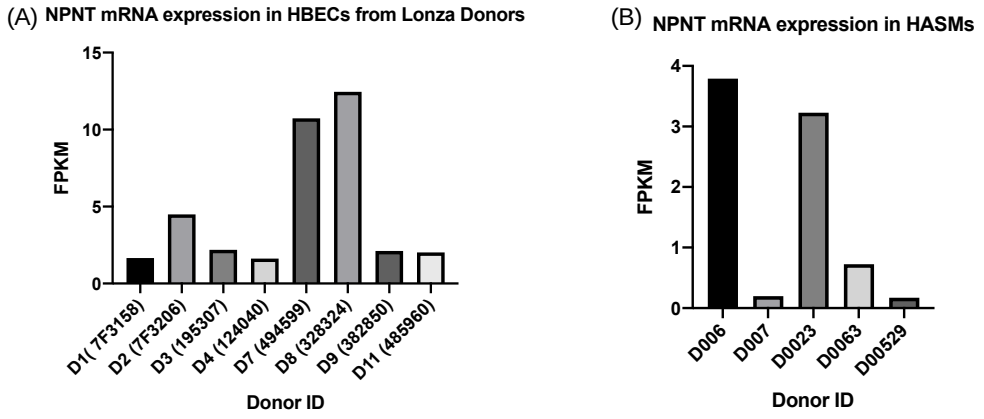


Figure 3-12 *NPNT* mRNA expression in HBECs and HASMs. The *NPNT* mRNA expression by FPKM in individuals HBECs from Lonza donors (A), and HASMs isolated from human lung tissue in Respiratory Medicine Division, University of Nottingham (B). FPKM, fragments per kilobase of transcript per million fragments mapped.

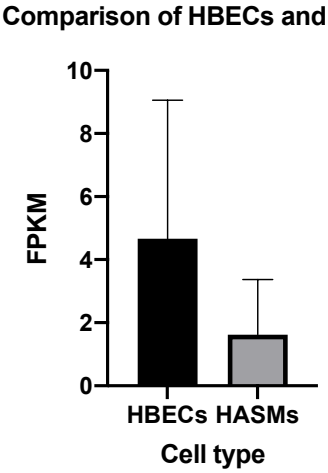


Figure 3-13 Comparison of *NPNT* mRNA expression both in HBECs and HASMs donors. There is no significant difference in average FPKM values between HBECs (n=8, mean=4.66) and HASMs (n=5, mean=1.62) donors. FPKM, fragments per kilobase of transcript per million fragments mapped.

3.4.7 *NPNT* isoform expression in HBECs and HASMs using RNA-seq

RNA-seq was used to define relative abundance of *NPNT* isoforms in cultured HBECs and HASMs from 8 and 5 donors, respectively. The RNA-seq data was analysed by Cuffnorm which gives details on isoform expression. The results confirmed the presence and abundance of *NPNT* transcripts in HBECs and HASMs. Moreover, it demonstrated the heterogeneity in expression among different individual donors. Transcript ENST00000379987 is the most abundant and had the highest transcript expression among 8 HBECs donors (Table **3**-**4**). In the HASMs donors, Transcript ENST00000504787 and ENST00000503451 are the top 2 highest transcripts expression among 5 HASMs donors. Furthermore, the donor D006 is the highest *NPNT* expression value among 5 donors (Table **3**-**5**).

Gene	Nearest ref ID	Gene	D1	D2	D3	D4	D7	D8	D9	D11	**Average**	**Protein Product**
NPNT	ENST00000504304	NPNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	No protein
NPNT	ENST00000503451	NPNT	0.17	0.17	0.09	0.00	0.62	0.31	0.06	0.00	0.18	Protein coding
NPNT	**ENST00000379987**	**NPNT**	**1.15**	**3.32**	**1.43**	**1.28**	**7.29**	**9.33**	**1.11**	**1.34**	**3.28**	**Protein coding**
NPNT	ENST00000453617	NPNT	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000427316	NPNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000514622	NPNT	0.00	0.00	0.00	0.06	0.81	0.00	0.00	0.00	0.11	Protein coding
NPNT	ENST00000305572	NPNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000506666	NPNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000503451	NPNT	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000514837	NPNT	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.00	0.09	Protein coding
NPNT	ENST00000513430	NPNT	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.01	No protein
NPNT	ENST00000506056	NPNT	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.01	No protein
NPNT	ENST00000511518	NPNT	0.18	0.37	0.36	0.00	0.46	0.69	0.00	0.41	0.31	No protein
NPNT	ENST00000505821	NPNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	No protein
NPNT	ENST00000514632	NPNT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	No protein
NPNT	ENST00000504787	NPNT	0.00	0.59	0.18	0.28	1.44	1.92	0.08	0.11	0.58	No protein
NPNT	ENST00000505917	NPNT	0.15	0.00	0.00	0.00	0.00	0.20	0.19	0.17	0.09	No protein
	Total		1.65	4.49	2.19	1.62	10.73	12.45	2.12	2.02		

Table 3-4 NPNT transcript in 8 HBECs donors. Transcript ENST00000379987 demonstrated the highest transcript expression among 8 HBECs donors.

Gene	Nearest ref ID	Class code	D006	D007	D0023	D0063	D00529	average	Protein product
NPNT	ENST00000504304	=	0.00	0.00	0.00	0.00	0.00	0.00	No protein
NPNT	**ENST00000503451**	**j**	**1.10**	**0.20**	**1.50**	**0.40**	**0.05**	**0.65**	Protein coding
NPNT	ENST00000503451	j	0.00	0.00	0.12	0.00	0.00	0.02	Protein coding
NPNT	ENST00000503451	j	0.00	0.00	0.00	0.00	0.03	0.01	Protein coding
NPNT	ENST00000503451	j	0.00	0.00	0.01	0.00	0.00	0.00	Protein coding
NPNT	ENST00000453617	=	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000379987	=	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000453617	=	0.00	0.00	0.00	0.11	0.00	0.02	Protein coding
NPNT	ENST00000427316	=	0.00	0.00	0.00	0.00	0.02	0.00	Protein coding
NPNT	ENST00000514622	=	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000305572	=	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000506666	=	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000503451	=	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000514837	=	0.00	0.00	0.00	0.13	0.00	0.03	Protein coding
NPNT	ENST00000513430	=	0.00	0.00	0.00	0.00	0.00	0.00	No protein
NPNT	ENST00000503451	j	0.00	0.00	0.00	0.00	0.00	0.00	Protein coding
NPNT	ENST00000506056	=	0.00	0.00	0.00	0.00	0.00	0.00	No protein
NPNT	ENST00000511518	=	0.00	0.00	0.00	0.00	0.00	0.00	No protein
NPNT	ENST00000505821	=	0.00	0.00	0.00	0.08	0.00	0.02	No protein
NPNT	ENST00000514632	=	0.00	0.00	0.00	0.00	0.00	0.00	No protein
NPNT	**ENST00000504787**	**=**	**2.69**	**0.00**	**1.60**	**0.00**	**0.06**	**0.87**	No protein
NPNT	ENST00000505917	=	0.00	0.00	0.00	0.00	0.00	0.00	No protein
	Total		3.79	0.20	3.23	0.72	0.17		

Table 3-5 NPNT transcript in 5 HASMs donors. Transcript ENST00000504787 and ENST00000503451 are the top 2 highest transcripts expression among 5 HASMs donors.

3.4.8 *NPNT* mRNA expression in COPD lung tissue using GEO dataset

To identify the expression level of *NPNT* mRNA in COPD lung tissues I used publicly available expression microarray data to assess *NPNT* mRNA expression in lung tissues taken from 23 COPD patients with a history of smoking and nine individuals with a history of smoking but no evidence of obstructive lung disease. *NPNT* mRNA *was* decreased in individuals with COPD who smoked compared with smokers who had no history of airway obstruction (adjusted p value <0.001 and logFC = -0.97, Figure 3-14.

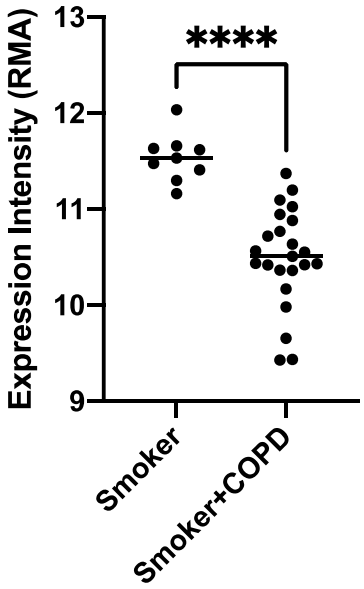


Figure 3-14 *NPNT* expression intensity in human lung tissues. Comparison the lung tissues NPNT expression in smoker without obstructive evidence (n=9) to smoker with COPD (n=26) which confirmed by spirometer. ****<0.001

3.4.9 NPNT protein expression in human lung tissues

3.4.9.1 NPNT protein expression in foetal lung tissues using immunohistochemistry

During the different stages of foetal development, NPNT protein expression was next investigated by immunohistochemistry. In the earliest stage, as shown in Figure **3**-**15**, NPNT could be observed weakly in embryonic samples (16 days), while protein showed higher expression around outer areas in developing bronchioles at 23 days development. NPNT was also detected in the outer areas at the pseudoglandular stage (10 weeks) (Figure **3**-**16**).

As shown in Figure 3-17, strong NPNT protein expression was detected throughout the canalicular stage (17 and 20 weeks), where the protein was localised to the nucleus and cytoplasm in most airway cell types. Taken together, NPNT protein is differentially expressed with increasing expression across foetal lung developmental stages.

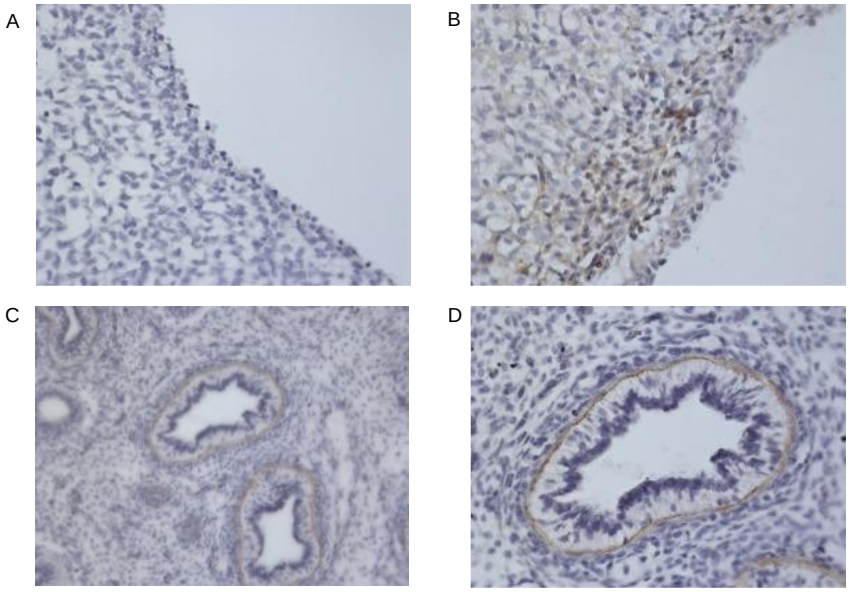


Figure 3-15 Images of NPNT protein expression in tissue from human foetuses at a range of development stages: embryonic. Representative images from foetal lung age at 16 days (A and B, ×10 magnification) and 23 days (C ×20 and D ×40 magnification) from a single donor, respectively.

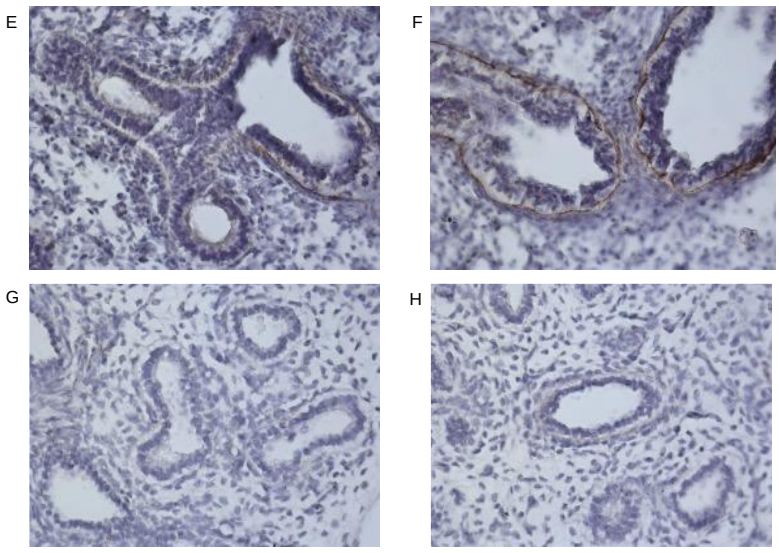


Figure 3-16 Images of NPNT protein expression in tissue from human foetuses at a range of development stages. Representative images from foetal lung age at 10 weeks (E and F, ×40 magnification) and 14 weeks (G and H, ×20 magnification) from a single donor, respectively.

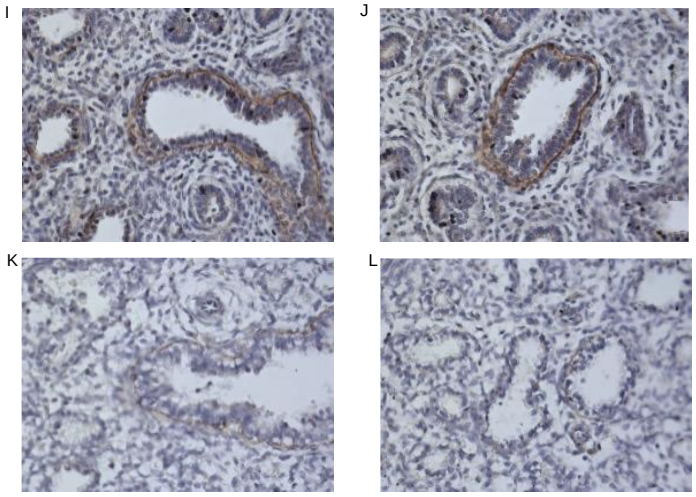


Figure 3-17 Images of NPNT protein expression in tissue from human foetuses at a range of development stages: canalicular (I and J. 17 weeks; K and L. 20 weeks). Representative images from foetal lung age at 17 weeks (I and J, ×40 magnification) and 20 weeks (K and L, ×40 magnification) from a single donor, respectively.

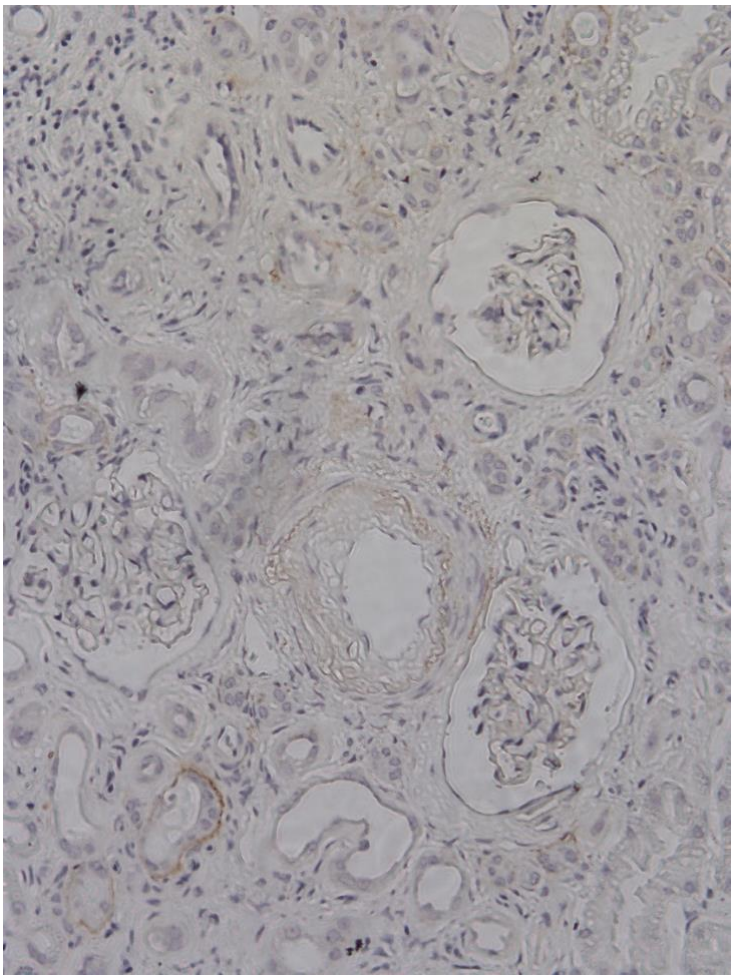


Figure 3-18 Adult kidney tissue was used as a positive control sample. ×40 magnification.

3.4.9.2 NPNT protein expression in COPD lung tissues using Immunohistochemistry

In order to investigate the expression of NPNT protein in the human lung tissue immunohistochemistry was carried out on human lung tissues. The NPNT antibody final concentration of 1:30 was used for this work due to overall weak staining.

The expression of NPNT protein in lung samples from healthy (control) donors (Figure **3**-**19** and Figure **3**-**20**) and COPD patients (Figure **3**-**21** and Figure **3**-**22**) was detected by immunohistochemistry. NPNT is expressed in the alveolar and fibro-elastic layers, especially in pneumocytes, fibroblasts from healthy and COPD donors. However, NPNT protein expression was higher in pneumocytes and fibroblasts than epithelial cells both in healthy and COPD lung tissue (Figure **3**-**23**). In all cell types, NPNT is predominantly expressed in nucleus and cytoplasm.

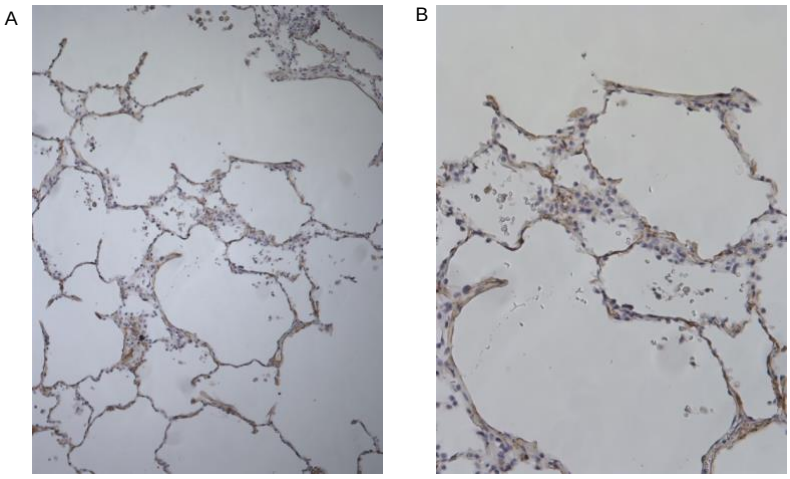


Figure 3-19 NPNT protein expression detected by immunohistochemistry in human lung tissue from one healthy control lung tissue (representative images from n=6 donors). NPNT expressed in alveolar, especially in pneumocytes, interstitial connective tissue of the lung (images A to B). ×40 magnification.

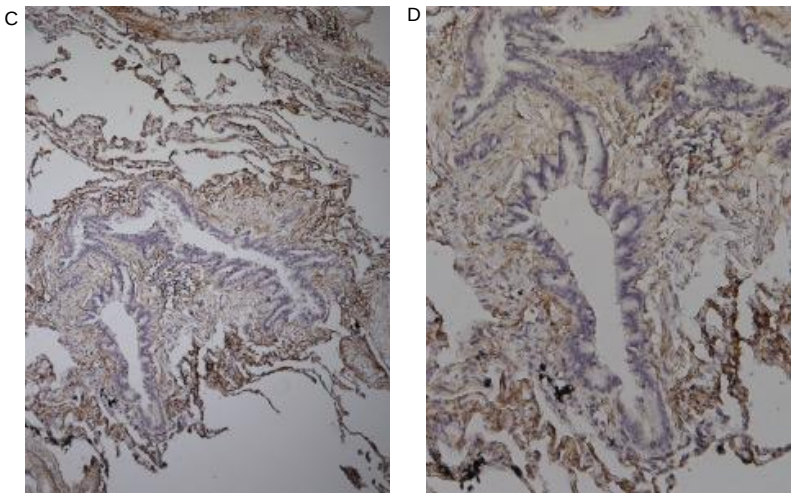


Figure 3-20 NPNT protein expression detected by immunohistochemistry in human lung tissue from one healthy control lung tissue (representative images from n=6 donors). NPNT lowly expressed in the epithelium of the mucosa, and highly detected in fibro-elastic layer, alveolar, interstitial connective tissue (images C to D). ×40 magnification.

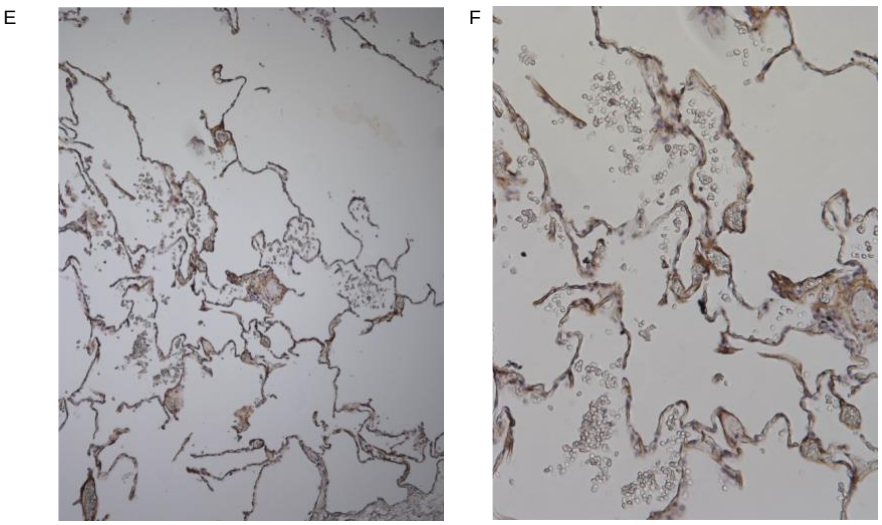


Figure 3-21 NPNT protein expression detected by immunohistochemistry in human lung tissue from one COPD patient's lung tissue (representative images from n=6 donors). NPNT expressed in alveolar, especially in pneumocytes, interstitial connective tissue of the lung (images E to F). ×40 magnification.

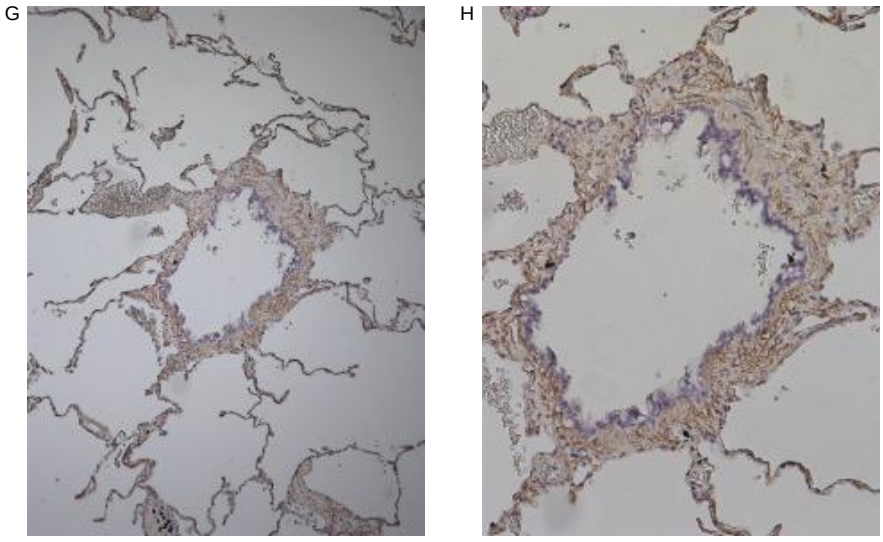


Figure 3-22 NPNT protein expression detected by immunohistochemistry in human lung tissue from one COPD patient's lung tissue (representative images from n=6 donors). NPNT lowly expressed in the epithelium of the mucosa, and highly detected in fibro-elastic layer, alveolar, interstitial connective tissue (images G, H).

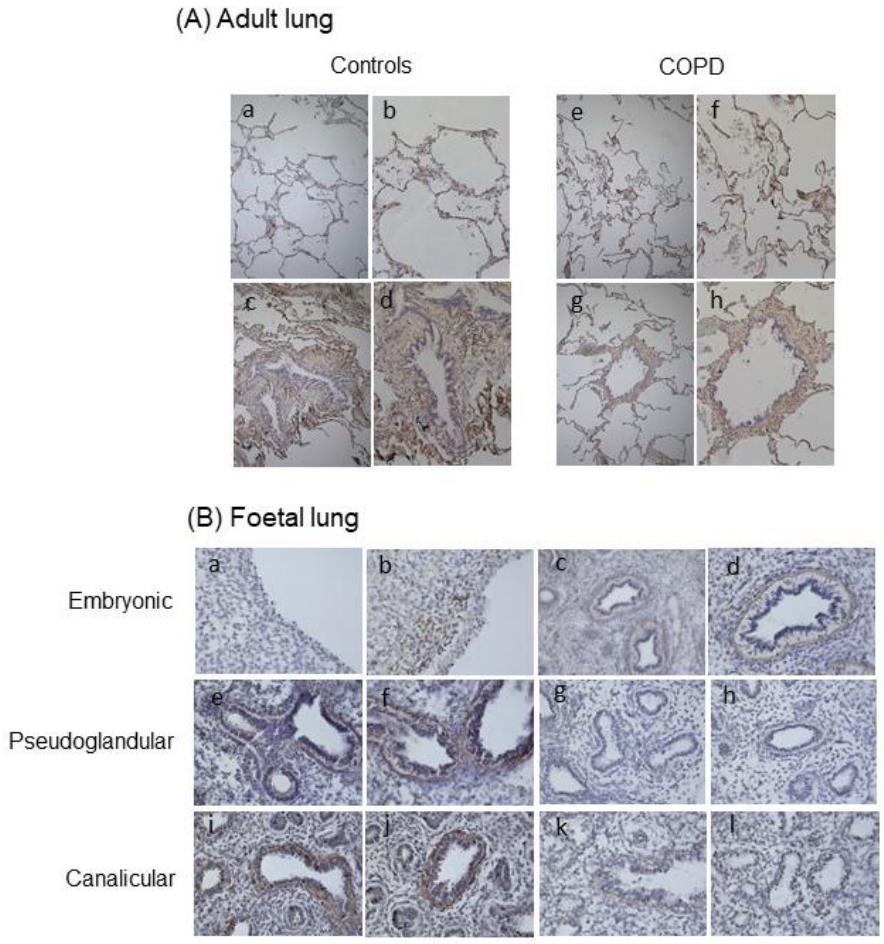


Figure 3-23. NPNT protein expression in human lung tissue. NPNT protein expression detected by immunohistochemistry in human lung tissue from controls, COPD and multi-stages of foetal tissue. (A) Images of NPNT expression in lung tissue from healthy donors (images a to d) and individuals with COPD (images e to h). (B) Images of NPNT protein expression in tissue from human foetuses at a range of development stages: embryonic (a and b.16 days, c and d. 23 days), pseudoglandular (e and f. 10 weeks; g and h. 14 weeks), and canalicular (i and j. 17 weeks; k and l. 20 weeks). NPNT expression in the human lung tissue was described above.

3.5 Discussion

In this chapter, the aim was to identify potentially causal variants in the 4q24 locus associated with lung function measures and to investigate the wider relevance by confirming the effect of variants on eQTL effects and association with phenotype by PheWAS. I also aimed to investigate the expression profiles of NPNT mRNA or protein in foetal and adult lung tissue and in airway structural cells by qRT-PCR, RNA sequencing and immunohistochemistry.

Both SNP rs17331332 and rs34712979 in the locus 4q24 are significantly associated with lung function and risk of COPD (10, 80). I explored SNP rs17331332 near *NPNT* for functional validation (10) and I found that the eQTL covered *GSTCD*, *INTS12*, and *NPNT*, and this SNP was associated with lung, eosphagus gastroesophageal junction, and esophagus muscularis expression of NPNT. On the other hand, the SNP rs34712979 located in the *NPNT* locus was the most significantly genetic variant which associated with FEV_1/FVC, although LD also spans *INTS12* and *GSTCD* **(80)**. From the analysis of PheWAS in this chapter, SNP rs34712979 was associated with COPD, asthma, arm fat percentage, as well as weight. Henry and colleagues have reported the expression and functions of *GSTCD* in airway relevant cells (165), and *GSTCD* has potential to affect airway response via inflammatory signals in *GSTCD* knockout mouse model (166). In addition, Kheirallah et al showed that *INTS12* can regulate translation through altered activity in protein synthesis pathways (130). However, the most current

association data, taken together with the eQTL and pQTL data, strongly suggest NPNT is a causal gene at this locus. Therefore, given the strong evidence suggesting a functional role for *NPNT* in the airways, I aimed to investigate the expression and function of *NPNT* in airway structural cells and lung.

NPNT was highly expressed in lung of foetal and adult. Gene expression was identified in lung tissues and relevant airway structural cells particularly in HBECs in public database by both RNA-seq and TaqMan PCR. Both mRNA and protein expression were observed using RNA-seq and antibody-based detection in Human Protein Atlas. In all I included 27 tissues or organs, and found higher levels expression of *NPNT* mRNA and protein in lung tissue than in most other tissues. Nearly 80% of the human protein-coding genes were integrated with RNA in the Human Protein Atlas study by RNA-seq and antibody-based detection (162). However, the value of RNA-seq of *NPNT* was the highest in thyroid of 27 tissues, whilst interestingly the NPNT protein could not be detected in thyroid.. In the lung, I found that NPNT protein expression was generally lower outside the airways.

The studies described in this chapter suggest that NPNT is implicated in foetal lung development. I found that *NPNT* mRNA expression intensity increased steadily with increasing foetal lung development across the Pseudoglandular and into the Canalicular stages of growth. NPNT protein was expressed during the these stages of foetal lung development. Furthermore, Hancock suggested that the *NPNT* SNP

rs10516529 is located in a binding site for the transcription factor *POU6F1* (also known as *mPOU* homeobox protein), which is predicted to play a key role in the developing lung (10, 167). In transgenic mice, Linton et al reported that they found that the first exon of *Npnt* was be excised mice lacked kidneys at birth (103). They suggested that *Npnt* plays a crucial role of renal agenesis at birth and is implicated in kidney development (103). Furthermore, Nephronectin as novel ECM protein could promote mitotic rounding and cytokinesis of postnatal cardiomyocytes in vitro (116). However, these studies did not assess lung development.

The data presented in this chapter also suggest that NPNT expression is regulated in COPD. NPNT was highly expressed in lung tissue confirmed by both RNA-seq and immunohistochemistry. However, the *NPNT* mRNA levels were significantly decreased in COPD lungs compared to smokers with no evidence of airway obstruction. NPNT protein was detected by immunohistochemistry in both COPD and control healthy lungs. The NPNT protein was detected in pneumonocytes, fibroblasts, and epithelial cells, particularly the sub-epithelial fibroblasts which surround airway. However, I could not compare the expression between COPD and healthy controls because immunohistochemistry based on enzymatic amplification technology is not sufficiently quantitative.

The lung extracellular matrix (ECM) forms the main structure of the lung and plays a key role in the COPD airway remodelling which results in

disturbance of the ECM (168). Kon and colleagues reported that anti-nephronectin antibody could ameliorate anti-type II collagen-induced arthritis (169), which suggest the interaction between nephronectin and type II collagen. The collagens increased deposition in the lung and implicated in lung remodelling. Collagen IV and laminin are the main components of the basement membrane. The main component of lamina propria and interstitium are fibrillar collagens, elastin, fibronectin, glycosaminoglycans and proteoglycans (170, 171). Furthermore, the cells that produce of ECM are mainly fibroblasts and activated myofibroblasts, with some contribution from airway smooth muscle and epithelial cells (172). The results of the lung tissue IHC studies showed that the nephronectin protein was expressed higher in the fibro-elastic layer than epithelial cells.

On the other hand, MMPs are elevated in COPD and are able to cleave all ECM components (173). MMP-1, MMP-7, MMP-9 are highly expressed in bronchiolar and alveolar epithelial cells of COPD (171, 174, 175). Moreover, growth factors, ECM-modifying enzymes, or other ECM-association protein are associated with ECM and can be stored in ECM as secreted proteins, which do not affect ECM structure but can still function. The ECM in COPD can be destroyed by ECM-degrading enzymes such as elastase released by inflammatory cells. Recent studies demonstrated that NPNT was downregulated by TGF-β1 in osteoblastic cell line (108) and Gao et al reported that the serum TGF-β1 level of acute exacerbation COPD was significantly elevated when compared to the stable stage of disease (176). These studies

potentially suggest the secreted protein TGF-β1 in ECM could down-regulate the expression of nephronectin in COPD. Thus, interactions between the ECM and NPNT in lung homeostasis deserve further research.

In summary, the data shown in this chapter implicate genetic variants at the NPNT locus in COPD risk and control of lung function. In the next chapter I therefore set out to develop approaches to explore the potential functional role for NPNT in the lung.

4. Gene expression changes measured by Microarray in human bronchial epithelial cells engineered with NPNT knockdown

4.1 Introduction

The extracellular matrix (ECM) provides structural support and stability to the lung (173). The ECM plays an important role in pathological alterations and functional changes in chronic lung diseases, such as COPD and asthma (177). Nephronectin (NPNT) is a novel ECM first described in kidney and highly expressed in lung (178), which potentially plays an important role in the development of the human lung as investigated in Chapter 3. Thus, in this chapter, I aim to assess the potential biological functions and pathways of NPNT using transcriptomics in human bronchial epithelial cells (HBECs).

The ECM regulates cell adhesion, differentiation, and proliferation and NPNT has been reported to increase cell proliferation and mediate several inflammatory pathways in the kidney, in addition to bone homeostasis (106, 179-181). Many studies report that ECM protein NPNT modulates differentiation and proliferation in the osteoblastic cell line (123) and is linked to the regenerating process of renal tubular epithelial cell (95). Furthermore, NPNT KO (*Npnt*Δ^{ex1}) mouse model displayed kidney defects and it has been demonstrated that ECM protein NPNT promotes kidney development and stimulates glial cell line-derived neurotrophic factor (Gdnf) expression via binding integrin α8β1 (NPNT-$\alpha 8\beta 1$-Gdnf pathway) (103).

In the chapter 3, I have shown that NPNT mRNA expression is highest in differentiated epithelial cells among different types of human airway cells, including cultured HBECs, BEAS-2B, and human airway smooth

muscle cells (HASMs). Furthermore, NPNT expression is decreased in COPD lung tissue compared to non-COPD control. However, the downstream genes and pathways involved in driving functional effects regulated by NPNT are unknown and require further study, particularly in HBECs.

To evaluate the functions of NPNT, I initially applied double-stranded RNA-mediated interference (RNAi) to knockdown NPNT expression in HBECs. RNAi is a rapid and widely used method of silencing gene expression in cell model research (Figure **4**-**1**) (182). RNAi has previously been used to knockdown genes which have been identified as susceptibility loci for COPD by GWAS (88, 161). To assess the consequences of NPNT knockdown, I initially used microarrays (183). On a Microarray, tens of thousands of complementary DNA (cDNA) fragments or oligonucleotides (probes) hybridize with the specific RNA which has been labelled. Thus, the Microarray can be used for profiling expression levels of known genes.

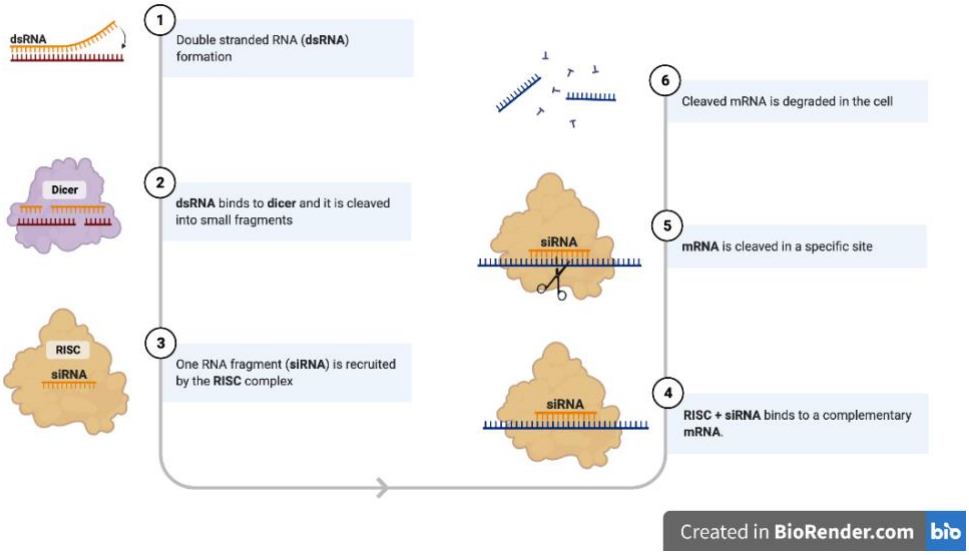


Figure 4-1 Gene silencing by double-stranded RNA. Firstly, dsRNA is cleaved by the Dicer enzyme to produce siRNAs as 20-25nt fragments. The siRNAs bind to the RNA-induced silencing complex (RISC). A helicase present in complex activates RISC by unwinding the siRNAs. The antisense component of siRNA in the RISC guides the complex towards a complementary mRNA, resulting in endonucleolytic cleavage of the mRNA in a specific site and then the cleaved mRNA is degraded in the cell. The figure created in BioRender.com (182).

4.2 Aims

In this chapter, the aims were to investigate the differentially expressed genes and pathways following *NPNT* knockdown by siRNA in HBECs using Microarray to provide insight into the potential biological role of *NPNT* in lung structural cells.

4.3 Methods

4.3.1 Cell Biology Methods

4.3.1.1 Cell culture of human bronchial epithelial cells

Passage 3 of Donors D195307 HBECs were cultured in T75 flask until 70%-80% confluent as detailed in Chapter 2. Prior to siRNA transfected HBECs were seeded to 6 well plates at a density of 130, 000 cells/well.

4.3.1.2 NPNT gene knock-down using siRNA

4.3.1.2.1 *NPNT* siRNA Preparation

The *NPNT* siRNA was designed as descripted in Chapter 2. siRNAs were stored at a stock concentration of 20µM at -80℃. Each siRNA (siRNA a, siRNA b, siRNA c) was diluted 1:100 using 0.5µl of each stock in a 50µl final solution. And the INTERFERin® (catalogue number: 409-10, VWR international, UK) was purchased, which is a powerful siRNA reagent which provides more than 90% transfection efficiency at 1nM siRNA in a wide range of cells.

4.3.1.2.2 NPNT siRNA Treatment

Passage 3 HBECs (repeated 3 times in 2 donors,) were cultured in 6-well plates, incubated for 24h or until cells were approximately 30%-50% confluent and ready to be knockdown by siRNA. Cells were transfected with 1nM of siRNA a, b and c, or 1nM of scrambled control and one untransfected condition. The transfection mix was prepared as follows: 120,000 cells, 21µl 0.2µM siRNA, 4.2µl INTERFERin® and 394.8µl PneumaCult-Ex Medium or Bronchial Epithelial Cell Growth

Medium (BEGM). Each transfection mix was vortexed and incubated for 15mins at room temperature. Each condition has 2 wells with 200 µl of transfection mix added to each well. After 24h, 48h or 72h incubation, the media was aspirated and discarded. The cells were collected as descripted in Chapter 2 and the cell lysates were stored at -80℃ until further use.

4.3.2 Western blot

The Western Blot was descripted in Chapter 2. Briefly, total cell protein was collected from HBECs, SDS-PAGE gel was prepared, proteins were transferred to PVDF membrane, and then separated by SDS-PAGE. The primary antibody of NPNT was diluted to 1:250 and incubated overnight. Finally, the protein was detected by the ECL method.

4.3.3 Molecular Biology Methods

4.3.3.1 RNA extraction

RNA was extracted from cell lysates as detailed in the general material and methods chapter 2 using the Sigma GenElute™ Mammalian Total RNA Miniprep Kit (RTN70- 1KT, Sigma Aldrich), cDNA synthesised by DNase I, Amplification Grade (18068015, Thermo Fisher Scientific). The RNA samples were stored at -80℃ freezer.

4.3.3.2 RNA quality control

RNA was detected for quality and quantity with a Nanodrop ND-2000 full spectrum spectrophotometer and RNA integrity number (RIN)

scores were measured by Agilent 2100 Bioanalyser and Agilent 4200 TapeStation as explained in Chapter 2. RIN>8 were considered to be of sufficient quality for Microarray.

4.3.3.3 TaqMan PCR

TaqMan PCR has been describted in chapter 2. Briefly, the RNA was extracted from HBECs by the GenElute™ Mammalian Total RNA Miniprep Kit (RTN70- 1KT, Sigma Aldrich), cDNA synthesised by DNase I, Amplification Grade (18068015, Thermo Fisher Scientific) and Superscript First Strand Synthesis System for RT-PCT kit (11904018, Invitrogen), and qRT-PCR was performed as previously described for NPNT by NPNT PDAR (Hs01568131_m1, Thermo Fisher Scientific).

4.3.3.4 Microarray

The method of Microarray was explained in Chapter 2. In brief, RNA was extracted and evaluated for RIN. Microarrays were run using Affymetrix GeneArray Scanner 3000 7G Plus (Affymetrix, USA) in Sutton Bonington, University of Nottingham. The human gene 2.1 array (Catalog number: 902114, Thermo Fisher Scientific, USA) is a whole-transcript array which includes probes to measure both mRNA and long non-coding RNA transcripts (lncRNA). R and Bioconductor were used for annotation, analysis and visualization of data (141, 183).

4.3.3.4.1 Gene Set Enrichment Analysis

As described previously in Chapter 2. GSEA was used to identify pathways when comparing between scrambled and siRNA treatment.

The curated dataset and Hallmark dataset were applied to generate gene sets which were regulated by involved in NPNT knockdown steam. The commands can be found in chapter 8 appendix.

4.3.3.4.2 Graphical visualisation of Microarray results

Volcano plot and heatmaps were used to represent differentially expressed genes, which were generated from R studio. From the R Bioconductor, the ***ggplots***, ***ggrepel***, ***limma***, ***edgeR***, *and* ***Glimma*** packages were installed and used the commands which can be found in chapter 9 appendix (184-186).

4.4 Results

4.4.1 Identifying the optimal concentration of *NPNT* siRNA in human bronchial epithelial cells model

4.4.1.1 qRT-PCR confirms NPNT knockdown efficiency using siRNAs

In order to determine the function of NPNT in the HBECs I engineered NPNT knockdown using anti-NPNT siRNAs. siRNA mediated knockdown was performed at three different concentrations (0.1nM, 1nM and 10nM), and using three different siRNAs (a, b, c and a+b+c), incubated for 24h, 48h and 72h. qRT-PCR was used to measure *NPNT* mRNA expression in controls and siRNA knockdown HBECs. The knockdown efficiency was higher in cells transfected with 1nM than 0.1nM or 10nM siRNA while at 48h (Figure **4**-**2**), *NPNT* mRNA expression in 48h was lower than at 24h or 72h after knockdown. *NPNT* mRNA was successfully reduced in HBECs after transfection with the siRNA concentration of 1nM siRNA a (84.0±5.7%, n=3, $p<0.05$) and 1nM siRNA b (81.0±2.0%, n=3, $p<0.05$) when compared to untransfected at 48h. There was extensive cell death seen by microscopy when the combined 1nM siRNA abc was used. Therefore, the optimal conditions for *NPNT* knockdown was therefore chosen as use of 1nM siRNA a or 1nM siRNA b at 48h.

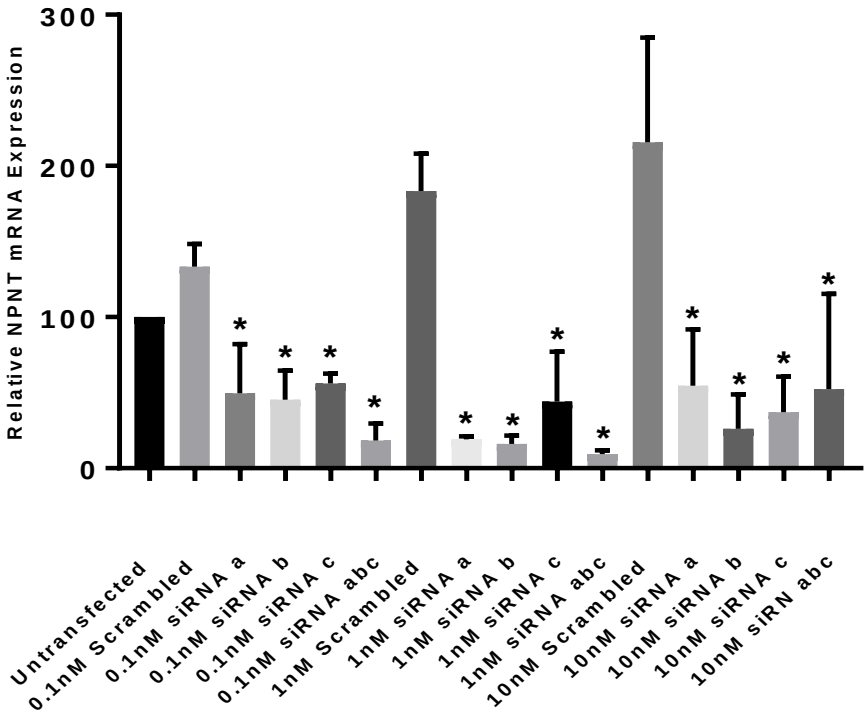


Figure 4-2. Expression of *NPNT* mRNA in HBECs following *NPNT* siRNAs mediated knock-down at 48h. The knockdown efficiency was confirmed by Taqman PCR of *NPNT* mRNA from untransfected, scrambled, and different siRNAs and concentrations (compared NPNT levels to scramble, * $p < 0.05$, n=3).

4.4.1.2 Western Blot confirms NPNT protein knockdown

To further investigate the knockdown efficiency, I detected NPNT protein levels in HBECs by Western blotting. The results of Western blot are consistent with mRNA knockdown results achieved in the TaqMan PCR experiments and identify that the NPNT specific RNAi down-regulate NPNT protein levels in HBECs (Figure **4-3**). The NPNT protein was normalised to β-actin and I found that siRNA b (0.02±0.01)

significantly downregulated NPNT protein expression in HBECs when compared to scrambled control (0.41±0.16) ($p<0.05$) (fig 4.3).

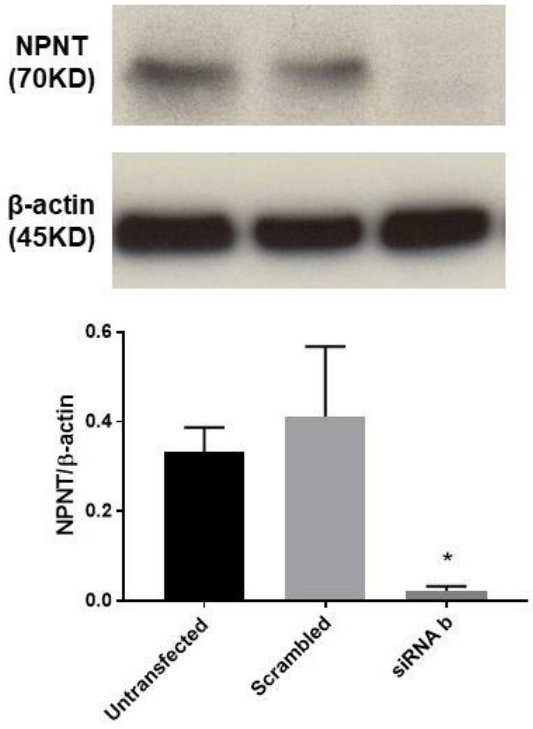


Figure 4-3. Representative example of down-regulation of NPNT protein levels by *NPNT* siRNA b. Western blotting was performed and normalised to $\boldsymbol{\beta}$-actin to analyse the levels of NPNT protein in HBECs (see lower panel, compared to scrambled *$p < 0.05$, $n=3$).

4.4.2 Quality control of RNA samples for Microarray

RNA samples were extracted and the quality control information for Microarray is shown in Table 4-1. The table shows treatment conditions, the RNA concentration, wavelength ratio (260/280 ratio) detected by Nanodrop, and RIN scores determined by Agilent 2100 Bioanalyser and Agilent 4200 TapeStation as detailed in chapter 2. All samples (RIN>8) were considered to be of sufficient quality for Microarray by Sutton Bonington Campus of University of Nottingham.

Sample No.	Treatment (48h)	Con.ng/µl	260/280	RIN
E3.1	Scrambled 1nM	166.4	2.06	10
E3.2	Scrambled 1nM	307.2	2.04	10
E3.3	Scrambled 1nM	382.3	2.03	10
E3.1	siRNA A 1nM	122.2	2.02	10
E3.2	siRNA A 1nM	310.7	2.04	10
E3.3	siRNA A 1nM	373.6	2.04	10
E3.1	siRNA B 1nM	55.9	2.04	10
E3.2	siRNA B 1nM	345.6	2.04	10
E3.3	siRNA B 1nM	383.2	2.03	10
E3.1	untransfected	315.4	2.03	10
E3.2	untransfected	391.6	2.02	10
E3.3	untransfected	333.6	2.02	10

Table 4-1 Information of RNA samples quality control for Microarray using siRNA to knockdown *NPNT* in HBECs. Donor: 195307, passage 3.

4.4.3 Microarray data analysis using siRNA to knockdown *NPNT* in HBECs

4.4.3.1 Microarray Data Quality Control Analysis

All the images of quality control (QC), including Boxplots of raw data and normalized data, cluster dendrogram, histogram and density, were generated by ***ggplot2*** package in R studio environment (185). Box plots were applied to examine the distribution of feature intensities. Comparing box plots of normalised data to raw data, the median absolute value for each sample were similar (Figure 4-4). The cluster dendrogram shows the illustration the hierarchical organisation of samples (Figure **4**-**5**). The density plot shows the comparison of probe intensity behaviour between different arrays from Affymetrix. Quality control of histogram and density were showed in Figure **4**-**6**, the shape and centre of the distribution among different arrays were similar, which means the normalization result was suitable for analysis. As shown in Figure **4**-**7**, the Principle component analysis (PCA) identified some level of clustering, however siRNA a and siRNA b did not cluster together, potentially suggesting these siRNAs have different effects on gene expression in contrast to that expected.

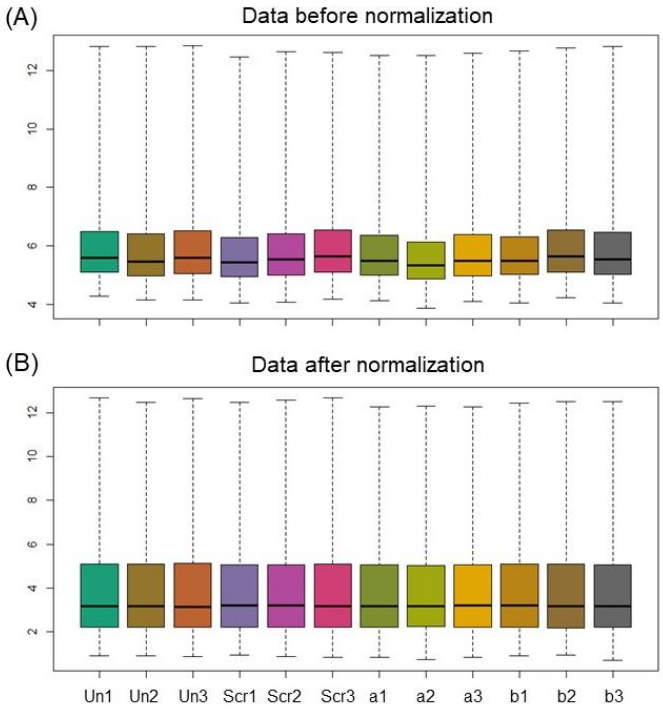


Figure 4-4 Illustration of the Boxplots from raw data of microarray and after normalization. Boxes indicate the interquartile range, a median in the middle of the box as the line, 25th percentile at the bottom, and 75th percentile at the top. (A) Data before normalization. (B) Data after normalization. Un, untransfected; Scr, scrambled; a, siRNA a; b, siRNA b.

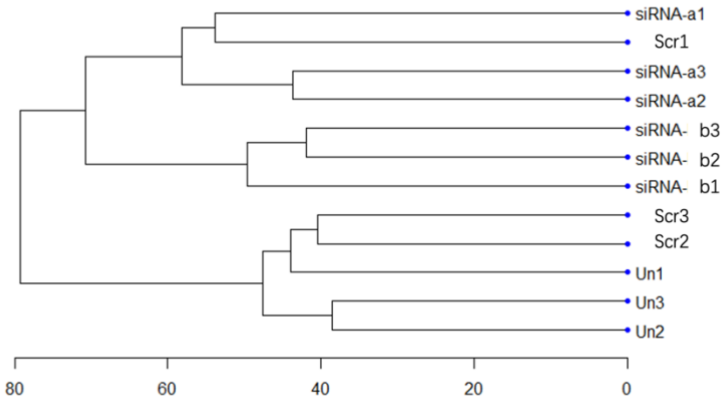


Figure 4-5. Cluster Dendrogram, which shows the illustration the hierarchical organisation of samples. The Scr1 was clustering with siRNA samples, but not excluded when analysed the data. The x axis represents the distance or dissimilarity between clusters. The y axis shows sample clustering.

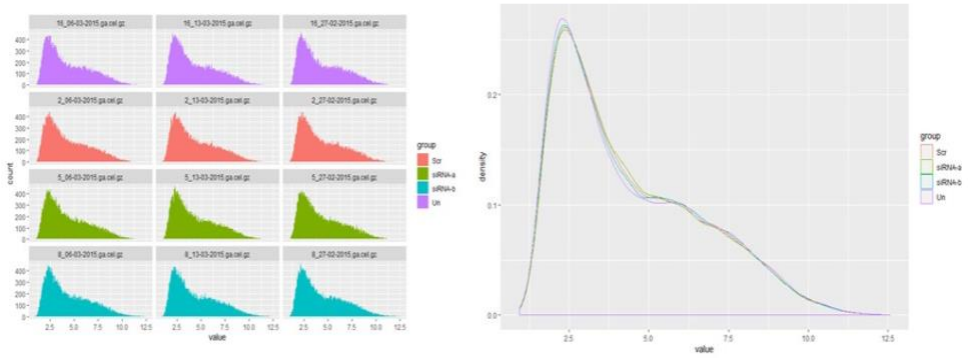


Figure 4-6. Quality control of histogram and density. (A) Histogram of each sample. (B) Density plot of the log-intensities of Affymetrix arrays.

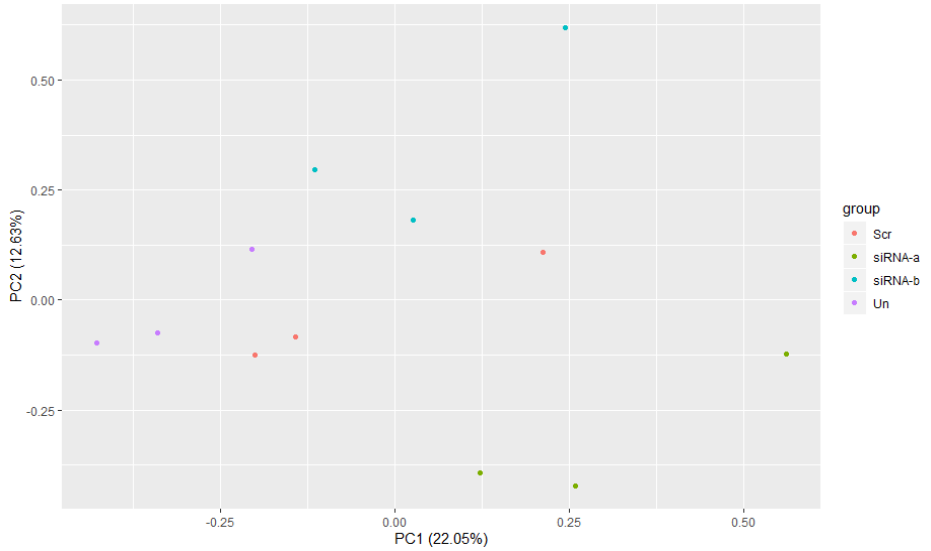


Figure 4-7. PCA is a statistical technique applied to identify the underlying variability in a dataset. The samples from the four groups are shown in different colours.

4.4.3.2 Gene expression profiles in Microarray

4.4.3.2.1 Differential gene expression analysis with siRNA a

The *limma* package was performed to generate differentially expressed genes (DEGs) from untransfected, scrambled, siRNAs group (140). Among the 26,376 genes identified from the gene expression profile microarray analysis, 0 DEGs were identified between scrambled and untransfected group (Fold change ≥1.5 or ≤ -1.5 and *adj.P.Val* < 0.05), 182 DEGs were identified between siRNA a and scrambled. The volcano plot was constructed by plotting the DEGs between the adjusted *P* value (y axis) and fold change (x axis) (Figure **4-8**). Among these 182 genes, the top 20 genes were listed in Table **4-2**, which were in order of adjusted p value as well as the fold change.

Among these top 20 genes are CXCL14 (C-X-C motif chemokine ligand 14, FC 4.5), OLFML2A (olfactomedin like 2A, Fold Change 3.4) and TMPRSS11 (transmembrane serine protease 13, Fold Change 3.01) which were the most upregulated genes. ADGRL4 (adhesion G protein-coupled receptor L4, Fold Change -2.55), VAMP7 (vesicle associated membrane protein 7, Fold Change -2.33) and CDK6 (cyclin dependent kinase 6, Fold Change -2.17) were the most downregulated genes based on adjusted p value and fold change.

Figure 4-8. Volcano plot of NPNT specific siRNA a compared to scrambled. The top 10 genes both up- and down-regulated based on adjusted p value are labelled. Log2 fold change is x-axis, and the –log10 adjusted p-value as y-axis.

Gene	Gene Name	Fold Change	Up/Down regulation	adj.P.Val
MAF	MAF bZIP transcription factor	2.78	Up	0.00028
IMPA2	inositol monophosphatase 2	2.38	Up	0.00028
IL6R	interleukin 6 receptor	2.37	Up	0.00028
ENDOD1	endonuclease domain containing 1	2.29	Up	0.00028
OLFML2A	olfactomedin like 2A	3.40	Up	0.00069
TMPRSS11F	transmembrane serine protease 11F	3.01	Up	0.00069
VANGL2	VANGL planar cell polarity protein 2	2.14	Up	0.00117
FRRS1	ferric chelate reductase 1	1.91	Up	0.00164
BLNK	B cell linker	2.05	Up	0.00175
MXRA5	matrix remodeling associated 5	2.97	Up	0.00198
VAMP7	vesicle associated membrane protein 7	-2.33	Down	0.00032
CDK6	cyclin dependent kinase 6	-2.17	Down	0.00117
ADGRL4	adhesion G protein-coupled receptor L4	-2.55	Down	0.00135
RAB28	RAB28, member RAS oncogene family	-2.11	Down	0.00135
GAS2L3	growth arrest specific 2 like 3	-1.87	Down	0.00147
OXR1	oxidation resistance 1	-1.89	Down	0.00381
ATP11C	ATPase phospholipid transporting 11C	-2.19	Down	0.00404
C20orf197	chromosome 20 open reading frame 197	-1.89	Down	0.00632
CASC4	cancer susceptibility 4	-1.62	Down	0.00657
GEMIN2	gem nuclear organelle associated protein 2	-2.18	Down	0.00706

Table 4-2. Top 20 DEG up- or down-regulation in HBECs. These genes were differentially expressed in HBECs from NPNT siRNA a (n=3) knock down compared to scrambled (n=3) group. Fold change ≥1.5. adj.P.Val (adjusted p value).

4.4.3.2.2 Differential gene expression analysis with siRNA b

I identified 413 DEGs by *limma* package in R environment (Fold change ≥1.5 or ≤ -1.5 and adj.*P*.Val < 0.05) between siRNA b and scrambled. Among these top 20 genes, IL33 (interleukin 33, FC 5.15), SYT11 (synaptotagmin 11, Fold Change 3.53), VCAN (versican, Fold Change 3.20) were the most upregulated genes and SOAT1 (sterol O-

acyltransferase 1, Fold Change -2.58), MAIP1 (matrix AAA peptidase interacting protein 1, Fold Change -2.53), GAS2L3 (growth arrest specific 2 like 3, Fold Change -2.46) were the most downregulated genes based on adjusted p value and fold change. The volcano plot was generated by plotting adjusted *P* value and fold change (Figure **4**-**9**). As showed in Table **4**-**3**, the top 20 genes were listed.

Figure 4-9. Volcano plot of anti-NPNT siRNA b compared to scrambled. The top 10 genes both up- and down-regulated based on adjusted *p* value and fold change are labelled. Log2 fold change is x-axis, and the –log10 adjusted *p* value as y-axis.

Gene symbol	Gene name	Fold Change	Up/Down regulation	adj.P.Val
SYT11	synaptotagmin 11	3.53	Up	4.09E-05
SMOC1	SPARC related modular calcium binding 1	3.11	Up	4.09E-05
RUNDC3A	RUN domain containing 3A	2.75	Up	8.22E-05
GAS2L3	growth arrest specific 2 like 3	-2.46	Down	8.22E-05
EHD3	EH domain containing 3	3.05	Up	0.000113
TDG	thymine DNA glycosylase	-2.15	Down	0.000113
RCAN3	RCAN family member 3	-2.2	Down	0.000113
SOAT1	sterol O-acyltransferase 1	-2.58	Down	0.000129
IL33	interleukin 33	5.15	Up	0.000186
VCAN	versican	3.2	Up	0.000186
CLIP3	CAP-Gly domain containing linker protein 3	2.83	Up	0.000186
CYBRD1	cytochrome b reductase 1	2.61	Up	0.000186
TOR1AIP2	torsin 1A interacting protein 2	-1.93	Down	0.000242
FAM171B	family with sequence similarity 171 member B	2.67	Up	0.000527
CBX1	chromobox 1	2.2	Up	0.000527
STK38	serine/threonine kinase 38	-1.85	Down	0.000527
RECQL	RecQ like helicase	-2.28	Down	0.000527
MAIP1	matrix AAA peptidase interacting protein 1	-2.53	Down	0.000527
MAGEF1	MAGE family member F1	-1.84	Down	0.000555
CCDC43	coiled-coil domain containing 43	-2.39	Down	0.000677

Table 4-3. Top 20 DEGs up- or down-regulation in HBECs selected by fold change ≥1.5 or ≤ -1.5 and adj.P.Val < 0.05. These genes were differentially expressed in HBECs from NPNT siRNA b (n=3) knock down compared to scrambled (n=3).

4.4.3.2.3 Differential gene expression in both siRNA a and siRNA b

Figure **4**-**10** shows the overlap between DEGs found with the NPNT siRNA a and siRNA b samples in Venn diagram. There were 24 genes replicated in the DEG analyses both in siRNA a and siRNA b groups which are listed in Table **4**-**4**. The DEGs protein-protein interaction was analysed using the STRING website (Figure **4**-**11**). However, three genes (ZBED6, CRABP2, DEPDC1) were expressed in the opposite direction from both groups. Therefore, there are 21 differentially expressed genes both in the siRNA a and siRNA b datasets. 15.15% of genes from the siRNA a occur in b and 6.17% of DEGs in siRNA b occur in a. Importantly, no significant reduction in NPNT expression (siRNA a, logFC=-0.391, adj.p.val=0.1136; siRNA b, logFC=-0.495, adj.p.val=0.0307) was noted in these analyses.

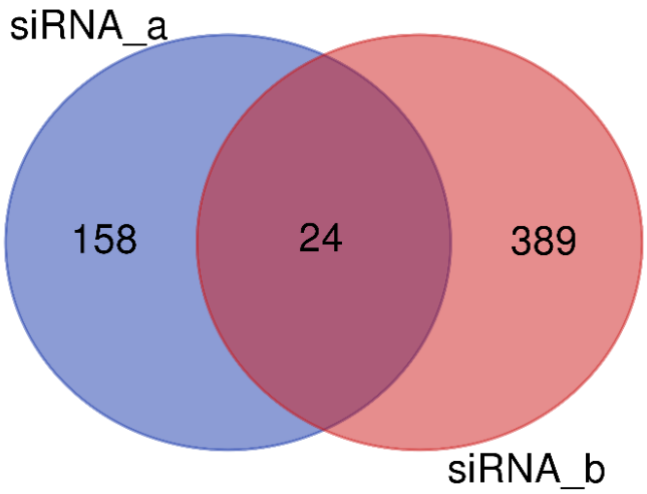


Figure 4-10. Venn diagram of DEGs in common between two groups. There are 24 DEGs in common between siRNA a and siRNA b, however, there is no differentially expressed genes between Scrambled and Untransfected group.

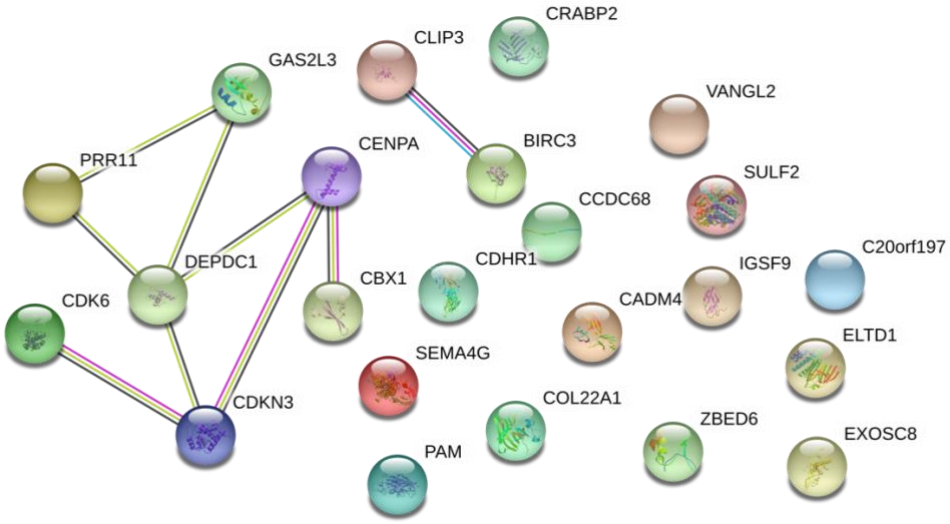


Figure 4-11 The protein-protein interaction network nodes was analysed by STRING (https://cn.string-db.org/). Edges represent protein-protein associations. The blue edge represents from curated databases. The pink edge represents experimentally determined.

Gene	Gene name	Location	Gene function/lung biology
GAS2L3	growth arrest specific 2 like 3	12q23.1	a locus at GAS2L3 for lung squamous cell carcinoma; actin cytoskeleton orgaization; can detect in Serum.
CLIP3	CAP-Gly domain containing linker protein 3	19q13.12	as a cytoplasmic linker protein, plays a role in glucose transport in adipocytes.
CBX1	Chromobox 1	17q21.32	Component of heterochromatin, the protein is enriched in the heterochromatin and associated with centromeres.
SULF2	sulfatase 2	20q13.12	Exhibits arylsulfatase activity and highly specific endoglucosamine-6-sulfatase activity; promotes the growth and metastasis of colorectal cancer by activating Akt and Erk1/2 pathways.
VANGL2	VANGL planar cell polarity protein 2	1q23.2	Involved in the control of early morphogenesis and patterning of both axial midline structures and the development of neural plate. Plays a role in the regulation of planar cell polarity.
ADGRL4	adhesion G protein-coupled receptor L4	1p31.1	Endothelial orphan receptor that acts as a key regulator of angiogenesis, a potential new biomarker for gliomas.
IGSF9	immunoglobulin superfamily member 9	1q23.2	Functions in dendrite outgrowth and synapse maturation.
CRABP2	cellular retinoic acid binding protein 2	1q23.1	Regulates the access of retinoic acid to the nuclear retinoic acid receptors, as a novel biomarker in plasma of non-small cell lung cancer.
SEMA4G	semaphorin 4G	10q24.31	May play a role in axon guidance.
EXOSC8	exosome component 8	13q13.3	involves in proper maturation of stable RNA species, involves in motor neuron and cerebellum development of zebrafish.
CENPA	centromere protein A	2p23.3	This gene encodes a centromere protein which contains a histone H3 related histone fold domain that is required for targeting to the centromere.
PRR11	proline rich 11	17q22	Proline rich 11 (PRR11) can promote cell proliferation, migration, and invasion; This novel gene regulated cell cycle, implicated in tumorigenesis via upregulating in lung cancer, and suggested as a potential therapeutic target in lung cancer.

CADM4	cell adhesion molecule 4	19q13.31	low CADM4 expression is associated with bone marrow metastasis in neuroblastoma.
CDK6	cyclin dependent kinase 6	7q21.2	Cyclin dependent kinase 6 (CDK6) protein plays an important role in cell cycle G1 phase progression and G1/S transition, and it also has been identified in many human cancers. Cyclin dependent kinase inhibitor 3 (CDKN3) as a cyclin-dependent kinase inhibitor, promotes cancer cell proliferation and tumorigenesis by targeting p27. This gene has been reported to be mutated or overexpressed in some cancers.
CDHR1	cadherin related family member 1	10q23.1	belongs to the cadherin superfamily of calcium-dependent cell adhesion molecules. May be required for the structural integrity of the outer segment (OS) of photoreceptor cells.
CDKN3	cyclin dependent kinase inhibitor 3	14q22.2	identified as a cyclin-dependent kinase inhibitor, prevent the activation of CDK2 kinase.
CCDC68	coiled-coil domain containing 68	18q21.2	as a tumor-suppressive role in pancreatic ductal adenocarcinoma, nove causal gene of schizophrenia by GWAS.
BIRC3	baculoviral IAP repeat containing 3	11q22.2	inhibits cell apoptosis and senescence, as a biomarker of mesenchymal habitat of glioblastoma.
COL22A1	collagen type XXII alpha 1 chain	8q24.23-q24.3	to contribute to the stabilization of myotendinous junctions and strengthen skeletal muscle attachments during contractile activity, to maintain vascular integrity.
LOC101927355	uncharacterized LOC101927355	20q13.12	highly expressed in placenta.
DEPDC1	DEP domain containing 1	1p31.3	regulates tumor proliferation and metastasis in hepatocellular cancinoma, promotes cell proliferation and prostate tumor growth, is a novel cell cycle related gene that regulates mitotic progression.

Table 4-4 The overlap 24 differential genes expressed in common siRNA a and siRNA b groups. The location and gene functions or lung biology were summarised from GeneCards (187) and Entrez (188).

4.4.3.3 Gene set enrichment analysis (GSEA) suggests NPNT involvement in inflammation response pathways

4.4.3.3.1 Curated dataset analysis of DEGs after siRNA a transfection suggests roles for NPNT in several inflammation response pathways

227 pathways were upregulated and 353 pathways were downregulated in HBECs cultured from NPNT siRNA a treatment compared to scrambled group (FDR <5%). Table **4**-**5** demonstrates a summary of top 20 upregulated (A) and top 20 downregulated (B) pathways which altered in NPNT siRNA a treated HBECs based on analyses using GSEA Curated database. The curated database identified several pathways involved in inflammation response after *NPNT* knockdown by siRNA a in HBECs.

(A)

GENE SET	SIZE	ES	NES	FDR q-val	LEADING EDGE		
BROWNE_INTERFERON_RESPONSIVE_GENES	63	0.88	3.21	0	tags=76%	list=7%	signal=82%
BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE	73	0.80	2.97	0	tags=53%	list=7%	signal=57%
MOSERLE_IFNA_RESPONSE	31	0.93	2.95	0	tags=90%	list=4%	signal=94%
SANA_RESPONSE_TO_IFNG_UP	72	0.79	2.93	0	tags=61%	list=10%	signal=68%
TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_3D_UP	166	0.70	2.92	0	tags=35%	list=5%	signal=37%
HECKER_IFNB1_TARGETS	86	0.78	2.92	0	tags=53%	list=5%,	signal=56%
DAUER_STAT3_TARGETS_DN	47	0.87	2.90	0	tags=70%	list=5%,	signal=74%
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	61	0.80	2.89	0	tags=59%	list=12%,	signal=67%
DER_IFN_ALPHA_RESPONSE_UP	69	0.77	2.84	0	tags=39%	list=4%,	signal=41%
SEITZ_NEOPLASTIC_TRANSFORMATION_BY_8P_DELETION_UP	65	0.76	2.77	0	tags=40%	list=3%	signal=41%
REACTOME_INTERFERON_SIGNALING	147	0.66	2.73	0	tags=32%	list=10%	signal=35%
DER_IFN_BETA_RESPONSE_UP	96	0.70	2.70	0	tags=28%	list=4%	signal=29%
BENNETT_SYSTEMIC_LUPUS_ERYTHEMATOSUS	29	0.84	2.66	0	tags=62%	list=5%	signal=65%
TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_10D_UP	168	0.63	2.64	0	tags=27%	list=5%	signal=29%
FARMER_BREAST_CANCER_CLUSTER_1	40	0.79	2.62	0	tags=38%	list=1%	signal=38%
UROSEVIC_RESPONSE_TO_IMIQUIMOD	21	0.92	2.61	0	tags=76%	list=4%	signal=79%
RADAEVA_RESPONSE_TO_IFNA1_UP	48	0.76	2.61	0	tags=56%	list=11%	signal=63%
ZHANG_INTERFERON_RESPONSE	21	0.91	2.59	0	tags=81%	list=4%	signal=84%
STAMBOLSKY_TARGETS_OF_MUTATED_TP53_DN	43	0.77	2.57	0	tags=51%	list=9%	signal=56%
EINAV_INTERFERON_SIGNATURE_IN_CANCER	25	0.85	2.54	0	tags=80%	list=11%	signal=89%

(B)

NAME	SIZE	ES	NES	FDR q-val	LEADING EDGE		
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER	127	-0.78	-3.35	0	tags=69%	list=11%	signal=77%
FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN	174	-0.73	-3.34	0	tags=58%	list=9%	signal=63%
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	135	-0.76	-3.26	0	tags=71%	list=16%	signal=84%
KOBAYASHI_EGFR_SIGNALING_24HR_DN	229	-0.68	-3.21	0	tags=62%	list=16%	signal=72%
WHITEFORD_PEDIATRIC_CANCER_MARKERS	105	-0.75	-3.18	0	tags=60%	list=11%	signal=67%
FISCHER_DREAM_TARGETS	846	-0.61	-3.18	0	tags=45%	list=18%	signal=53%
CROONQUIST_NRAS_SIGNALING_DN	69	-0.79	-3.11	0	tags=62%	list=8%	signal=68%
CROONQUIST_IL6_DEPRIVATION_DN	93	-0.75	-3.11	0	tags=60%	list=9%	signal=66%
DUTERTRE_ESTRADIOL_RESPONSE_24HR_UP	290	-0.64	-3.08	0	tags=55%	list=16%	signal=65%
SARRIO_EPITHELIAL_MESENCHYMAL_TRANSITION_UP	155	-0.69	-3.06	0	tags=46%	list=11%	signal=51%
LEE_EARLY_T_LYMPHOCYTE_UP	95	-0.74	-3.05	0	tags=55%	list=9%	signal=60%
KONG_E2F3_TARGETS	87	-0.76	-3.05	0	tags=60%	list=9%	signal=65%
CHANG_CYCLING_GENES	135	-0.71	-3.05	0	tags=59%	list=12%	signal=66%
KANG_DOXORUBICIN_RESISTANCE_UP	48	-0.83	-3.02	0	tags=75%	list=8%	signal=82%
ZHANG_TLX_TARGETS_DN	83	-0.74	-3.00	0	tags=65%	list=14%	signal=75%
ZHOU_CELL_CYCLE_GENES_IN_IR_RESPONSE_6HR	75	-0.77	-3.00	0	tags=79%	list=15%	signal=93%
RUIZ_TNC_TARGETS_DN	133	-0.69	-2.99	0	tags=53%	list=16%	signal=63%
ZHOU_CELL_CYCLE_GENES_IN_IR_RESPONSE_24HR	117	-0.70	-2.98	0	tags=69%	list=16%	signal=82%
FISCHER_G2_M_CELL_CYCLE	207	-0.64	-2.97	0	tags=42%	list=11%	signal=47%
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	145	-0.67	-2.97	0	tags=53%	list=16%	signal=63%

Table 4-5. “Curated” pathways of top 20 upregulated (A) and top 20 downregulated (B) gene sets regulated in NPNT siRNA a treated HBECs. SIZE=number of genes in the gene set after filtering out those not in the expression dataset. ES=enrichment score for the gene set. NES=normalised enrichment score. Q-value=false discovery rate – the estimated probability that the NES represents a false positive. Leading edge subset = subset of members that contribute the most to the ES. Tags= the percentage of genes contributing to the ES. List= percentage of genes in the ranked gene list before the peak in the running ES. This gives an indication of where in the list the ES is attained. Signal = the ES strength that combines the two previous statistics.

4.4.3.3.2 Curated dataset analysis of DEGs after siRNA b transfection suggests roles for NPNT in several inflammation response pathways

92 pathways were upregulated and 143 pathways were downregulated in HBECs cultured from NPNT siRNA b treatment compared to scrambled group (FDR <5%). **Table 4-6** shows a summary of the top 20 upregulated (A) and top 20 downregulated (B) pathways which altered in NPNT siRNA b treated HBECs based on analyses using GSEA Curated database. Similarly, curated database identifies several pathways are involved in inflammation response after NPNT knockdown by siRNA b in HBECs.

(A)

NAME	SIZE	ES	NES	FDR q-val	LEADING EDGE		
BROWNE_INTERFERON_RESPONSIVE_GENES	63	0.68	2.56	0	tags=63%	list=17%	signal=77%
SANA_RESPONSE_TO_IFNG_UP	72	0.64	2.49	0	tags=56%	list=19%	signal=68%
MOSERLE_IFNA_RESPONSE	31	0.75	2.42	0	tags=77%	list=17%	signal=94%
DAUER_STAT3_TARGETS_DN	47	0.66	2.33	0	tags=60%	list=18%	signal=72%
HELLEBREKERS_SILENCED_DURING_TUMOR_ANGIOGENESIS	72	0.59	2.32	0	tags=31%	list=9%	signal=34%
SEITZ_NEOPLASTIC_TRANSFORMATION_BY_8P_DELETION_UP	65	0.61	2.31	0	tags=49%	list=11%	signal=55%
WORSCHECH_TUMOR_EVASION_AND_TOLEROGENICITY_UP	28	0.73	2.31	0	tags=50%	list=15%	signal=59%
HECKER_IFNB1_TARGETS	86	0.57	2.31	0	tags=44%	list=13%	signal=51%
UROSEVIC_RESPONSE_TO_IMIQUIMOD	21	0.78	2.30	0	tags=67%	list=9%	signal=73%
ZHANG_INTERFERON_RESPONSE	21	0.75	2.29	2.59E-04	tags=81%	list=17%	signal=98%
EINAV_INTERFERON_SIGNATURE_IN_CANCER	25	0.74	2.26	2.35E-04	tags=68%	list=17%	signal=82%
BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE	73	0.58	2.24	4.28E-04	tags=51%	list=18%	signal=61%
JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP	67	0.59	2.24	3.95E-04	tags=39%	list=16%	signal=46%
WONG_ENDMETRIUM_CANCER_DN	70	0.57	2.18	0.001197	tags=31%	list=12%	signal=35%
TURASHVILI_BREAST_LOBULAR_CARCINOMA_VS_ LOBULAR_NORMAL_DN	70	0.55	2.14	0.002326	tags=31%	list=11%	signal=35%
DER_IFN_ALPHA_RESPONSE_UP	69	0.54	2.10	0.004042	tags=43%	list=20%	signal=55%
STEGER_ADIPOGENESIS_DN	23	0.70	2.09	0.004339	tags=35%	list=9%	signal=38%
BENNETT_SYSTEMIC_LUPUS_ERYTHEMATOSUS	29	0.66	2.09	0.004243	tags=72%	list=26%	signal=97%
KRASNOSELSKAYA_ILF3_TARGETS_UP	37	0.63	2.09	0.004222	tags=51%	list=17%	signal=62%
ANASTASSIOU_MULTICANCER_INVASIVENESS_SIGNATURE	60	0.56	2.09	0.004011	tags=30%	list=10%	signal=33%

(B)

NAME	SIZE	ES	NES	FDR q-val	LEADING EDGE		
ROSTY_CERVICAL_CANCER_PROLIFERATION _CLUSTER	127	-0.69	-2.91	0	tags=50%	list=11%	signal=56%
FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN	174	-0.65	-2.81	0	tags=49%	list=12%	signal=55%

SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	135	-0.67	-2.80	0	tags=58%	list=20%	signal=71%
LEE_EARLY_T_LYMPHOCYTE_UP	95	-0.68	-2.73	0	tags=46%	list=10%	signal=51%
ZHAN_MULTIPLE_MYELOMA_PR_UP	39	-0.78	-2.65	0	tags=62%	list=10%	signal=68%
CROONQUIST_NRAS_SIGNALING_DN	69	-0.68	-2.60	0	tags=51%	list=15%	signal=59%
FARMER_BREAST_CANCER_CLUSTER_2	29	-0.81	-2.59	0	tags=62%	list=9%	signal=68%
ZHOU_CELL_CYCLE_GENES_IN_IR_RESPONSE_24HR	117	-0.62	-2.56	0	tags=56%	list=22%	signal=72%
FISCHER_G2_M_CELL_CYCLE	207	-0.58	-2.55	0	tags=33%	list=10%	signal=37%
AMUNDSON_GAMMA_RADIATION_RESPONSE	38	-0.76	-2.55	0	tags=61%	list=9%	signal=66%
REICHERT_MITOSIS_LIN9_TARGETS	26	-0.85	-2.54	0	tags=77%	list=10%	signal=86%
ODONNELL_TFRC_TARGETS_DN	119	-0.61	-2.53	0	tags=40%	list=11%	signal=45%
KONG_E2F3_TARGETS	87	-0.64	-2.50	0	tags=63%	list=21%	signal=80%
REACTOME_DEPOSITION_OF_NEW_CENPA_ CONTAINING_NUCLEOSOMES_AT_THE_CENTROMERE	49	-0.71	-2.50	0	tags=63%	list=20%	signal=79%
PID_PLK1_PATHWAY	40	-0.72	-2.46	0	tags=45%	list=10%	signal=50%
CHANG_CYCLING_GENES	135	-0.58	-2.46	0	tags=50%	list=20%	signal=63%
KANG_DOXORUBICIN_RESISTANCE_UP	48	-0.70	-2.46	0	tags=60%	list=15%	signal=71%
CHIANG_LIVER_CANCER_SUBCLASS_ PROLIFERATION_UP	160	-0.57	-2.45	0	tags=33%	list=11%	signal=37%
WHITEFORD_PEDIATRIC_CANCER_MARKERS	105	-0.59	-2.43	0	tags=50%	list=20%	signal=63%
EGUCHI_CELL_CYCLE_RB1_TARGETS	23	-0.80	-2.39	0	tags=70%	list=11%	signal=78%

Table 4-6. “Curated” pathways of top 20 upregulated (A) and top 20 downregulated (B) gene sets regulated in NPNT siRNA b treated HBECs. SIZE=number of genes in the gene set after filtering out those not in the expression dataset. ES=enrichment score for the gene set. NES=normalised enrichment score. Q-value=false discovery rate – the estimated probability that the NES represents a false positive. Leading edge subset = subset of members that contribute the most to the ES. Tags= the percentage of genes contributing to the ES. List= percentage of genes in the ranked gene list before the peak in the running ES. This gives an indication of where in the list the ES is attained. Signal = the ES strength that combines the two previous statistics.

4.4.3.4 Overlap pathways between siRNA a and siRNA b using Curated dataset analysis

4.4.3.4.1 Upregulated pathways overlap

227 pathways and 92 pathways were upregulated in HBECs cultured from NPNT siRNA a and siRNA b compared to scrambled group (FDR <5%), respectively. There are 50 “replicated” pathways upregulated by NPNT knockdown (Figure **4-12**). The top 20 upregulated pathways are listed in Table 4-7.

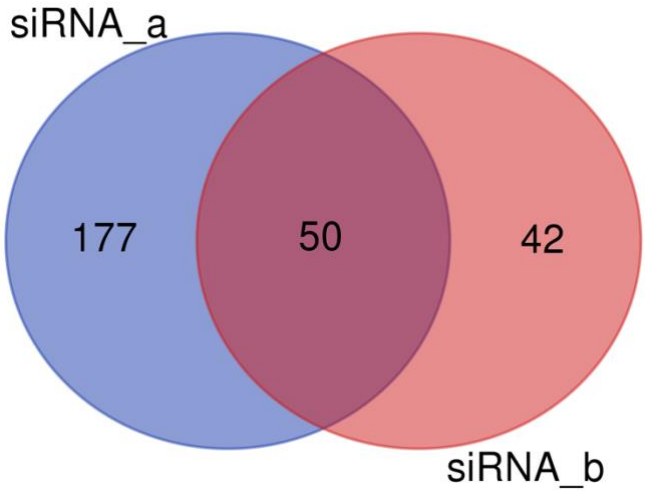


Figure 4-12 Venn diagram showing the overlap of upregulated pathways in cells treated with siRNA a and siRNA b for 48h. There were 227 and 92 pathways were upregulated in HBECs when knockdown NPNT by siRNA a and siRNA b, respectively. There were 50 pathways upregulated in common between both groups.

NAME OF PATHWAYS
BROWNE_INTERFERON_RESPONSIVE_GENES
SANA_RESPONSE_TO_IFNG_UP
MOSERLE_IFNA_RESPONSE
DAUER_STAT3_TARGETS_DN
HELLEBREKERS_SILENCED_DURING_TUMOR_ANGIOGENESIS
SEITZ_NEOPLASTIC_TRANSFORMATION_BY_8P_DELETION_UP
WORSCHECH_TUMOR_EVASION_AND_TOLEROGENICITY_UP
HECKER_IFNB1_TARGETS
UROSEVIC_RESPONSE_TO_IMIQUIMOD
ZHANG_INTERFERON_RESPONSE
EINAV_INTERFERON_SIGNATURE_IN_CANCER
BOSCO_INTERFERON_INDUCED_ANTIVIRAL_MODULE
JECHLINGER_EPITHELIAL_TO_MESENCHYMAL_TRANSITION_UP
DER_IFN_ALPHA_RESPONSE_UP
BENNETT_SYSTEMIC_LUPUS_ERYTHEMATOSUS
KRASNOSELSKAYA_ILF3_TARGETS_UP
ANASTASSIOU_MULTICANCER_INVASIVENESS_SIGNATURE

Table 4-7. The top 20 pathways upregulated overlap between siRNA a and siRNA b identified by Curated dataset.

4.4.3.4.2 Downregulated pathways overlap

353 pathways and 143 pathways were downregulated in HBECs cultured from NPNT siRNA a and siRNA b respectively compared to the scrambled group (FDR <5%). There are 128 “replicated” pathways downregulated by NPNT knockdown (Figure **4-13**). The top 20 downregulated pathways are listed in Table 4-8.

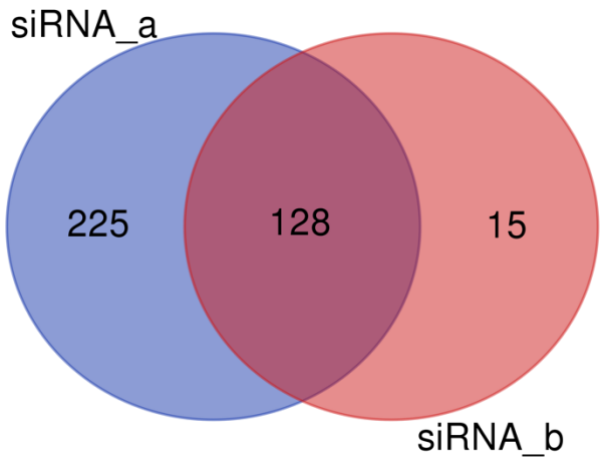


Figure 4-13 Venn diagram showing the overlap of downregulated pathways in cells treated by siRNA a and siRNA b for 48h. There were 353 and 143 pathways downregulated in HBECs with knockdown NPNT by siRNA a and siRNA b, respectively. There are 128 pathways downregulated in both groups.

NAME OF PATHWAYS
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER
FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP
LEE_EARLY_T_LYMPHOCYTE_UP
ZHAN_MULTIPLE_MYELOMA_PR_UP
CROONQUIST_NRAS_SIGNALING_DN
FARMER_BREAST_CANCER_CLUSTER_2
ZHOU_CELL_CYCLE_GENES_IN_IR_RESPONSE_24HR
FISCHER_G2_M_CELL_CYCLE
AMUNDSON_GAMMA_RADIATION_RESPONSE
REICHERT_MITOSIS_LIN9_TARGETS
ODONNELL_TFRC_TARGETS_DN
KONG_E2F3_TARGETS
REACTOME_DEPOSITION_OF_NEW_CENPA_CONTAINING_ NUCLEOSOMES_AT_THE_CENTROMERE
PID_PLK1_PATHWAY
CHANG_CYCLING_GENES
KANG_DOXORUBICIN_RESISTANCE_UP
CHIANG_LIVER_CANCER_SUBCLASS_PROLIFERATION_UP
WHITEFORD_PEDIATRIC_CANCER_MARKERS
EGUCHI_CELL_CYCLE_RB1_TARGETS

Table 4-8. The top 20 pathways downregulated in HBECs both siRNA a and siRNA b identified by Curated dataset.

4.5 Discussion

In this chapter, the main aims were to investigate the differentially expressed genes and gene sets (pathways) following *NPNT* knockdown in HBECs by siRNA to provide novel insight into the potential biological roles of *NPNT* in the lung and identify pathways implicated in human bronchial epithelial cells homeostasis.

To identify the contribution of NPNT to human bronchial epithelial cell homeostasis, I established the use of NPNT RNAi for knockdown in HBECs. NPNT knockdown by siRNAs (siRNA a, b, and c) at different concentrations (0.1, 1, and 10nM) for 24h, 48h or 72h, respectively was assessed. NPNT knockdown was observed in HBECs at the mRNA level using Taqman PCR. The optimal concentration of NPNT siRNA a and b were 1nM at 48h. Therefore, efficient and reliable knockdown of NPNT can be achieved in primary HBECs.

From Microarray data analysis, a total of 24 differentially expressed genes were identified after *NPNT* knockdown consistently after siRNA a and siRNA b knockdown in HBECs after 48h treatment. This included 11 up-regulated, 10 down-regulated, 3 inconsistently regulated (*ZBED6*, *C20orf197*, *PAM*) genes. Furthermore, 158 DEGs were identified between *NPNT* siRNA a samples and scrambled samples, and 389 DEGs were identified between *NPNT* siRNA b and scrambled samples.

GAS2L3, CBX1, CENPA, PRR11, CDK6, CDKN3 and DEPDC1 were connected network nodes among 24 DEGs by STRING website (https://cn.string-db.org/). These six genes were downregulated except CBX1, when NPNT was knocked down in HBECs. Growth arrest specific 2 like 3 (**GAS2L3**) plays an important role of cardiomyocyte cell cycle regulation in embryonic development (189). A GWAS of lung squamous cell carcinoma has reported that a SNP rs12296850 located in GAS2L3, which is a target of DREAM complex and plays an important role in accurate cell division (190, 191). Chromobox 1 (**CBX1**) protein may perform a vital role in epigenetic control of chromatin structure and gene expression and have an involvement in oncogenic activity (192). Centromere protein A (**CENPA**) performs as an epigenetic mark and is involved in replication and cell division (193, 194). Proline rich 11 (**PRR11**) can promote cell proliferation, migration, and invasion (195). This novel gene regulated cell cycle, implicated in tumorigenesis via upregulating in lung cancer, and suggested as a potential therapeutic target in lung cancer (196, 197). Cyclin dependent kinase 6 (**CDK6**) protein plays an important role in cell cycle G1 phase progression and G1/S transition, and it also has been identified in many human cancers (198). Cyclin dependent kinase inhibitor 3 (**CDKN3**) as a cyclin-dependent kinase inhibitor, promotes cancer cell proliferation and tumorigenesis by targeting p27 (199). This gene has been reported to be mutated or overexpressed in some cancers (200). DEP domain containing 1 (**DEPDC1**) is required for cell cycle progression and promotes cell proliferation in some cancers, such as hepatocellular

carcinoma and breast cancer cells (201, 202). A recent study reported that miR-130a induces A549 lung adenocarcinoma cells apoptosis via inhibiting DEPDC1 expression (203). Taken together, all these differentially expressed genes all implicated in the cell cycle, cell proliferation, and migration consistent with other studies (204, 205). This evidence potentially provides insight into the novel function of NPNT gene and protein in human bronchial epithelial cells as well as COPD.

GSEA pathway analysis was carried out using “Curated” databases. There are 50 “replicated” pathways were significantly upregulated and 128 “replicated” pathways were downregulated by NPNT knockdown (both siRNA a and siRNA b samples). The most upregulated pathways were the interferon (IFN) signalling, suggesting involvement in inflammatory process. STAT3, extracellular matrix, and epithelial to mesenchymal transition (EMT) were also upregulated which might represent an inflammatory or airway remodelling phenotype. The downregulation gene sets consisted of cervical cancer proliferation cluster, breast cancer grade, E2F/E2F3, EMT, chang cycling genes, G1/S and G2/M cell cycle, senescence TP53 target, which demonstrate regulation cell proliferation, division, senescence and tumorigeneses. In addition, EGFR signalling and Interleukin 6 (IL-6) pathways, potentially leading to altered inflammation were also identified.

In summary, Microarray studies do identify genes which are potentially differentially regulated, and for those genes which were present in both

experiments 21 out of 24 showed consistent directions of effect which is encouraging. However, the lack of significant knockdown in NPNT itself and the cluster dendrogram and PCA analyses suggest that Microarray based approaches do not show enough reproducibility overall. The enrichment analyses were not performed directly using the DEGs that overlapped between siRNA a and siRNA b because there were only 21 genes. Hence, I went on to undertake RNA-seq analyses in an attempt to provide more definitive data.

5. Comparison of gene expression changes measured by Microarray and RNA-sequencing in human bronchial epithelial cells engineered with NPNT knockdown

5.1 Introduction

In chapter 4, I described experiments designed to assess the downstream DEGs and gene sets (pathways) using Microarray after NPNT knockdown by siRNAs. The result of RNAi KD NPNT in HBECs indicates that NPNT plays important roles in several inflammation response pathways and cell proliferation. Gene expression profiles can be assessed by Microarray and, more commonly now, by next generation sequencing technologies (RNA-seq) (183). In Microarray, tens of thousands of complementary DNA (cDNA) fragments or oligonucleotides (probes) hybridize with the specific RNA sequences which have been labelled. The raw data from Microarray is then normalized before any analysis. An alternative approach is to use RNA sequencing. RNA sequencing is a widely used method for the exploration of differential gene expression, differential splicing of mRNAs and other aspects RNA biology (206). RNA sequencing can directly explore the transcript of each sample, without the need to design probes in advance as in Microarray (207, 208). RNA sequencing identifies novel transcripts and evaluates thousands of genes rather than being limited to prior designed targets, and coverage of all genes and mRNA splice forms (Figure **5**-**1**). An obvious drawback of the data presented in chapter 4 was the failure to observe significant knockdown of NPNT using Microarrays, despite having previously validated the siRNAs to be used. Therefore, in this chapter, investigated differential gene expression following NPNT knockdown by siRNA using RNA sequencing, and then compared the downstream DEGs and pathways

identified by Microarray and RNA sequencing. Furthermore, I found in Chapter 4 that the differentially expressed genes after knockdown NPNT were implicated in the cell cycle, cell proliferation, and migration. NPNT could promote cell differentiation and proliferation, including MC3T3-E1 cells (123), and MDPC-23 (209). Thus, in this chapter the cell viability, proliferation or migration of HBECs were measured by MTT and scratch wound assay.

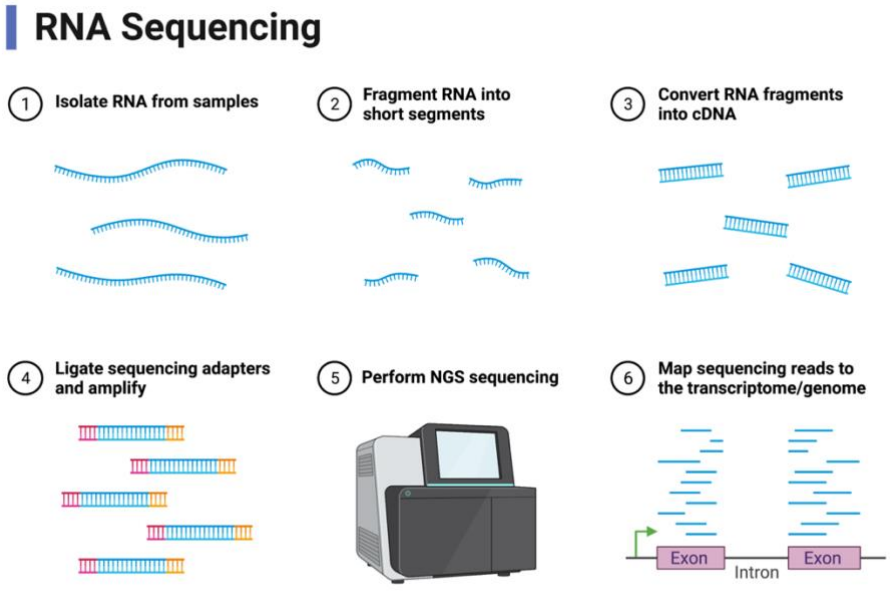


Figure 5-1 Overview of RNA sequencing. RNA extracted from HBECs was then converted into complementary DNA (cDNA). Sequencing adapters were ligated to the ends of the cDNA fragments and NGS sequencing performed using a NovaSeq 6000 system. Finally, sequencing reads were mapped to the transcriptome/genome for data analysis. The figure was created in BioRender (https://app.biorender.com/).

5.2 Aims

In this chapter, the main aims were to i) identify the downstream differentially expressed genes and pathways using siRNA knockdown of NPNT in HBECs using RNA sequencing, and ii) confirm whether the differentially expressed genes and pathways detected by Microarray and RNA-seq would provide insight of the biological mechanisms underlying the role of NPNT in HBECs. iii) detect the potential functions of NPNT on cell viability, proliferation and migration after knockdown of NPNT in HBECs.

5.3 Methods

5.3.1 Cell Biology Methods

5.3.1.1 Cell culture of human bronchial epithelial cells

Passage 3 of Donors D494599 and D328324 HBECs were cultured in T75 flask until 70%-80% confluent as detailed in Chapter 2 for RNA sequencing. Prior to siRNA transfected HBECs were seeded to 6 well plates at a density of 130, 000 cells/well.

5.3.1.2 NPNT gene knock-down using siRNA

The *NPNT* siRNA was designed and knocked down as described in Chapter 2. Briefly, passage 3 HBECs (repeated 3 times in 2 donors,) were cultured and transfected with 1nM of siRNA a and siRNA b, or 1nM of scrambled control. For an additional control I used untransfected cells. Each condition was studied in 2 wells with 200µl of transfection mix added to each well. After 48h incubation, the media was aspirated and discarded. The cells were collected as described in Chapter 2 and the cell lysates were stored at -80℃ until further use.

5.3.2 MTT assay

The Methyl Thiazolyl Tetrazolium bromide, or thiszolyl blue tetrazolium bromide (MTT) assay was used to quantify mitochondrial dehydrogenase activity. This assay measures cellular respiration which has been previously been extrapolated to indicate the rate of cell viability(210).

In brief, HBECs were cultured as described previously. 0.5mg/mL MTT was made by dissolving 5mg MTT in 10mL BEGM media and filtered through 0.2μm filter and warmed to 37℃. The BEGM media of 6-well-plate was aspirated, replaced with 2mL of the 5mg MTT solution, and returned to the incubator for 3-4h. Then, the 0.5mg MTT solution was aspirated and discarded, and 1mL isopropanol was added to dissolve the resulting MTT formazan crystals. The isopropanol solution was aliquoted into a 96-well-plate at the volume 200μL/well. Isopropanol (200ul/well) was added into 3 wells as control so that background absorbance can be measured and substracted from the result. The solution was incubated for 10min at room temperature. The 96-well-plate was measured at a wavelength of 570nm (with a background subtraction of 690nm) using the plate reader flexstation.

5.3.3 Scratch Wound assay

The scratch wound assay has been widely applied as a model for cell injury and as an indicator of cell migration or proliferation (211, 212). This assay involves the disruption of nearly confluent HBEC monolayer in a 6-well-plate by scratching with a sterile p200 pipette tip vertically in the bottom and then culture in incubator to repair the scratch wound.

Briefly, the HBECs were cultured in 6-well-plates, marked with 3 lines in the bottom of 6-well-plate by blade before cell culture. The cells were cultured in bronchial epithelial wounding media (BEWM, without Hydrocortisons, Epinephrin or Epithelial Growth Factor Supplements) 24h before scratch wounding. When the HBECs grew to >90%

confluence, cultures were wounded with a sterile p200 pipette tip to form 3 vertical wounds. The wells were washed with warmed sterile PBS and fed with fresh warmed 2mL BEWM. The wounded HBECs were visualised over 24h (0h, 2h, 4h, 8h, 12h, 24h) at 9 sites in each well using a Nikon Diaphot 300 Inverted Phase Contrast Microscope with the 4× objective lens. The wounds areas were calculated by ImageJ. The data was presented as percentage wound remaining rather than area healed in order to compensate for variation in wound size.

5.3.4 Molecular Biology Methods

5.3.4.1 RNA extraction and quality control

RNA was extracted from cell lysates and cDNA synthesised as detailed in the general material and methods chapter 2. RNA was detected for quality and quantity as explained in Chapter 2.

5.3.4.2 RNA sequencing

5.3.4.2.1 Library preparation and sequencing

Briefly, library preparation and sequencing were carried out by Oxford Genomics Centre. Paired and sequencing was then carried out using NovaSeq 6000 system to generate ≥30 million read pairs per sample for species with reference genome. Full details of this method can be found in chapter 2.

5.3.4.2.2 Trimming and alignment

In brief, the fastq files were transferred into the desired directory on the University of Nottingham High Performance Computer (HPC) using WinSCP. Alignment of RNA-Seq reads to genome build GRCh37 was carried out as described in chapter 2. A file containing the adaptor sequence (as shown below) which was provided by Oxford Genome Centre was used to adapter reads by Trimmomatic. The STAR package is a large mapping set of high-throughput sequencing reads to a reference genome in RNA sequencing data analysis (144, 145). STAR generates several output files including files containing counts for each gene and BAM files.

Oxford Genomics uses standard Illumina multiplexing Y adapter sequences. Multiplex adapter1 5’ (A)GAT CGG AAG AGC ACA CGT CT (trim read 1 for this sequence). Multiplex adapter2 5’ ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TC*T (trim read 2 for the reverse compliment of this sequence). Reverse compliment (AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT)

5.3.4.2.3 Differential expression genes

The Bioconductor project offers a number of efficiently RNA-seq data analysis tools (184). In the work described in this chapter, I used ***edgeR*** package to import, organise, and filter data (213), and the ***limma*** package with its voom method to assess differentially expressed genes (141, 184). Furthermore, the ***Glimma*** package was used to present differential expression results by graphical plots. Significance

was defined using an adjusted p-value cut-off set at 5% by default. The R studio script can be found in the chapter 9 appendix.

5.3.4.2.4 Gene Set Enrichment Analysis

As described previously in Chapter 2. GSEA was used to identify pathways when comparing between scrambled and siRNA treatment. The curated dataset and Hallmark dataset were applied to generate gene sets which were regulated by NPNT knockdown. The command can be found in chapter 9 appendix.

5.3.4.2.5 Graphical visualisation of RNA-seq results

Volcanos and heatmaps were used to represent differentially expressed genes, which were generated from R studio. From the R Bioconductor, the ***ggplots***, ***ggrepel***, ***limma***, ***edgeR***, ***Glimma*** and ***vennDiagram*** packages were installed and used: the commands can be found in chapter 9 appendix (184-186).

5.3.5 Statistical analysis

Gene count values generated from fastq files were used in independent linear regression models (empirical bayes method) including within subject correlation for matched samples across each experiment using the limma package to identify differentially expressed genes. For RNA-seq data, log-CPM mean-variance relationship is accommodated using precision weights calculated by the voom function. Significant genes were classified based on a Benjamini-Hochberg false discovery rate (FDR) less than 0.05 to correct for multiple testing.

5.4 Results

5.4.1 qRT-PCR confirms NPNT knockdown efficiency using siRNAs

NPNT expression was successfully reduced after transfection with 1nM siRNA a (remaining expression 23.40±8.91%) and siRNA b (remaining expression15.80±8.99%) and confirmed by Taqman PCR (Figure **5-2**). There were significantly different compare to scrambled siRNA group (remaining expression 94.00±43.50%).

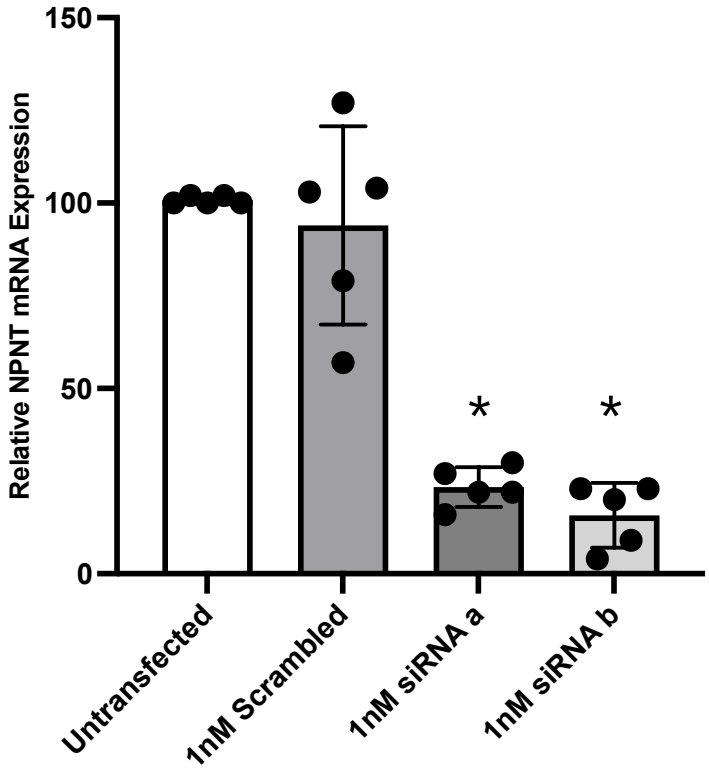


Figure 5-2 Expression of NPNT mRNA in HBECs following NPNT siRNA medicated knockdown at 48h. The knockdown efficiency was detected by qRT-PCR of NPNT mRNA from Untransfected, scrambled and 2 types of siRNA (compared NPNT levels to scramble, *p<0.05, n=5).

5.4.2 Quality control of RNA samples for RNA-seq

RNA samples were extracted and the quality control information for RNA sequencing is shown in Table 5-1 . The table shows treatment conditions, the RNA concentration, wavelength ratio (260/280 ratio) detected by Nanodrop, and RIN scores determined by Agilent 2100 Bioanalyser and Agilent 4200 TapeStation as detailed in chapter 2. All samples (RIN>8) were considered to be of sufficient quality for RNA sequencing by Oxford Genomics Centre. 6 sets of experiments were performed, 3 replicates in each of two donors (D494599, D328324), however, the RNA sample quality of one of the experiments (E8.4) of D328324 was not of high enough quality for RNA-seq and hence this sample was not used.

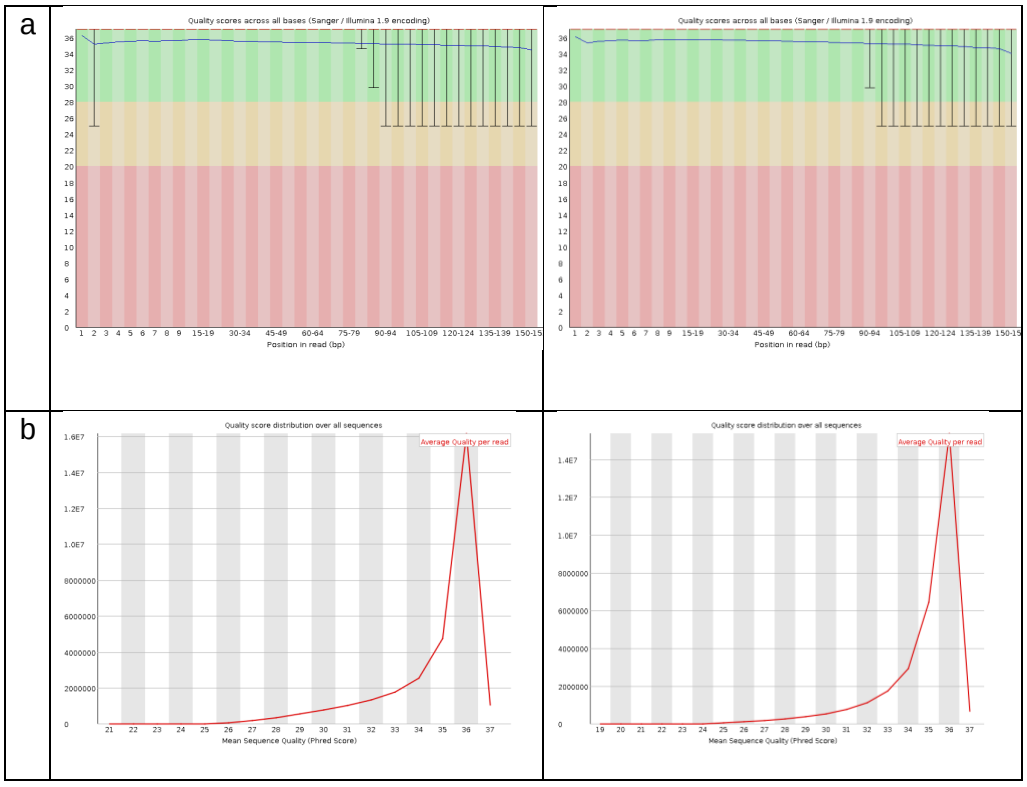
Donor	Treatment (48h)	Conc.ng//μl	260/280	RIN
D494599 P3	E8.1 Untransfected HBEC	208.0	2.05	9.7
	E8.1 1nM scrambled HBEC	229.0	2.05	9.7
	E8.1 1nM NPNT siRNA-A HBEC	185.0	2.06	9.3
	E8.1 1nM NPNT siRNA-B HBEC	202.0	2.05	9.5
D328324 P3	E8.2 Untransfected HEBC	263.0	2.03	9.8
	E8.2 1nM scrambled HBEC	278.0	2.02	9.7
	E8.2 1nM NPNT siRNA-A HBEC	196.0	2.03	9.7
	E8.2 1nM NPNT siRNA-B HBEC	262.3	2.03	9.7
D328324 P3	E8.3 Untransfected HBEC	159.0	2.05	9.5
	E8.3 1nM scrambled HBEC	199.0	2.04	9.7
	E8.3 1nM NPNT siRNA-A HBEC	136.0	2.05	9.1
	E8.3 1nM NPNT siRNA-B HBEC	174.0	2.04	9.5
D494599 P3	E8.5 Untransfected HBEC	252.0	2.06	10
	E8.5 1nM scrambled HBEC	194.0	2.06	10
	E8.5 1nM NPNT siRNA-A HBEC	120.0	2.08	10
	E8.5 1nM NPNT siRNA-B HBEC	265.0	2.03	10
D494599 P3	E8.6 Untransfected HBEC	230.0	2.04	10
	E8.6 1nM Scrambled HBEC	242.0	2.04	10
	E8.6 1nM NPNT siRNA-A HBEC	121.0	2.04	10
	E8.6 1nM NPNT siRNA-B HBEC	267.0	2.02	10

Table 5-1 Information of RNA samples quality control for RNA-seq using siRNAs to knockdown *NPNT* in HBECs. The details included treatment conditions, concentration of RNA, 260/280 ratio, and RIN. P3, passage 3.

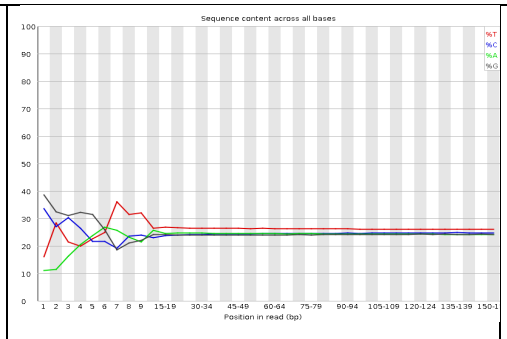
5.4.3 RNA-seq analysis using siRNA to knockdown NPNT in HBECs

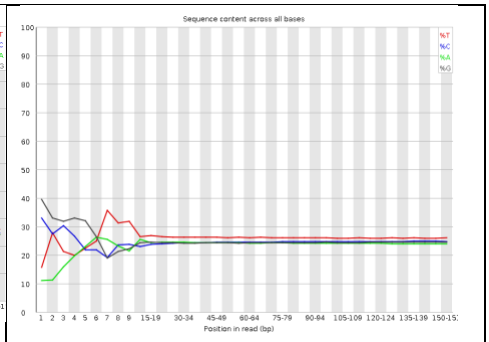
5.4.3.1 Sequencing data quality control

The quality of sequencing data was confirmed using the FastQC tool. This tool was utilised and it provides a number of outputs, such as (a) Per base sequence quality, (b) per sequence quality scores, (c) per base sequence content, (d) per base GC content, (e) per sequence GC content, (f) per base N content, (g) sequence length distribution, (h) sequence duplication levels (i) Kmer content. An example of good quality sequencing data is shown in Figure 5-3.

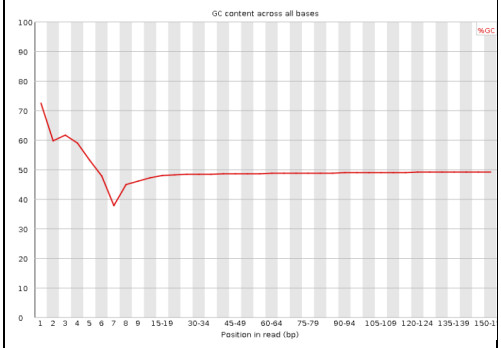


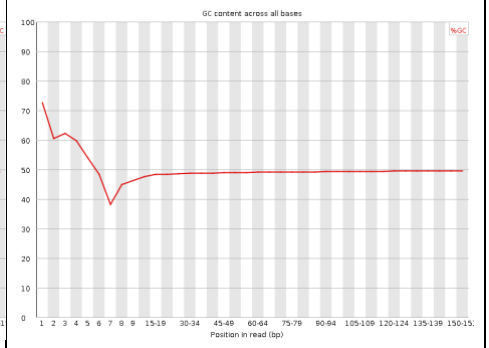
c


Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)

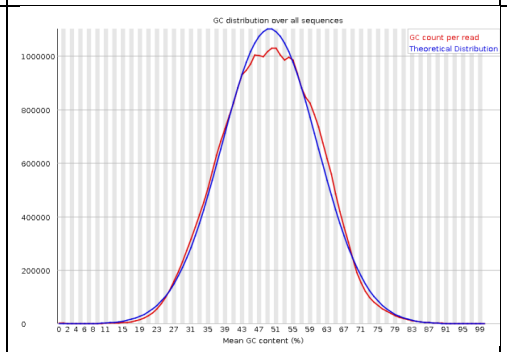

Sequence content across all bases
%T
%C
%A
%G
Position in read (bp)

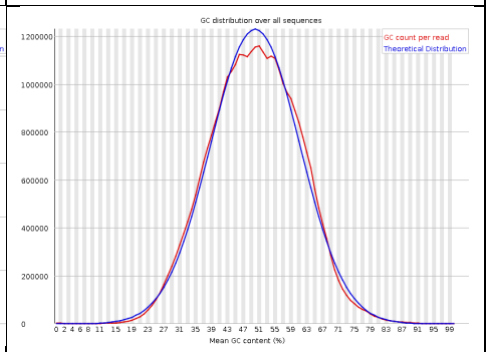
d


GC content across all bases
%GC
Position in read (bp)

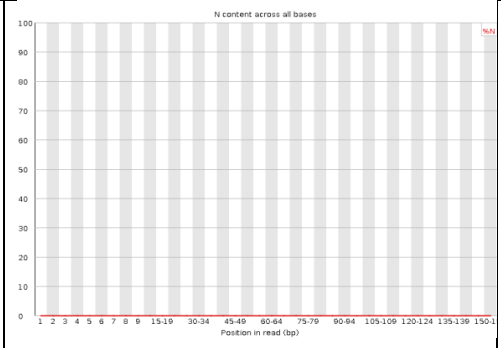

GC content across all bases
%GC
Position in read (bp)

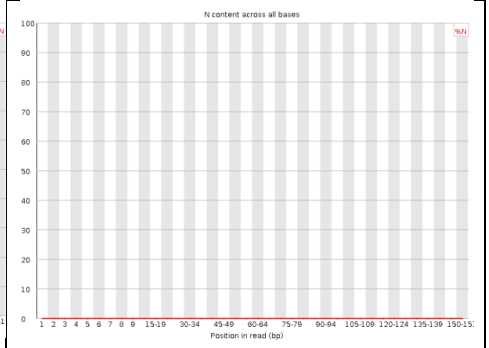
e


GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)

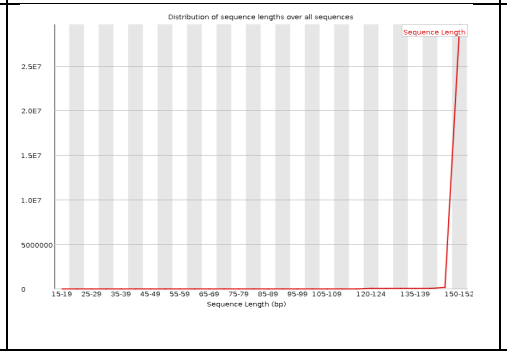

GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)

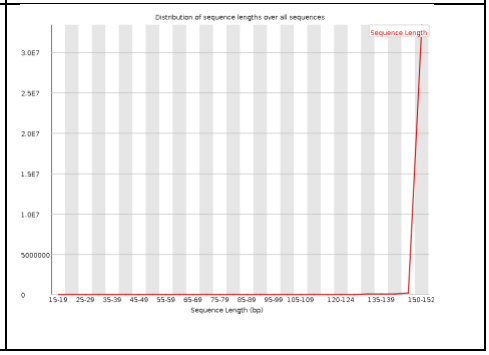
f


N content across all bases
%N
Position in read (bp)


N content across all bases
%N
Position in read (bp)

g


Distribution of sequence lengths over all sequences
Sequence Length
Sequence Length (bp)


Distribution of sequence lengths over all sequences
Sequence Length
Sequence Length (bp)

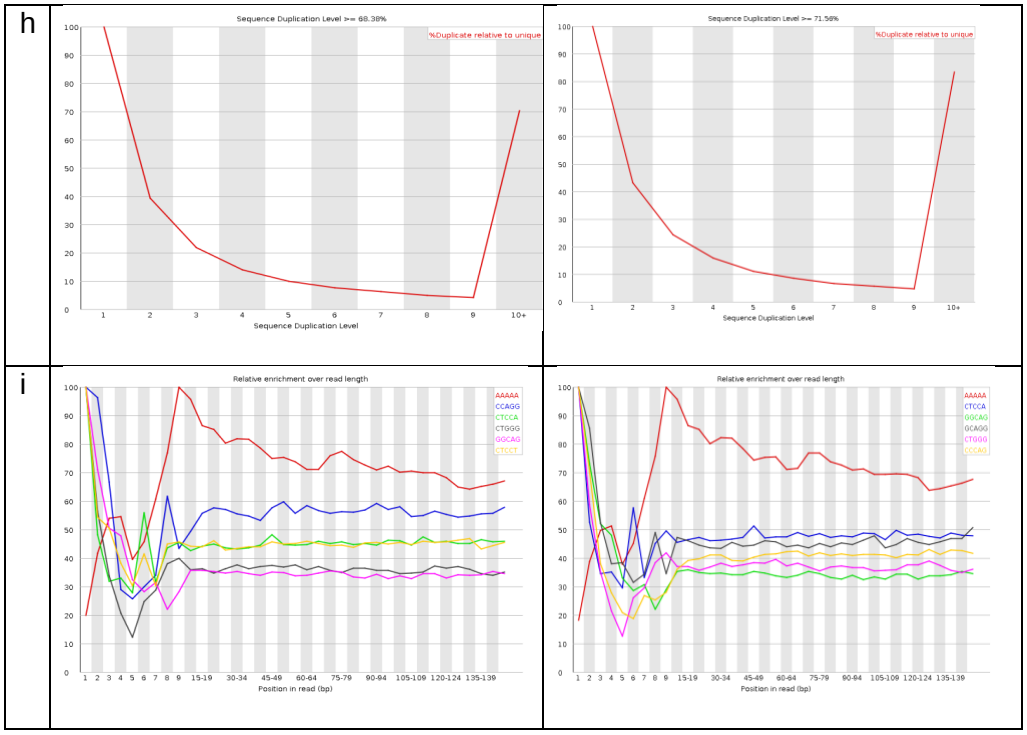


Figure 5-3 FastQC confirmed Read files for quality control. All parameters detected passed quality control analysis as follows: (a) Per base sequence quality assessment showed that all reads met q score >20, (b) per sequence quality scores ≥27, (c) per base sequence content ≤20%, (d) per base GC content ≤10%, (e) per sequence GC content 15-30%, (f) per base N content =0 (failure ≥20%), (g) sequence length distribution analysis did not identify read sequences of varying lengths, (h) sequence duplication levels ≤20%, (i) Kmer content ≤10 fold. Read 1 files (V1, forward reads) and Read 2 files (V2, reverse reads) are shown on the left and right, respectively. Data shown are a representative example from one set of Read files from untransfected (left) and siRNA a (right) treated cells.

5.4.3.2 Differential Gene Expression Analysis

5.4.3.2.1 Removing genes that are lowly expressed

The gene dataset includes the expressed and not expressed genes. There were 25,723 genes in this data set which have zero counts across all 20 samples. Genes which are lowly expressed across all samples are filtered out because they are not biologically relevant and this reduces the number of statistical tests that are performed. The ***edgeR*** package filterByExpr provides an automatic way to filter genes in order to keeping as many genes as possible with worthwhile counts (Figure **5-4**). Using the package filterByExpr, grouping samples based on siRNA condition, 15,656 genes remained for analysis.

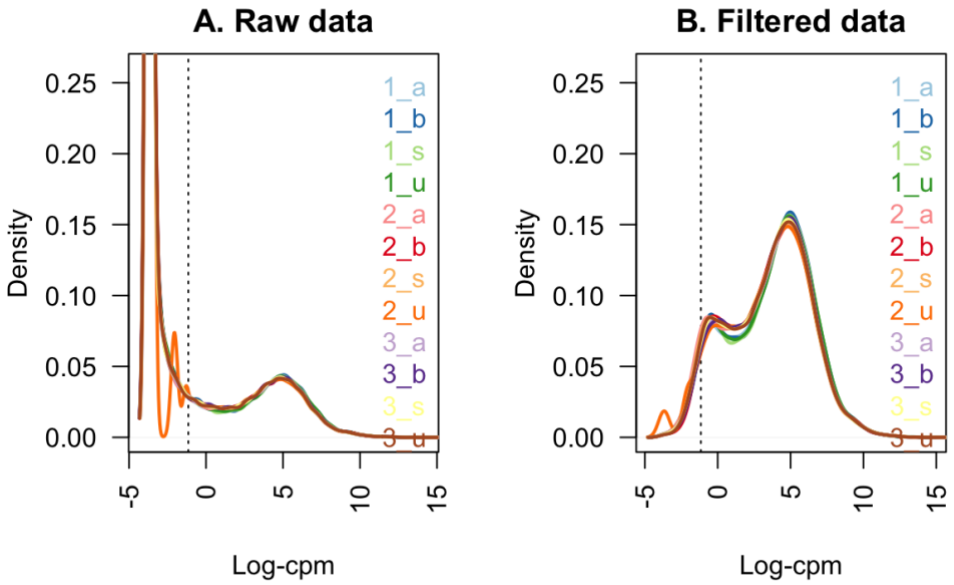


Figure 5-4 The density of log-CPM values for raw pre-filtered data (A) and post filtered data (B). CPM, counts per million. Log2-CPM, log2-counts per million.

5.4.3.2.2 Unsupervised clustering of samples

Multi-dimensional scaling plots were constructed to look at sample clustering. The *plotMDS* function of *limma* package in R studio was used to draw the plots as shown in Figure **5**-**5** which shows that samples clustered mainly on donor rather than transfection condition.

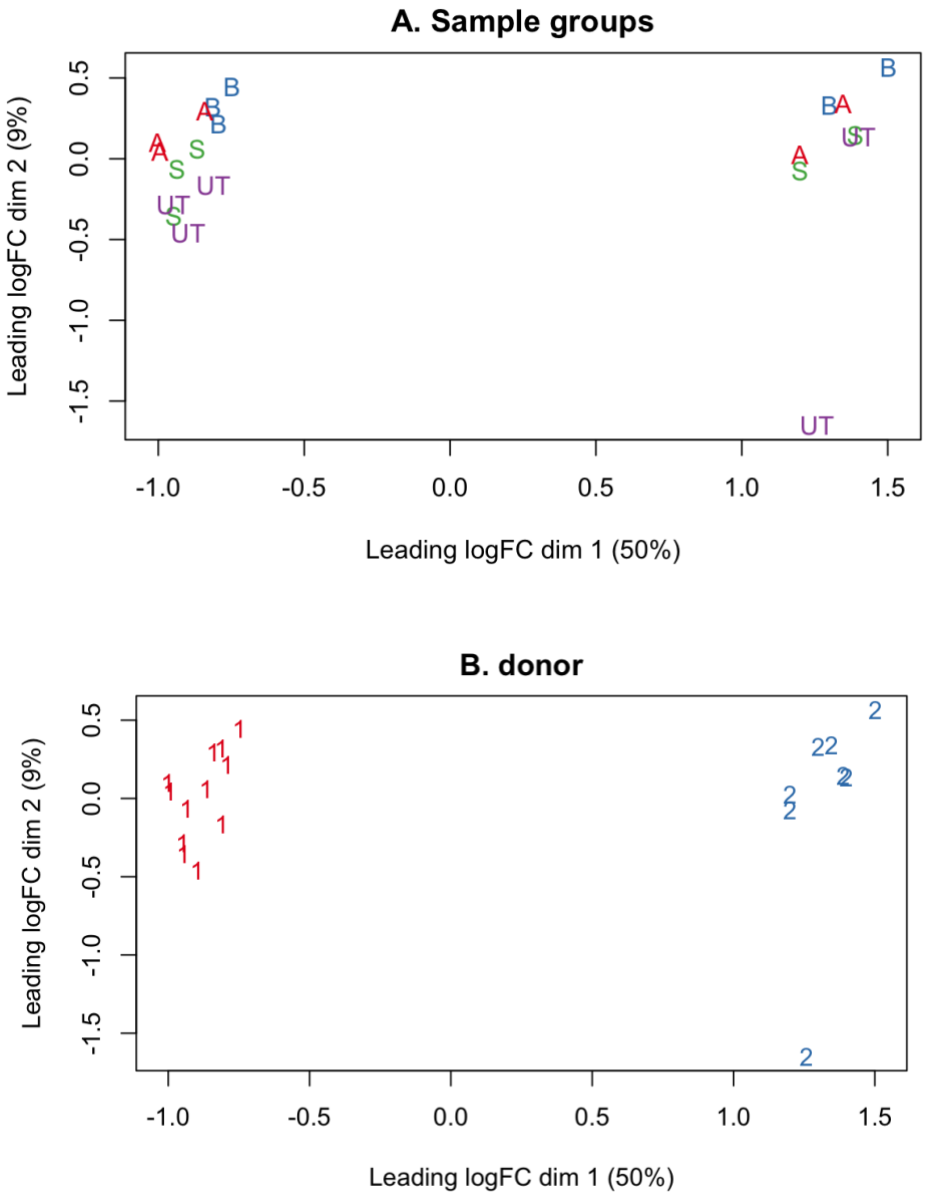


Figure 5-5 Multi-dimensional scaling (MDS) plots of log-CPM values over dimensions with samples coloured and labelled by sample groups (A) and by donors (B). CPM, counts per million. There is clear clustering by donor number.

5.4.3.3 Differentially expressed genes in RNA-seq

In this chapter, the differentially expressed genes in each different treatment group was profiled. Initially, a design matrix is set up. The ***makeContrasts*** function of ***limma*** package was applied to compare condtions (siRNA a or b) to scrambled control.

Initial analyses showed that in contrast with the Microarray experiments described in the previous chapter, significant NPNT knockdown was seen in the RNA-seq data following treatment with siRNA a and siRNA b.

The number of significantly up- and down-regulated genes are summarised in Table 5-2. In summary, 466 genes were detected to be down-regulated in siRNA a compared to scrambled with 566 genes up-regulated. In addition, there were 1460 down-regulated and 1346 up-regulated with siRNAb compared to scrambled.

	SvsUT	**AvsS**	**BvsS**
Down regulation	121	466	1460
NotSig	15530	14624	12850
Up regulation	5	566	1346

Table 5-2 Summary of the number of differentially expressed genes from RNA sequencing results. Significance is defined as adjusted p-value cut-off that is set at 0.05. The comparison of scrambled vs Untransfected was excluded from further analyses. SvsUT, scrambled

vs untransfected; AvsS, siRNA a vs Scrambled; siRNA b vs Scrambled. NotSig, not significant statistics.

5.4.3.3.1 DEGs analysis of Scrambled vs Untransfected

Interactive mean-difference plot generated using *Glimma* of scrambled versus untransfected is shown in Figure 5-6.There were 15530 genes not significantly different when scrambled versus untransfected cells, and only 121 differentially expressed genes down-regulated and 5 genes were up-regulated (Figure 5-7 and Figure **5**-**8**).

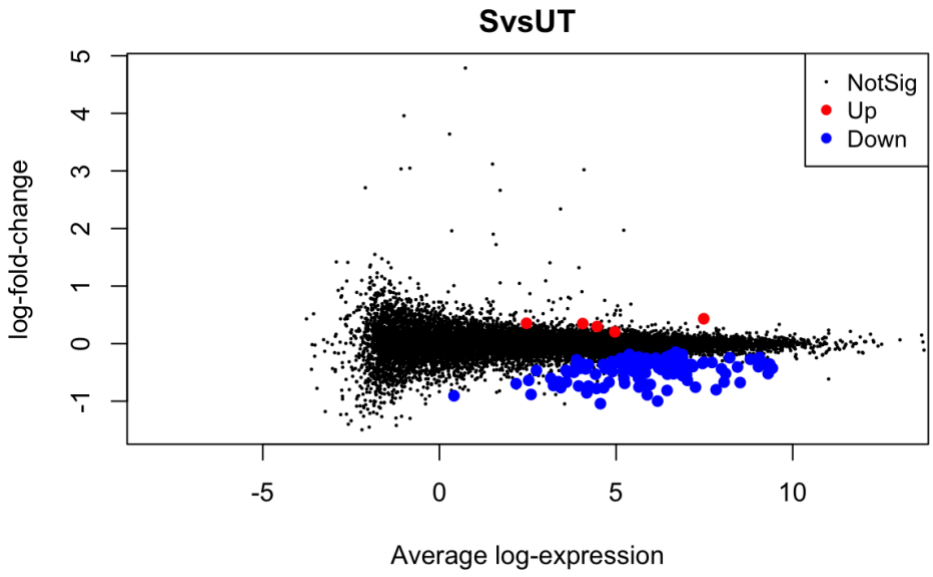


Figure 5-6 Interactive mean-difference plot generated using Glimma. Summary data (log-FCs verse Log-CPM value) are shown in this figure which is linked to each value per samples for a selected gene scrambled group compared to untransfected group. CPM, counts per million**.** SvsUT, Scrambled vs Untransfected.

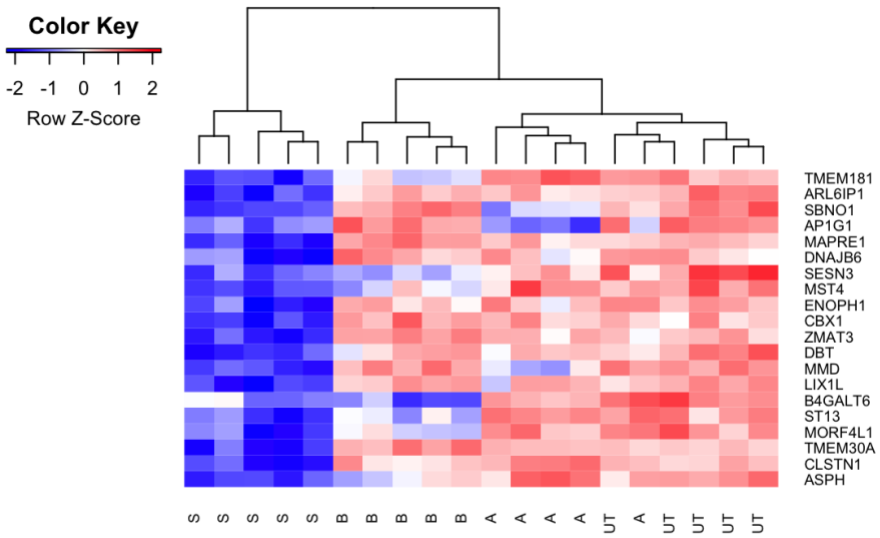


Figure 5-7 Heatmap of Z-score values for top 20 differentially expressed genes from scrambled group compared to other 3 groups in HBECs for 48h. Samples with absolute higher expression of a gene are showed in red and samples with absolute lower expression of a gene are showed in blue .

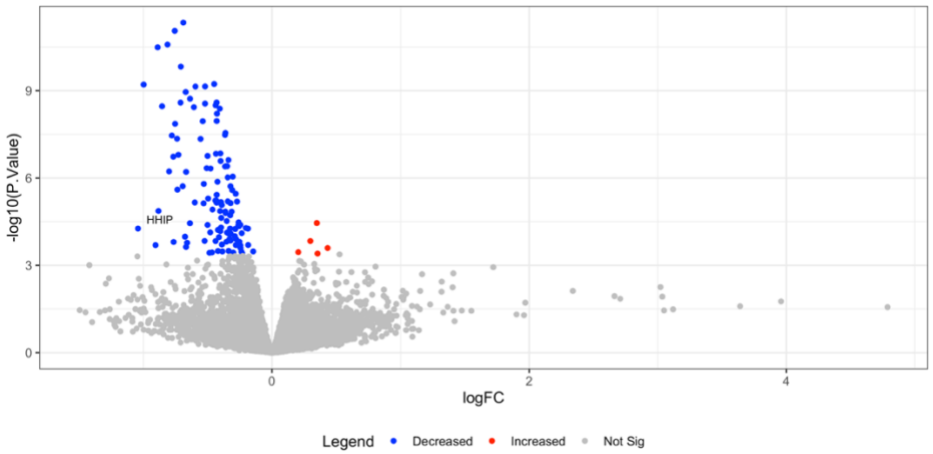


Figure 5-8 Volcano plot of scrambled group compared to untransfected group. Each point represents a transcript. Only significant genes in (5% FDR) and log FC >1 or <-1 are colour coded in the volcano plot.

5.4.3.3.2 DEGs analysis in siRNA a

The *plotMD* function of ***Glimma*** package generates log-FC from the linear model fit against the average log-CPM expression values plots as shown in Figure **5**-**9**.

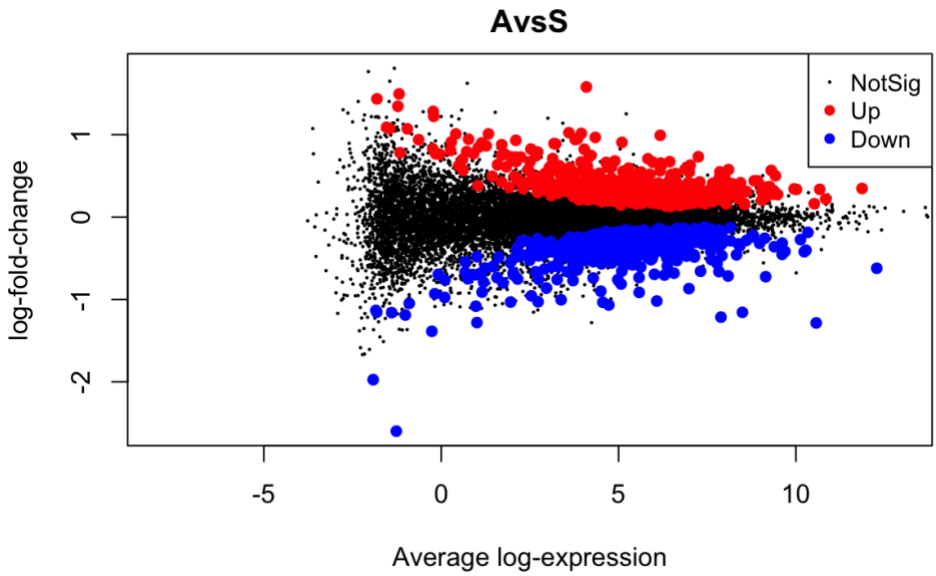


Figure 5-9 Interactive mean-difference plot following treatment with siRNA a generated using Glimma. Summary data (log-FCs verse Log-CPM expression value) are shown in this figure which is linked to each value per samples for a selected gene siRNA a group compared to scrambled. CPM, counts per million. AvsS, siRNA a vs Scrambled.

The most significant DEGs are visualised in heatmaps, which provides insight into the differential gene expression across individual groups and samples. Heatmaps were created for the top 20 DEGs (plotted by log-CPM values) from siRNA a versus scrambled. The heatmap of siRNA a compared to scrambled is shown in Figure **5**-**10**. The top

differentially expressed genes were selected by log FC >1 or <-1 and were used to label in the volcano plot (Figure 5-11).

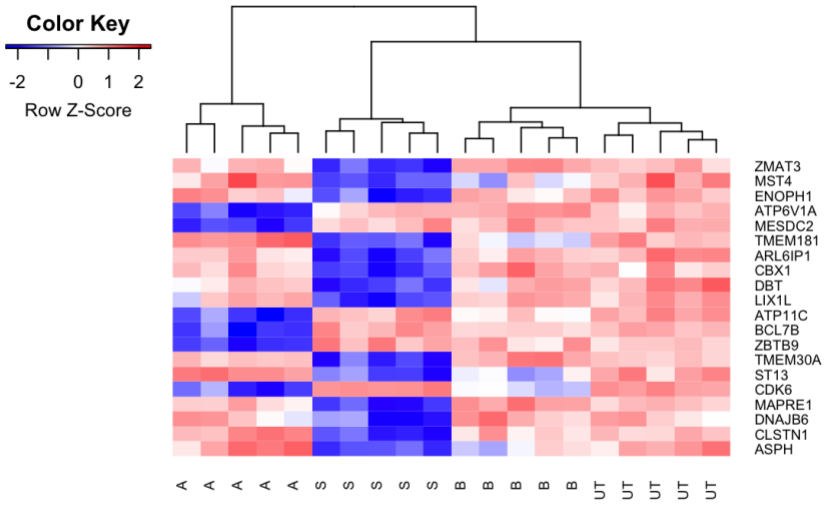


Figure 5-10 Heatmap of Z-score values for top 20 differentially expressed genes from siRNA a compared to other 3 groups in HBECs for 48h. Red marked genes are those with absolute higher gene expression and the blue marked those genes with absolute lower gene expression.

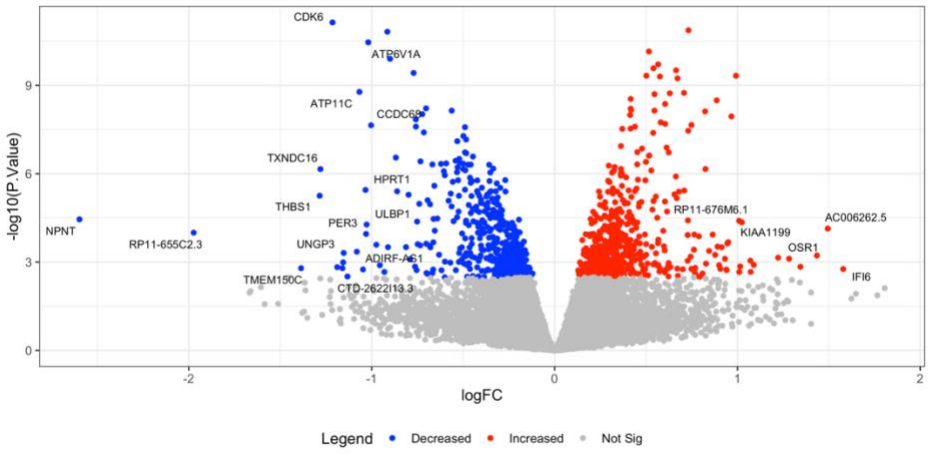


Figure 5-11 Volcano plot of siRNA a compared to scrambled group. Each point represents a transcript. Only significant genes in (adjusted

BH p<0.05) and log FC >1 or <-1 are labelled with text. P values were computed with adjusted BH p<0.05.

Among the 566 up-regulated differentially expressed genes, IFI6 (Interferon Alpha Inducible Protein 6, logFC 1.58), AC006262.5 (lncRNA, logFC 1.49) were the most upregulated genes which ranked by adjusted p-value firstly and then by logFC. Moreover, there are many lncRNAs (e.g., AC006262.5, LINC00240, RP11-353N4.4, RP11-565F19.2, and RP11-676M6.1) up-regulated (Table 5-3).

Gene Symbol	logFC	Gene function
IFI6	1.58	IFI6 (Interferon Alpha Inducible Protein 6) is associated with Interferon gamma signaling and Immune response IFN alpha/beta signaling pathway.
AC006262.5	1.49	AC006262.5 is a lncRNA associated with the risk of breast cancer.
RP11-353N4.4	1.43	lncRNA
RP11-565F19.2	1.34	lncRNA
OSR1	1.28	OSR1 (Odd-Skipped Related Transcription Factor 1) is related to this gene include nucleic acid binding.
SP140	1.22	SP140 (SP140 Nuclear Body Protein) is related to DNA-binding transcription factor activity and chromatin binding.
LINC00240	1.09	lncRNA, suppresses invasion and migration.
AC000068.5	1.07	-
PPP1R9A	1.07	PPP1R9A (Protein Phosphatase 1 Regulatory Subunit 9A) is related to actin binding and actin filament binding.
KIAA1199	1.02	CEMIP (Cell Migration Inducing Hyaluronidase 1) is a prognostic biomarker of cancer, involves in cellular migration and invasion of lung cancer.

ISG15	1.01	ISG15 (ISG15 Ubiquitin Like Modifier) is related pathways are Interferon gamma signaling and RIG-I/MDA5 mediated induction of IFN-alpha/beta pathways.
VTCN1	1.01	VTCN1 (V-Set Domain Containing T Cell Activation Inhibitor 1) is related pathways are Cell adhesion molecules and NF-κB Signaling.
RP11-676M6.1	1.01	lncRNA

Table 5-3 Top up-regulated genes by siRNA a versus scrambled after 48h ranked by adjusted p-value firstly and then by logFC. Adjusted p-value <0.05 and LogFC >1.0. Gene functions or biology were obtained from GeneCards (187) and Entrez Gene (188).

Among 466 differentially expressed genes, *NPNT* was the most downregulated genes, which demonstrates an efficient knockdown of *NPNT* expression by siRNA a in HBECs when this is assessed by RNA-seq.

THBS1 (Thrombospondin 1, logFC -1.28) encodes protein which mediates cell-to-cell and cell-to-matrix interactions, encodes protein also can bind to fibrinogen, laminin, type V collagen and integrins alpha-V/beta-1. CDK6 (Cyclin dependent kinase 6, logFC -1.21), which plays a crucial role in the cell cycle, was significantly downregulated, which suggests that NPNT may be involved in cell proliferation in HBECs. As shown in Table 5-4, the top 20 downregulation genes by siRNA a versus scrambled after 48h ranked by adjusted p-value firstly and then by logFC.

Gene SYMBOL	logFC	Gene function
NPNT	-2.60	Nephronectin
RP11-655C2.3	-1.97	lncRNA
TMEM150C	-1.39	TMEM150C (Transmembrane Protein 150C) involves in ion transmembrane transport.
THBS1	-1.28	THBS1 (Thrombospondin 1) is related to pathways of Proteoglycans in cancer and Degradation of the extracellular matrix.
TXNDC16	-1.28	TXNDC16 (Thioredoxin Domain Containing 16)
CDK6	-1.21	CDK6(Cyclin dependent kinase 6) protein plays an important role in cell cycle G1 phase progression and G1/S transition, and it also has been identified in many human cancers. Cyclin dependent kinase inhibitor 3 (CDKN3) as a cyclin-dependent kinase inhibitor, promotes cancer cell proliferation and tumorigenesis by targeting p27. This gene has been reported to be mutated or overexpressed in some cancers.
CCDC181	-1.19	CCDC181 (Coiled-Coil Domain Containing 181)
AP000692.9	-1.16	-
INHBA	-1.16	INHBA (Inhibin Subunit Beta A) involves in pathways NODAL and Peptide hormone metabolism.
UNGP3	-1.15	UNGP3 (Uracil-DNA Glycosylase Pseudogene 3) is a Pseudogene.
CTD-2622I13.3	-1.13	-
ADIRF-AS1	-1.08	lncRNA
ATP11C	-1.07	ATP11C (ATPase Phospholipid Transporting 11C) is related to pathways are Cardiac conduction and Ion channel transport.
ASGR1	-1.05	ASGR1 (Asialoglycoprotein Receptor 1) associates with cardiovascular risk.
HPRT1	-1.03	HPRT1 (Hypoxanthine Phosphoribosyltransferase 1) involves in ATP/ITP metabolism.

PER3	-1.03	PER3 (Period Circadian Regulator 3) is described as a core component of the circadian clock.
ULBP1	-1.03	ULBP1 (UL16 Binding Protein 1) leads to activation of several signal transduction pathways, including those of JAK2, STAT5, ERK and PI3K kinase/Akt.
ATP6V1A	-1.02	ATP6V1A (ATPase H+ Transporting V1 Subunit A) related pathways are Innate Immune System and Insulin receptor recycling.
CCDC68	-1.00	CCDC68 (Coiled-Coil Domain Containing 68) demonstrates a tumour-suppressive role.

Table 5-4 Top 20 downregulation genes by siRNA a versus scrambled after 48h ranked by adjusted p-value firstly and then by logFC. Adjusted p-value <0.05 and LogFC ≥1.0. Gene functions or biology were descripted from GeneCards (187) and Entrez Gene (188).

5.4.3.3.3 DEGs analysis in siRNA b

As shown in Figure 5-12, R studio generates log-FC from the linear model fit against the average log-CPM expression values plots. The interactive mean-difference plot was generated using *Glimma* of siRNA b versus scrambled as shown in Figure 5-12.

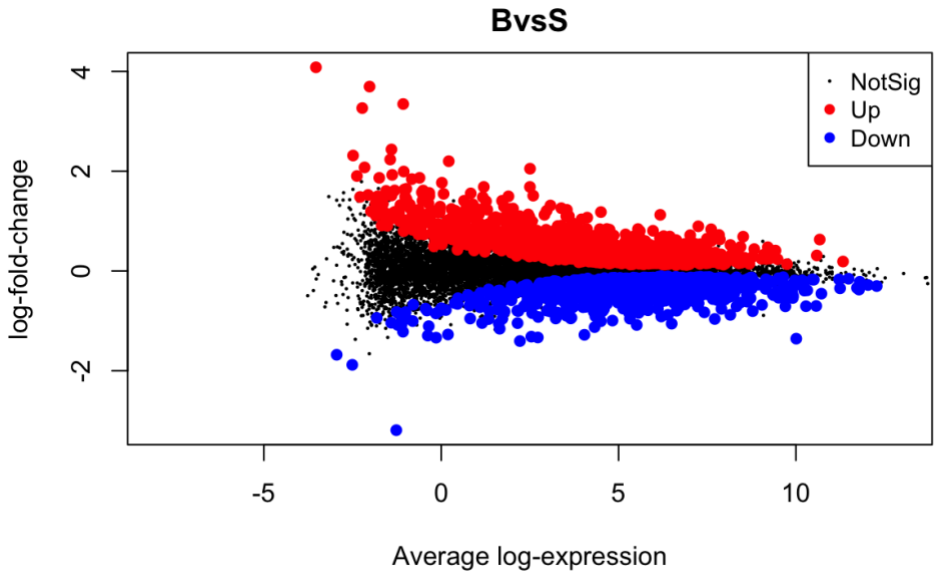


Figure 5-12 Interactive mean-difference plot generated using Glimma. Summary data (log-FCs verse Log-CPM value) are shown in this figure which is linked to each value per sample for a selected gene siRNA b group compared to scrambled. CPM, counts per million. BvsS, siRNA b vs Scrambled.

Among the 1346 upregulated DEGs, the top 20 genes were ranked by adjusted p-value and then by log-CPM as shown in Figure 5-13 as heatmap when compared siRNA b to scrambled. The top DEGs were selected by FDR <5% and log FC >1 or <-1 which then were used to label the volcano plot (Figure **5-14**).

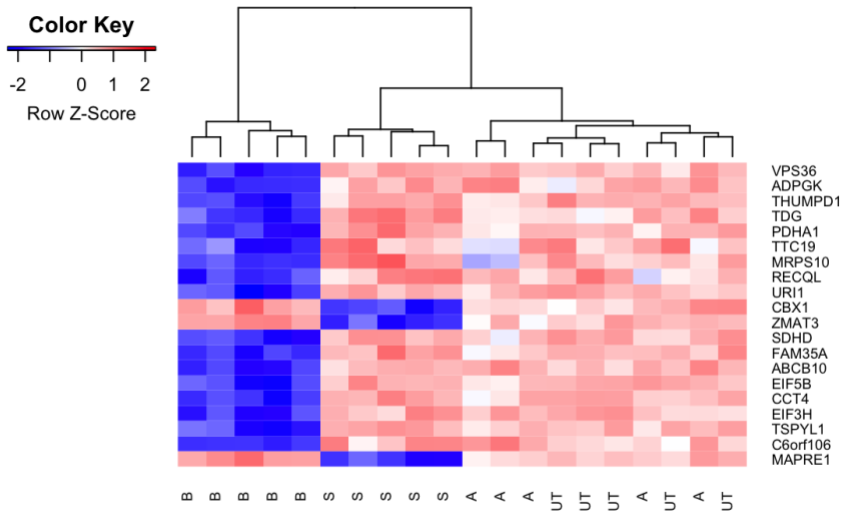


Figure 5-13 Heatmap of Z-score values for top 20 differentially expressed genes from siRNA b compared to other 3 groups in HBECs for 48h. Samples with absolute higher expression of a gene are showed in red and samples with absolute lower expression of a gene are showed in blue.

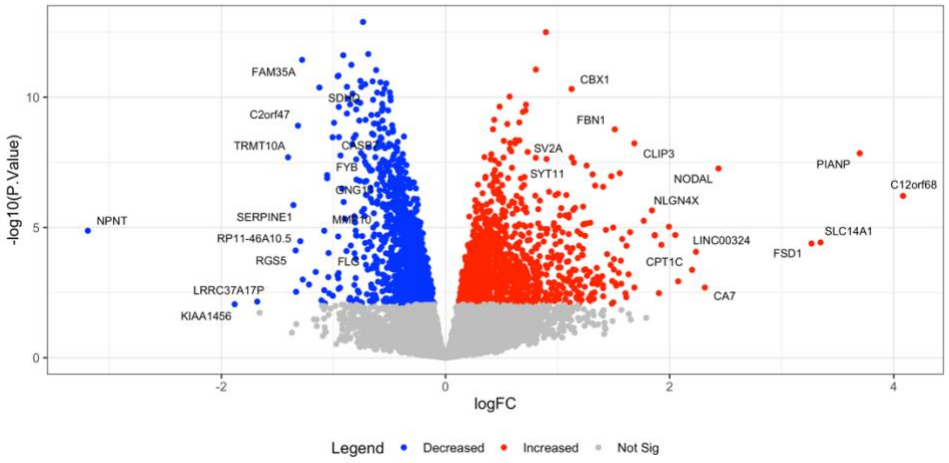


Figure 5-14 Volcano plot of siRNA b compared to Scrambled group. Each point represents a transcript. Only significant genes in (adjusted BH p<0.05) and log FC >1 or <-1 are labelled. P values were computed with adjusted BH p<0.05.

Among the 1346 up-regulated differentially expressed genes in siRNA b versus scrambled group, C12orf68(Coiled-Coil Domain Containing 184, log FC 4.08), PIANP(PILR Alpha Associated Neural Protein, logFC 3.70) and SLC14A1 (Solute Carrier Family 14 Member 1 (Kidd Blood Group, logFC 3.35) were the most up-regulated DE genes. As shown Table 5-5, the top 20 up-regulation genes by siRNA b versus scrambled after 48h ranked by adjusted p-value firstly and then by logFC.

Gene symbol	**logFC**	**Gene function**
C12orf68	4.08	Now also called CCDC184 (Coiled-Coil Domain Containing 184)
PIANP	3.70	PIANP (PILR Alpha Associated Neural Protein) involves in immune regulation.
SLC14A1	3.35	SLC14A1 (Solute Carrier Family 14 Member 1 (Kidd Blood Group)) is related to water transmembrane transporter activity and urea channel activity, could be induced by cytokines.
FSD1	3.27	FSD1 (Fibronectin Type III And SPRY Domain Containing 1) involves in the stability and organization of microtubules during cytokinesis (involves in cell cycle).
NODAL	2.44	NODAL (Nodal Growth Differentiation Factor) involves in developmental biology and Tgif disruption of Shh signalling pathways.
CA7	2.31	CA7 (Carbonic Anhydrase 7) involves in Nitrogen metabolism and Metabolism pathways.
LINC00324	2.23	lncRNA. It involves in cell proliferation, invasion, and migration.
CPT1C	2.20	CPT1C (Carnitine Palmitoyltransferase 1C) involves in Adipocytokine signalling pathway and AMP-activated Protein Kinase (AMPK) Signalling.
NOVA2	2.08	NOVA2 (NOVA Alternative Splicing Regulator 2) is related to nucleic binding and RNA binding.
THY1	2.05	THY1 (Thy-1 Cell Surface Antigen) associates with protein kinase binding and GTPase

		activator activity, may play a role in cell-cell or cell-ligand interactions.
DHRS2	1.99	DHRS2 (Dehydrogenase/Reductase 2) attenuates MDM2-mediated p53/TP53 degradation, leading to p53/TP53 stabilization and increased transcription activity, resulting in the accumulation of MDM2 and CDKN1A/p21.
B4GALNT1	1.93	B4GALNT1 (Beta-1,4-N-Acetyl-Galactosaminyltransferase 1) involved in the biosynthesis of gangliosides GM2, GD2, GT2 and GA2 from GM3, GD3, GT3 and GA3, respectively.
SPRN	1.90	SPRN (Shadow Of Prion Protein) is related to metabolism of proteins and post-translational modification-synthesis of GPI anchored proteins.
MIAT	1.87	MIAT (Myocardial Infarction Associated Transcript) is affiliated with lncRNA class. It associated myocardial infarction.
GRIA3	1.87	GRIA3 (Glutamate Ionotropic Receptor AMPA Type Subunit 3) is related to brain-derive neurotrophic factor pathway.
NLGN4X	1.84	NLGN4X (Neuroligin 4 X-Linked) is related to cell adhesion molecules pathway.
DISP2	1.77	DISP2 (Dispatched RND Transporter Family Member 2) is related pathways are CDK-mediated phosphorylation and removal of Cdc6 and Signalling by GPCR.
IGFN1	1.68	IGFN1 (Immunoglobulin Like And Fibronectin Type III Domain Containing 1) associated with asthma by GWAS.
CLIP3	1.68	CLIP3 (CAP-Gly Domain Containing Linker Protein 3) is a Protein Coding gene. Among its related pathways are Signalling by GPCR and TNF signalling.
KISS1	1.65	KISS1 (KiSS-1 Metastasis Suppressor) may regulate events downstream of cell-matrix adhesion, perhaps involving cytoskeletal reorganization.

Table 5-5 Top 20 up-regulation genes by siRNA **b** versus scrambled after 48h ranked by adjusted p-value firstly and then by logFC. Adjusted

p-value <0.05 and LogFC >1.0. Gene functions or biology were descripted from GeneCards (187) and Entrez Gene (188).

Among the 1460 down-regulated differentially expressed genes in siRNA b vs scrambled, *NPNT* was the most downregulated gene found by RNA-seq, which means the high efficiency of *NPNT* knocked down by siRNA b.

SERPINE1 (Serpin Family E Member 1, logFC -1.36) plays a role in respiratory diseases, such as COPD (214), and is involved in cell adhesion and ECM remodelling (126). MMP10 (Matrix Metallopeptidase 10, logFC -1.08) is involved in ECM breakdown, and was found to be downregulated after *NPNT* knockout. Thus, SERPINE1 and MMP10 might implicate in epithelial cell homeostasis and associate with NPNT expression.

SYMBOL	**logFC**	**Gene function**
NPNT	-3.19	Nephronectin
KIAA1456	-1.88	C8orf79, TRMT9B (TRNA Methyltransferase 9B (Putative)) which involves tRNA processing and gene expression, Acts as a tumour suppressor by promoting the expression of LIN9.
LRRC37A17P	-1.68	Pseudogene
TRMT10A	-1.41	TRMT10A (TRNA Methyltransferase 10A) is a protein coding gene. Among its related pathways are tRNA processing and Gene Expression.
SERPINE1	-1.36	SERPINE1 (Serpin Family E Member 1) involves in G-protein signalling Ras family GTPases in kinase cascades (scheme) and Cell adhesion ECM remodelling. As PLAU

		inhibitor, it is involved in the regulation of cell adhesion and spreading (PubMed:9175705). Acts as a regulator of cell migration, independently of its role as protease inhibitor. Plays a role in alveolar type 2 cells senescence in the lung.
RGS5	-1.34	RGS5 (Regulator of G Protein Signaling 5) involves in peptide ligand-binding receptors and signalling by GPCR.
MYBL2	-1.33	MYBL2 (MYB Proto-Oncogene Like 2) is a protein coding gene. Among its related pathways are Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors and EGF/EGFR Signaling Pathway. Transcription factor involved in the regulation of cell survival, proliferation, and differentiation.
C2orf47	-1.32	Also called MAIP1 (Matrix AAA Peptidase Interacting Protein 1) which promotes sorting of SMDT1/EMRE in mitochondria by ensuring its maturation.
RP11-46A10.5	-1.30	lncRNA
FAM35A	-1.28	Also called SHLD2 (Shieldin Complex Subunit 2). Component of the shieldin complex, which plays an important role in repair of DNA double-stranded breaks (DSBs). During G1 and S phase of the cell cycle, the complex functions downstream of TP53BP1 to promote non-homologous end joining (NHEJ) and suppress DNA end resection
AGMO	-1.27	AGMO (Alkylglycerol Monooxygenase). Among its related pathways are Fatty Acyl-CoA Biosynthesis and Metabolism.
RP11-337C18.8	-1.22	lncRNA
TRIML2	-1.16	TRIML2 (Tripartite Motif Family Like 2) enables protein binding.
SDHD	-1.13	SDHD (Succinate Dehydrogenase Complex Subunit D). Among its related pathways are Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins. and Metabolism.
CUX2	-1.11	CUX2 (Cut Like Homeobox 2) involves in the control of neuronal proliferation and

		differentiation in the brain, regulates dendrite development and branching, dendritic spine formation.
MMP10	-1.08	MMP10 (Matrix Metallopeptidase 10) are involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development, reproduction, and tissue remodelling. Can degrade fibronectin, gelatins of type I, III, IV, and V; weakly collagens III, IV, and V. Activates procollagenase.
RP5-1160K1.3	-1.08	-
EFCAB6	-1.08	EFCAB6 (EF-Hand Calcium Binding Domain 6) negatively regulates the androgen receptor.
FYB	-1.06	FYB1 (FYN Binding Protein 1) is a protein coding gene. Among its related pathways are Apoptotic Pathways in Synovial Fibroblasts and Cell junction organization.
GNG11	-1.06	GNG11 (G Protein Subunit Gamma 11) is a protein coding gene. Guanine nucleotide-binding proteins (G proteins) are involved as a modulator or transducer in various transmembrane signalling systems.

Table 5-6 Top down-regulation genes by siRNA b versus scrambled after 48h ranked by adjusted p-value firstly and then by logFC. Adjusted p-value <0.05 and LogFC >1.0. Gene functions or biology were descripted from GeneCards (187) and Entrez Gene (188).

5.4.3.3.4 DEGs analysis in both siRNA a and siRNA b

As shown in Figure **5-15**, there were 442 DEGs identified common to treatment with both NPNT siRNA a and siRNA b. There were 256 and 152 DE genes up-regulated or down-regulated, respectively after 48h treatment in HBECs (Figure **5-16** and Figure **5-17**). The top 10 up-regulated and down regulated DE genes were list as follows (Table **5-7** and Table **5-8**)

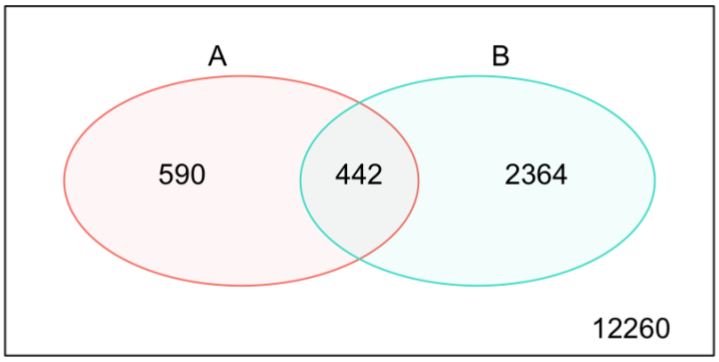


Figure 5-15 Venn diagram of total overlap differentially expressed genes in common both of siRNA a and siRNA b after 48h treatment. There are total 442 DEG overlap between siRNA a and siRNA b.

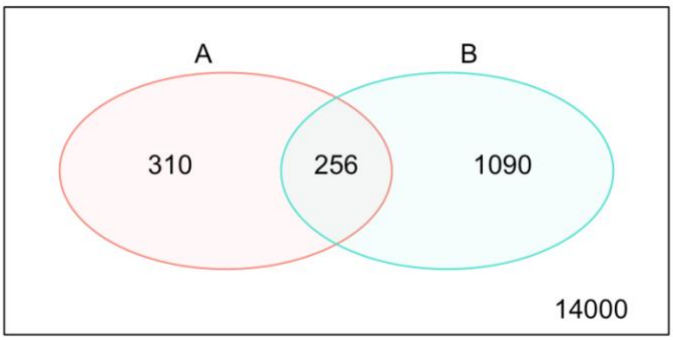


Figure 5-16 Venn diagram of up-regulated differentially expressed genes in common both of siRNA a and siRNA b after 48h treatment. There are 256 up-regulated differentially expressed genes overlap between siRNA a and siRNA b.

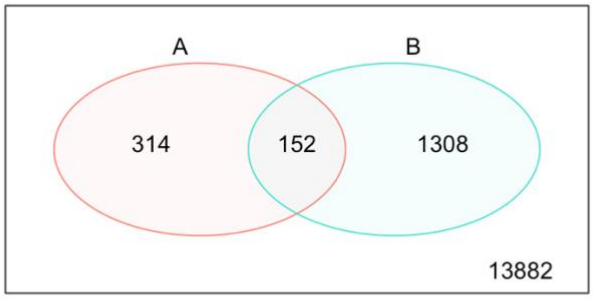


Figure 5-17 Venn diagram of down-regulated differentially expressed genes in common both of siRNA a and siRNA b after 48h treatment. There are 152 down-regulated differentially expressed genes overlap between siRNA a and siRNA b.

Gene symbol	logFC.a	logFC.b	Gene function
AC006262.5	1.49	1.17	AC006262.5 is a lncRNA associated with the risk of breast cancer.
RP11-353N4.4	1.43	1.31	lncRNA
KIAA1199	1.02	1.10	CEMIP (Cell Migration Inducing Hyaluronidase 1) is a prognostic biomarker of cancer, involves in cellular migration and invasion of lung cancer.
RP11-676M6.1	1.01	0.82	lncRNA
CBX1	0.99	1.12	CBX1 (Chromobox 1) leads to epigenetic repression, associated with chromatin regulation/acetylation.
IVL	0.94	0.81	IVL (Involucrin) involves in developmental biology.
RGS16	0.94	1.22	RGS16 (Regulator of G Protein Signaling 16) involves peptide ligand-binding receptors and Signalling by GPCR.
FAM43A	0.93	1.01	FAM43A (Family with Sequence Similarity 43 Member A) enables protein binding.
RP11-158H5.7	0.91	0.86	lncRNA
GPD1L	0.89	0.68	GPD1L (Glycerol-3-Phosphate Dehydrogenase 1 Like) plays a role in regulating cardiac sodium current.

Table 5-7 The top 10 up-regulated DE genes both of siRNA a and siRNA b comparison to scrambled. The genes ranked by adjusted p-value firstly and then by logFC siRNA a group.

Gene symbol	logFC.a	logFC.b	Gene function
NPNT	-2.60	-3.19	Nephronectin
THBS1	-1.28	-0.70	THBS1 (Thrombospondin 1) is related to pathways of Proteoglycans in cancer and Degradation of the extracellular matrix.
CDK6	-1.21	-0.61	CDK6(Cyclin dependent kinase 6) protein plays an important role in cell cycle G1 phase progression and G1/S transition, and it also has been identified in many human cancers. Cyclin dependent kinase inhibitor 3 (CDKN3) as a cyclin-

			dependent kinase inhibitor, promotes cancer cell proliferation and tumorigenesis by targeting p27. This gene has been reported to be mutated or overexpressed in some cancers.
UNGP3	-1.15	-0.94	UNGP3 (Uracil-DNA Glycosylase Pseudogene 3) is a Pseudogene.
ULBP1	-1.03	0.77	ULBP1 (UL16 Binding Protein 1) leads to activation of several signal transduction pathways, including those of JAK2, STAT5, ERK and PI3K kinase/Akt.
CCDC68	-1.00	-0.49	CCDC68 (Coiled-Coil Domain Containing 68) demonstrates a tumour-suppressive role.
SYT1	-0.96	0.66	SYT1 (Synaptotagmin 1) play a role in calcium ion binding and transport activity, uptake and actions of bacteria toxins.
FLG	-0.91	-1.04	FLG (Filaggrin) involves in developmental biology pathway.
LINC00888	-0.80	-0.81	lncRNA
PEX5L	-0.79	-0.63	PEX5L (Peroxisomal Biogenesis Factor 5 Like) involves in small GTPase binding and peroxisome matrix targeting signal-1 binding.
ANKRD46	-0.76	-0.83	ANKRD46 (Ankyrin Repeat Domain 46)

Table 5-8 The top 10 down-regulated DE genes both of siRNA a and siRNA b comparison to scrambled control. The genes ranked by adjusted p-value firstly and then by logFC in siRNA a.

5.4.3.4 Potential NPNT roles in cell proliferation and inflammation response as highlighted by GSEA pathway analysis

5.4.3.4.1 Curated dataset analysis shows NPNT might be involved in the inflammation response and cell proliferation

9 pathways were upregulated and 160 pathways were down-regulated in HBECs cultured from NPNT siRNA a treatment compared to scrambled group (FDR <5%). Table 5-9 (A and B) shows a summary of the top 9 upregulated (A) and top 20 downregulated (B) pathways which altered in NPNT siRNA a treated HBECs based on analyses using GSEA Curated database. The gene sets (pathways) demonstrated a potential role for NPNT in the interferon response pathway in cell cycle or proliferation.

(A)

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	LEADING EDGE
HECKER_IFNB1_TARGETS	94	0.77	2.02	0	0.001	tags=48%, list=7%, signal=51%
BENNETT_SYSTEMIC_LUPUS_ERYTHEMATOSUS	31	0.89	1.95	0	0.001	tags=81%, list=8%, signal=87%
MOSERLE_IFNA_RESPONSE	31	0.87	1.95	0	0.001	tags=90%, list=11%, signal=101%
BROWNE_INTERFERON_RESPONSIVE_GENES	67	0.76	1.93	0	0.002	tags=60%, list=7%, signal=64%
EINAV_INTERFERON_SIGNATURE_IN_CANCER	26	0.90	1.92	0	0.002	tags=81%, list=7%, signal=87%
STAMBOLSKY_TARGETS_OF_MUTATED_TP53_DN	46	0.78	1.86	0	0.009	tags=65%, list=12%, signal=74%
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	63	0.73	1.83	0	0.013	tags=59%, list=14%, signal=68%
TAKEDA_TARGETS_OF_NUP98_HOXA9_FUSION_3D_UP	173	0.64	1.82	0	0.017	tags=49%, list=12%, signal=56%
KOBAYASHI_EGFR_SIGNALING_24HR_UP	100	0.68	1.78	0	0.047	tags=64%, list=15%, signal=75%

(B)

Name of pathway	Size	ES	NES	FDR q-val	LEADING EDGE
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	146	-0.82	-2.24793	0	tags=79%, list=13%, signal=90%
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER	137	-0.81	-2.16478	0	tags=80%, list=13%, signal=91%
KONG_E2F3_TARGETS	96	-0.83	-2.15917	0	tags=73%, list=9%, signal=80%
FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN	185	-0.77	-2.14404	0	tags=71%, list=12%, signal=80%
KOBAYASHI_EGFR_SIGNALING_24HR_DN	244	-0.75	-2.12112	0	tags=70%, list=13%, signal=80%
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	157	-0.77	-2.11153	0	tags=74%, list=15%, signal=86%
CROONQUIST_IL6_DEPRIVATION_DN	98	-0.81	-2.10854	0	tags=82%, list=12%, signal=93%
ZHANG_TLX_TARGETS_DN	88	-0.82	-2.07531	0	tags=76%, list=12%, signal=86%
CHANG_CYCLING_GENES	143	-0.77	-2.07311	0	tags=75%, list=12%, signal=85%
FISCHER_DREAM_TARGETS	927	-0.67	-2.07055	0	tags=61%, list=15%, signal=70%
CROONQUIST_NRAS_SIGNALING_DN	72	-0.83	-2.06112	0	tags=82%, list=12%, signal=93%
PUJANA_BRCA_CENTERED_NETWORK	117	-0.78	-2.05891	0	tags=77%, list=15%, signal=90%

KANG_DOXORUBICIN_RESISTANCE_UP	53	-0.88	-2.05752	0	tags=91%, list=10%, signal=100%
ZHOU_CELL_CYCLE_GENES_IN_IR_RESPONSE_24HR	123	-0.77	-2.05	0	tags=72%, list=10%, signal=80%
WHITEFORD_PEDIATRIC_CANCER_MARKERS	115	-0.77	-2.05	0	tags=74%, list=11%, signal=83%
PUJANA_XPRSS_INT_NETWORK	167	-0.74	-2.04	0	tags=72%, list=15%, signal=85%
ZHANG_TLX_TARGETS_36HR_DN	184	-0.74	-2.04	0	tags=63%, list=14%, signal=72%
GRAHAM_CML_DIVIDING_VS_NORMAL_QUIESCENT_UP	178	-0.74	-2.04	0	tags=62%, list=13%, signal=71%
MISSIAGLIA_REGULATED_BY_METHYLATION_DN	119	-0.77	-2.04	0	tags=70%, list=10%, signal=78%
ZHAN_MULTIPLE_MYELOMA_PR_UP	44	-0.89	-2.04	0	tags=80%, list=7%, signal=85%

Table 5-9 “Curated” pathways of top 9 upregulated (A) and top 20 downregulated (B) gene sets regulated in NPNT siRNA a treated HBECs. SIZE=number of genes in the gene set after filtering out those not in the expression dataset. ES=enrichment score for the gene set. NES=normalised enrichment score. Q-value=false discovery rate – the estimated probability that the NES represents a false positive. Leading edge subset = subset of members that contribute the most to the ES. Tags= the percentage of genes contributing to the ES. List= percentage of genes in the ranked gene list before the peak in the running ES. This gives an indication of where in the list the ES is attained. Signal = the ES strength that combines the two previous statistics.

There was no pathway upregulated after treatment by siRNA b. Using the curated dataset, 81 gene sets are significantly enriched at FDR < 5% which implicate cancer-related pathways. As shown in Table 5-10, it demonstrates a summary of top 20 downregulated pathways which altered in NPNT siRNA b treated HBECs based on analyses using GSEA Curated database.

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	LEADING EDGE
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	146	-0.71	-1.9753	0	0.001157	tags=66%, list=15%, signal=78%
PUJANA_XPRSS_INT_NETWORK	167	-0.69	-1.96551	0	0.001157	tags=58%, list=15%, signal=68%
PUJANA_BRCA_CENTERED_NETWORK	117	-0.71	-1.92934	0	0.00273	tags=62%, list=15%, signal=73%
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	157	-0.68	-1.92234	0	0.002627	tags=65%, list=18%, signal=79%
PUJANA_CHEK2_PCC_NETWORK	765	-0.59	-1.89471	0	0.003992	tags=59%, list=19%, signal=72%
MILI_PSEUDOPODIA_HAPTOTAXIS_UP	503	-0.60	-1.87056	0	0.00685	tags=59%, list=16%, signal=70%
REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION	123	-0.69	-1.87008	0	0.005871	tags=69%, list=22%, signal=88%
ZHANG_TLX_TARGETS_36HR_DN	184	-0.65	-1.86356	0	0.005729	tags=67%, list=17%, signal=80%
REACTOME_PEPTIDE_CHAIN_ELONGATION	102	-0.70	-1.86135	0	0.005354	tags=71%, list=22%, signal=90%
REACTOME_TRANSLATION	165	-0.66	-1.85948	0	0.005052	tags=70%, list=22%, signal=89%
FISCHER_DREAM_TARGETS	927	-0.58	-1.85835	0	0.004911	tags=58%, list=17%, signal=69%
REACTOME_INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION	118	-0.68	-1.85116	0	0.005085	tags=69%, list=22%, signal=88%
REACTOME_CELL_CYCLE_MITOTIC	308	-0.61	-1.8435	0	0.005509	tags=59%, list=17%, signal=71%
ZHAN_MULTIPLE_MYELOMA_PR_UP	44	-0.78	-1.84015	0	0.005697	tags=77%, list=15%, signal=91%
KAUFFMANN_MELANOMA_RELAPSE_UP	60	-0.74	-1.83951	0	0.005394	tags=72%, list=19%, signal=88%
REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED	123	-0.67	-1.83436	0	0.005568	tags=71%, list=22%, signal=90%

_BY_THE_EXON_JUNCTION_COMPLEX						
JOHNSTONE_PARVB_TARGETS_3_DN	871	-0.57	-1.8336	0	0.00531	tags=58%, list=18%, signal=70%
FLORIO_NEOCORTEX_BASAL_RADIAL_GLIA_DN	185	-0.64	-1.8331	0	0.005081	tags=61%, list=14%, signal=71%
REICHERT_MITOSIS_LIN9_TARGETS	28	-0.84	-1.82463	0	0.005681	tags=82%, list=14%, signal=95%
ROSTY_CERVICAL_CANCER_PROLIFERATION_CLUSTER	137	-0.65	-1.817	0	0.006507	tags=63%, list=15%, signal=74%

Table 5-10. “Curated” pathways of top 20 downregulated gene sets regulated in NPNT siRNA b treated HBECs. SIZE=number of genes in the gene set after filtering out those not in the expression dataset. ES=enrichment score for the gene set. NES=normalised enrichment score. Q-value=false discovery rate – the estimated probability that the NES represents a false positive. Leading edge subset = subset of members that contribute the most to the ES. Tags= the percentage of genes contributing to the ES. List= percentage of genes in the ranked gene list before the peak in the running ES. This gives an indication of where in the list the ES is attained. Signal = the ES strength that combines the two previous statistics.

5.4.3.4.2 Hallmark dataset analysis shows NPNT implicates in several inflammation response pathways and cell proliferation

Using the Hallmark dataset, 11 gene sets were upregulated and 6 gene sets down-regulated at FDR < 5% after treated by siRNA a. All the 12 pathways are implicated in inflammation response, including interferon alpha/gamma, IL6-JAK-STAT3, Notch, TNF-alpha and P53 pathways. On the other hand, the down-regulated pathways are implicated in cell cycle or proliferation, such as Myc target, E2F trarget, G2M checkpoint, mitotic spindle.

(A)

NAME	SIZE	ES	NES	FDR q-val	LEADING EDGE
HALLMARK_INTERFERON_ALPHA_RESPONSE	94	0.773008	2.020891	0	tags=71%, list=12%, signal=81%
HALLMARK_INTERFERON_GAMMA_RESPONSE	197	0.614505	1.749533	0.002173	tags=57%, list=15%, signal=67%
HALLMARK_IL6_JAK_STAT3_SIGNALING	87	0.55352	1.432597	0.071168	tags=54%, list=16%, signal=64%
HALLMARK_NOTCH_SIGNALING	32	0.652325	1.421913	0.06162	tags=47%, list=14%, signal=55%
HALLMARK_TNFA_SIGNALING_VIA_NFKB	199	0.489008	1.407481	0.057988	tags=48%, list=16%, signal=57%
HALLMARK_ESTROGEN_RESPONSE_EARLY	200	0.490359	1.397269	0.053322	tags=44%, list=14%, signal=51%
HALLMARK_P53_PATHWAY	200	0.486179	1.384984	0.055325	tags=48%, list=17%, signal=58%
HALLMARK_INFLAMMATORY_RESPONSE	199	0.478686	1.359167	0.066055	tags=41%, list=15%, signal=48%
HALLMARK_KRAS_SIGNALING_DN	195	0.474041	1.351622	0.062013	tags=33%, list=14%, signal=38%
HALLMARK_GLYCOLYSIS	199	0.457868	1.310923	0.084589	tags=33%, list=12%, signal=37%
HALLMARK_HEME_METABOLISM	196	0.447023	1.283041	0.100673	tags=37%, list=14%, signal=43%

(B)

NAME	SIZE	ES	NES	FDR q-val	LEADING EDGE
HALLMARK_E2F_TARGETS	199	-0.78146	-2.17869	0	tags=69%, list=12%, signal=79%
HALLMARK_MYC_TARGETS_V1	199	-0.68346	-1.92515	0	tags=63%, list=20%, signal=79%
HALLMARK_G2M_CHECKPOINT	199	-0.65195	-1.82958	2.41E-04	tags=59%, list=13%, signal=68%
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	112	-0.59987	-1.58291	0.005214	tags=61%, list=20%, signal=76%
HALLMARK_MTORC1_SIGNALING	199	-0.51039	-1.42242	0.036704	tags=54%, list=18%, signal=65%
HALLMARK_MITOTIC_SPINDLE	199	-0.45572	-1.27278	0.157195	tags=31%, list=9%, signal=33%

Table 5-11. “Hallmark” pathways of top top 12 up-regulated (A) and 6 down-regulated (B) gene sets regulated in NPNT siRNA a treated HBECs. SIZE=number of genes in the gene set after filtering out those not in the expression dataset. ES=enrichment score for the gene set. NES=normalised enrichment score. Q-value=false discovery rate – the estimated probability that the NES represents a false positive. Leading edge subset = subset of members that contribute the most to the ES. Tags= the percentage of genes contributing to the ES. List= percentage of genes in the ranked gene list before the peak in the running ES. This gives an indication of where in the list the ES is attained. Signal = the ES strength that combines the two previous statistics.

Using the Hallmark dataset, there were no up-regulated pathways enriched after treated by siRNA b (Table 5-12). 15 downregulated pathways were significantly enriched at FDR < 25%. The pathways implicate NPNT in cell cycle or proliferation, including Myc target, E2F trarget, G2M checkpoint, mitotic spindle and DNA repair.

NAME	SIZE	ES	NES	FDR q-val	LEADING EDGE
HALLMARK_MYC_TARGETS_V1	199	-0.71252	-2.04925	0	tags=70%, list=19%, signal=86%
HALLMARK_E2F_TARGETS	199	-0.6372	-1.82995	6.19E-04	tags=63%, list=17%, signal=75%
HALLMARK_PROTEIN_SECRETION	96	-0.63367	-1.71007	0.001813	tags=61%, list=19%, signal=76%

HALLMARK_G2M_CHECKPOINT	199	-0.56742	-1.64687	0.0028	tags=54%, list=14%, signal=63%
HALLMARK_OXIDATIVE_PHOSPHORYLATION	199	-0.54668	-1.56373	0.008892	tags=53%, list=18%, signal=64%
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	112	-0.5695	-1.54274	0.008921	tags=54%, list=19%, signal=66%
HALLMARK_MTORC1_SIGNALING	199	-0.48663	-1.41755	0.03412	tags=48%, list=16%, signal=57%
HALLMARK_ANDROGEN_RESPONSE	100	-0.46927	-1.24882	0.176514	tags=39%, list=12%, signal=44%
HALLMARK_MITOTIC_SPINDLE	199	-0.43324	-1.24738	0.158449	tags=49%, list=18%, signal=60%
HALLMARK_MYC_TARGETS_V2	58	-0.49535	-1.23117	0.165651	tags=48%, list=18%, signal=59%
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	47	-0.5126	-1.21562	0.173085	tags=34%, list=13%, signal=39%
HALLMARK_FATTY_ACID_METABOLISM	158	-0.43063	-1.20326	0.177783	tags=45%, list=18%, signal=55%
HALLMARK_HYPOXIA	198	-0.40492	-1.16574	0.234815	tags=34%, list=12%, signal=38%
HALLMARK_DNA_REPAIR	143	-0.41267	-1.15767	0.234898	tags=45%, list=16%, signal=54%
HALLMARK_COMPLEMENT	196	-0.39883	-1.15107	0.233424	tags=32%, list=14%, signal=37%

Table 5-12. “Hallmark” pathways of top 15 down-regulated gene sets regulated in NPNT siRNA b treated HBECs. SIZE=number of genes in the gene set after filtering out those not in the expression dataset. ES=enrichment score for the gene set. NES=normalised enrichment score. Q-value=false discovery rate – the estimated probability that the NES represents a false positive. Leading edge subset = subset of members that contribute the most to the ES. Tags= the percentage of genes contributing to the ES. List= percentage of genes in the ranked gene list before the peak in the running ES. This gives an indication of where in the list the ES is attained. Signal = the ES strength that combines the two previous statistics.

5.4.4 Comparison of Microarray and RNA-seq platform gene expression

5.4.4.1 Absolute gene expression concordance between in the RNA-seq and Microarray platforms

The data shown in chapter 4 indicate that whilst Microarray analyses can identify genes with potentially differential regulation following treatment with siRNA targeting NPNT, there were major concerns about the quality of the data, driven largely by the failure to identify knockdown of NPNT itself in the Microarray experiments. This could potentially be due to inaccuracy in quantifying gene expression in general using Microarrays, but alternatively might reflect poor performance of the specific probes used for NPNT on the Microarray. I therefore undertook a comparison of data from the RNA-seq read counts and the fluorescence intensities of Microarray for gene expression. The scatter plot shows the relative gene expression in terms of logFC by RNA-seq and Microarray platforms when treated by siRNA a (Figure **5**-**18**), siRNA b (Figure **5**-**19**), and scrambled group (Figure **5**-**20**). Furthermore, the scatter plot of linear correlation between siRNA a and siRNA b were showed using RNA-seq (Figure 5-21) or Microarray platforms (Figure **5**-**22**).

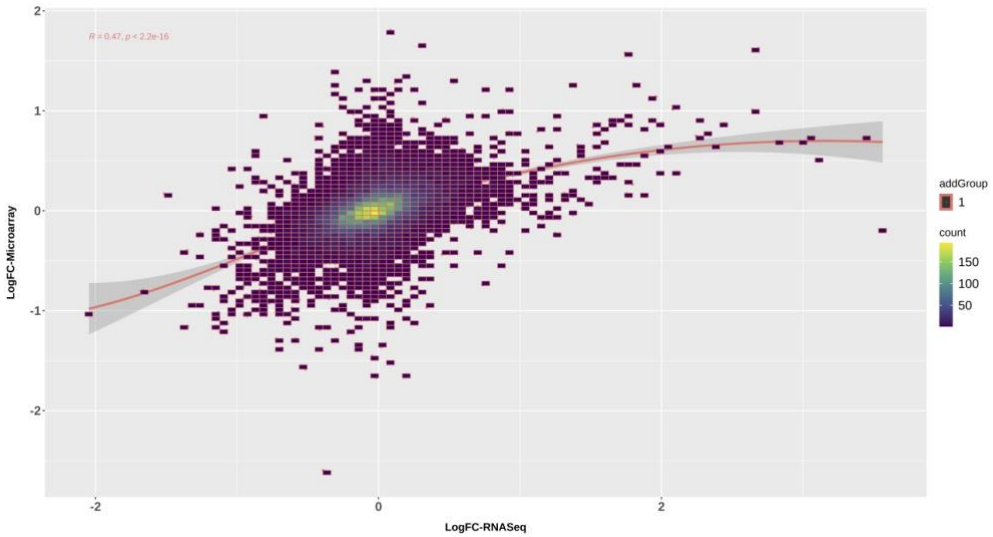


Figure 5-18 Scatter plot of linear correlation of relative gene expression in terms of logFC by RNA-seq (X axis) and Microarray platforms (Y axis) when treated by siRNA a. LogFC is computed by taking average of three samples and five samples of Microarray and RNA-seq, respectively. R=0.47, p<0.001.

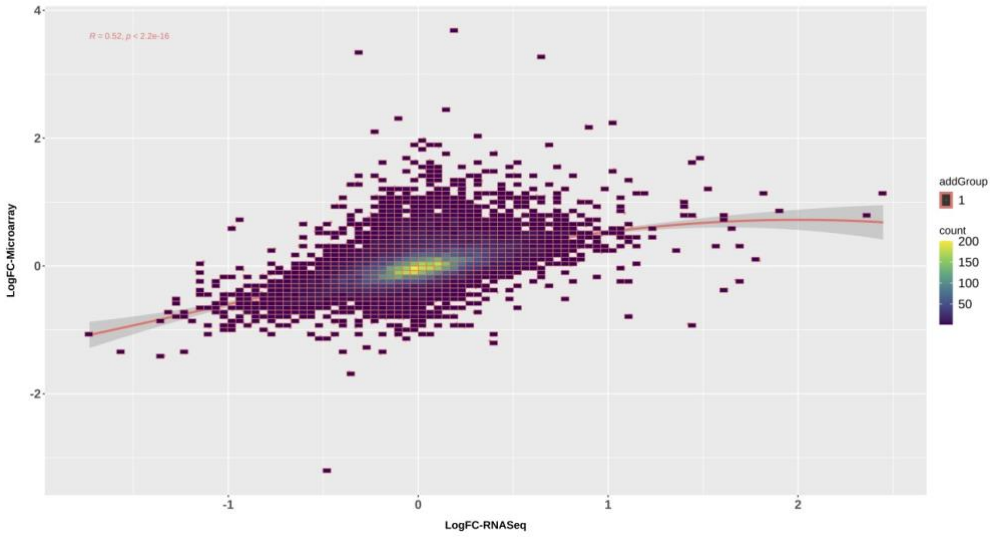


Figure 5-19 Scatter plot of linear correlation of relative gene expression in terms of logFC by RNA-seq (X axis) and Microarray platforms (Y axis) when treated by siRNA b. LogFC is computed by taking average

of three samples and five samples of Microarray and RNA-seq, respectively.R=0.52, p<0.001.

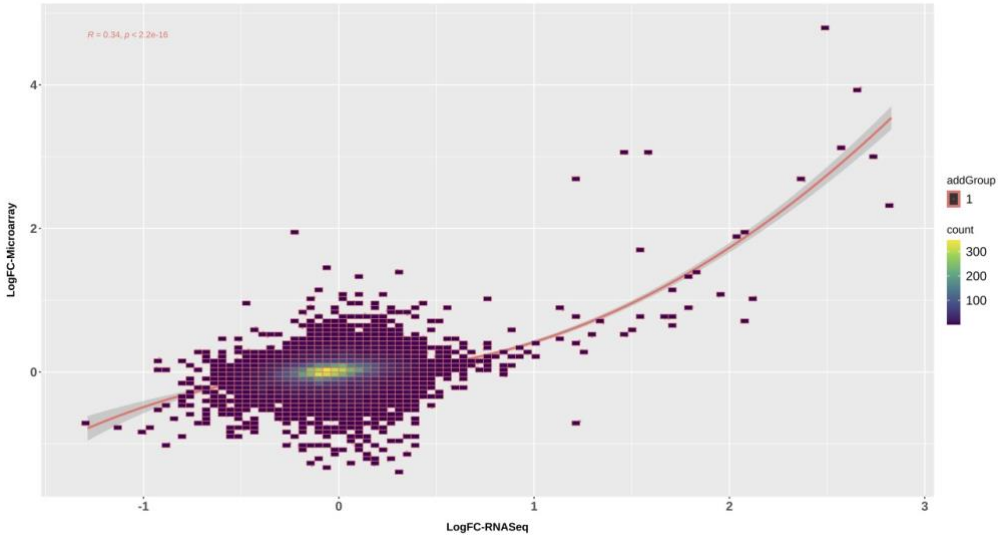


Figure 5-20 Scatter plot of linear correlation of relative gene expression in terms of logFC by RNA-seq (X axis) and Microarray platforms (Y axis) when treated by scrambled. LogFC is computed by taking average of three samples and five samples of Microarray and RNA-seq, respectively. R=0.34, p<0.001.

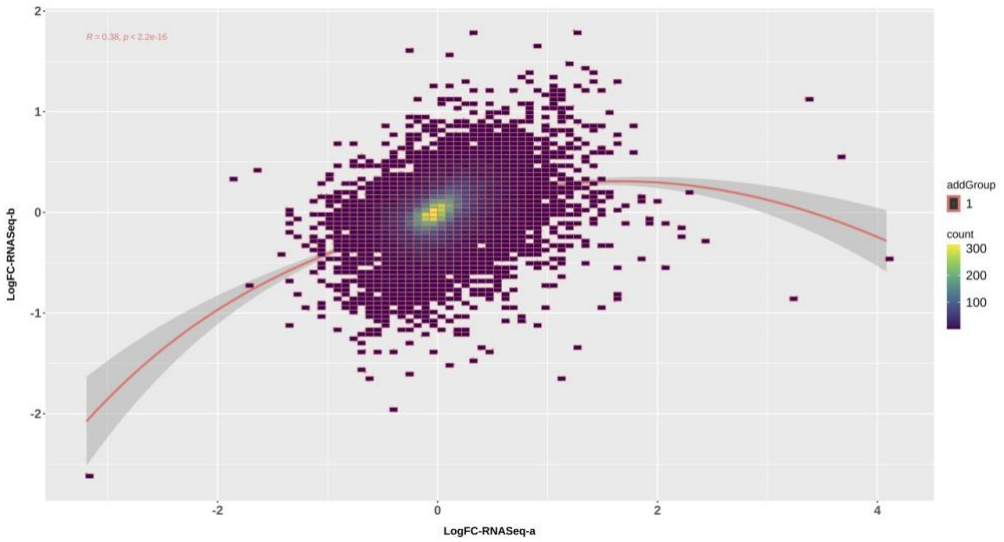


Figure 5-21 Scatter plot of linear correlation of relative gene expression in terms of logFC by RNA-seq platform when treated by anti-NPNT siRNA a (X axis) and siRNA b (Y axis). LogFC is computed by taking average of five samples of RNA-seq. R=0.38, p<0.001.

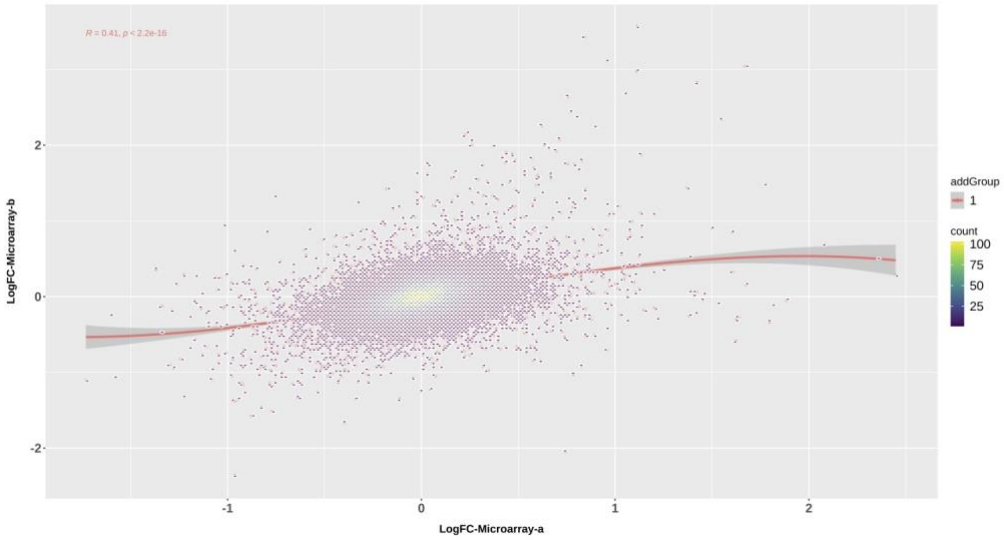


Figure 5-22 Scatter plot of linear correlation of relative gene expression in terms of logFC by Microarray platform when treated by anti-NPNT siRNA a (X axis) and siRNA b (Y axis). LogFC is computed by taking average of three samples of Microarray. R=0.41, p<0.001.

5.4.4.2 DEGs concordance between Microarray and RNA-seq platforms

The number of DEGs identified by RNA-seq platforms was significantly larger than with Microarray platforms (Table **5**-**13**).

	Platform	**Genes**	**DEG total**	**up-regulation**	**down-regulation**
siRNA a	Microarray	26,376	285	153	132
	RNA-seq	25,723	1022	556	466
siRNA b	Microarray	26,376	642	343	299
	RNA-seq	25,723	2806	1460	1346

Table 5-13 Summary of number of DEGs from Microarray and RNA-seq platforms. The HBECs were treated by 2 types of siRNAs, including siRNA a and siRNA b.

There were 24 DEGs from Microarray and 442 DEGs from RNA-seq, and 5 DEGs overlap between Microarray and RNA-seq (Figure 5-23). The VANGL2 and CBX1 were up-regulated, CDK6 and CCDC68 were down-regulated, while PAM was inconsistently regulated between siRNA a and siRNA b (Table 5-14). The details about fold change/log2FC and adjusted p-value showed in Table 5-14, and the gene location, functions showed in Table 5-15.

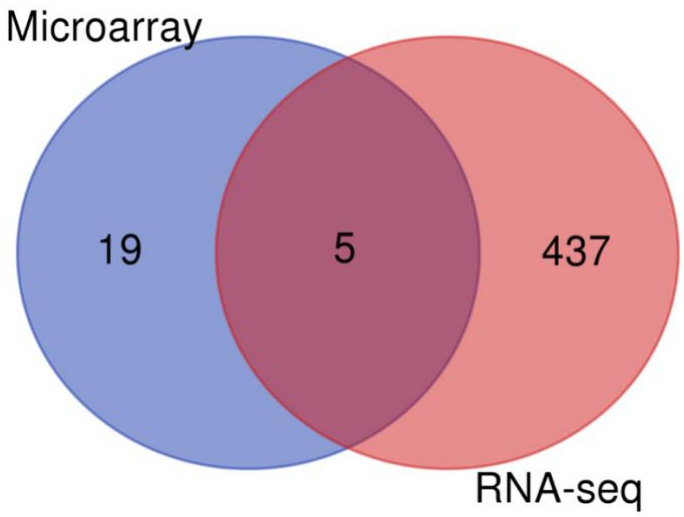


Figure 5-23 Venn diagram overlap of Microarray and RNA-seq DEGs in common. There were 24 DEGs from Microarray and 442 DEGs from RNA-seq, and 5 DEGs overlap between Microarray and RNA-seq.

	RNA-seq				Microarray			
Group	**siRNA a**		**siRNA b**		**siRNA a**		**siRNA b**	
Gene Symbol	logFC	adj.P.Val	logFC	adj.P.Val	logFC	adj.P.Val	logFC	adj.P.Val
PAM	0.35	0.0188	-0.49	0.0005	0.62	0.0110	-0.60	0.0058
CDK6	-1.21	0.0000	-0.61	0.0001	-1.12	0.0012	-0.64	0.0117
VANGL2	0.26	0.0109	0.52	0.0000	1.10	0.0012	0.98	0.0013
CBX1	0.99	0.0000	1.12	0.0000	0.93	0.0036	1.14	0.0005
CCDC68	-1.00	0.0000	-0.49	0.0023	-1.03	0.0098	-0.84	0.0122

Table 5-14 The logFC and adjusted p-value of overlap DEGs in common between Microarray and RNA-seq. Only PAM was inconsistently regulated in siRNA a and siRNA b groups. FC, fold change. Adj.p.val, adjusted p-value.

Gene Symbol	**Full name**	**Location**	**Gene function/lung biology**
CDK6	cyclin dependent kinase 6	7q21.2	Cyclin dependent kinase 6 (CDK6) protein plays an important role in cell cycle G1 phase progression and G1/S transition, and it also has been identified in many human cancers. Cyclin dependent kinase inhibitor 3 (CDKN3) as a cyclin-dependent kinase inhibitor, promotes cancer cell proliferation and tumorigenesis by targeting p27. This gene has been reported to be mutated or overexpressed in some cancers.
VANGL2	VANGL planar cell polarity protein 2	1q23.2	Involved in the control of early morphogenesis and patterning of both axial midline structures and the development of neural plate. Plays a role in the regulation of planar cell polarity.

CBX1	Chromobox 1	17q21.32	Component of heterochromatin, the protein is enriched in the heterochromatin and associated with centromeres.
CCDC68	coiled-coil domain containing 68	18q21.2	as a tumour-suppressive role in pancreatic ductal adenocarcinoma, novel causal gene of schizophrenia by GWAS.

Table 5-15 The gene location and functions of overlap DEGs in common between Microarray and RNA-seq. The VANGL2 and CBX1 were up-regulated, CDK6 and CCDC68 were down-regulated, and PAM was inconsistently regulated by siRNA a and siRNA b, and deleted in this table.

5.4.5 The role of NPNT in proliferation and migration of HBECs

5.4.5.1 siRNA NPNT knockdown does not change the cell proliferation in HBECs assessed by MTT assay

Given the data suggesting a potential role for NPNT in proliferative responses, I undertook some functional experiments to explore the ability of siRNAs targeting NPNT to alter proliferative signalling. I performed an MTT assay on HBECs treated with siRNAs (anti-NPNT siRNA a and b) to assess effects of NPNT knockdown on cell proliferation. As shown in Figure **5-24**, there is no significantly statistic difference between siRNA a or siRNA b compared with scrambled group. The p value of comparison of siRNA b to untransfected group was 0.0537.

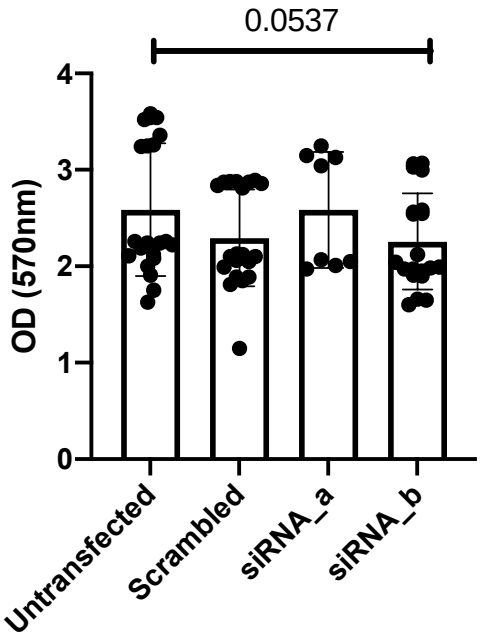


Figure 5-24 HBECs proliferation tested by MTT assay (n=5 experiments, the plots showed the replicates in every experiment). The

p value of siRNA b compared with untransfected group was the smallest (p=0.0537) among these 4 groups.

5.4.5.2 siRNA NPNT do not alter the rate of scratch wound repair in HBECs

I also performed the scratch wound repair assay on HBECs with siRNAs (anti-NPNT siRNA a and b) to assess effects of NPNT knockdown on cell migration. NPNT knockdown by siRNA b did not significantly alter the rate of scratch wound repair (Y axis) in human bronchial epithelial cells (Figure **5**-**25**).

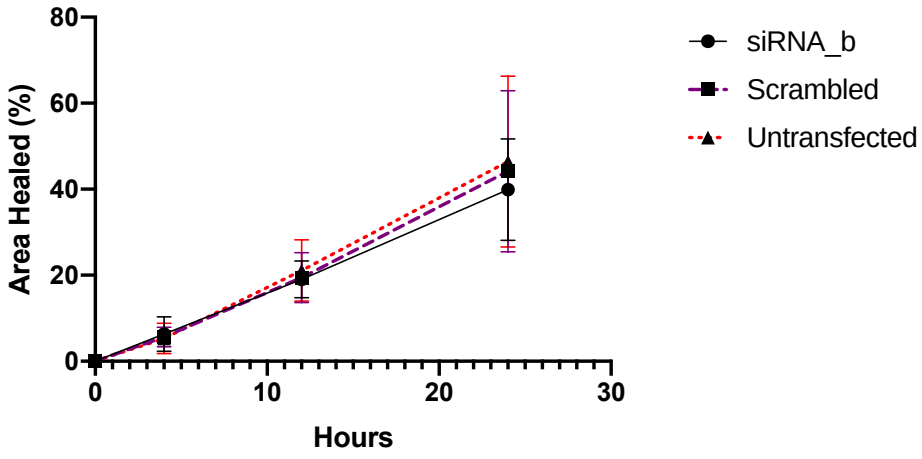


Figure 5-25 The effect of siRNA b on the rate of human bronchial epithelial cell migration. The area healed percentage comparison of siRNA b and Scrambled group was no significant difference at 24h (p >0.05). The data from 5 individual experiment carried out in one donor.

5.5 Discussion

In this chapter, the main aims were to compare the differentially expressed genes and gene sets (pathways) following *NPNT* knockdown in HBECs by siRNA using RNA-seq, and to identify novel potential biological roles of *NPNT* in human bronchial epithelial cells homeostasis.

From RNA sequencing, there were 442 overlap DEGs could be detected between the *NPNT* siRNA a and siRNA b groups in common, including 256 DEGs up-regulated and 152 DEGs down-regulated after 48h treatment in HBECs. Interestingly, the RNA-seq DEGs of siRNA compared to scrambled control were significantly higher than Microarray (in chapter 4). There were 556 up- and 466 down-regulated DEGs in siRNA a, while 1460 up- and 1346 down-regulated in siRNA b when compared to scrambled control. The different number of DEGs between siRNA a and siRNA b might be because of the different type of siRNA, different knockdown efficiency and the potential cultured conditions. The main function of siRNA is to regulate gene expression and the different types of siRNA have different sequences which and can recognise, cleaved and destroyed the different complementary dsRNA.

In addition, there were 5 DEGs in common between 24 DEGs from Microarray and 442 DEGs from RNA-seq data analysis. VANGL2 and CBX1 were up-regulated, CDK6 and CCDC68 were down-regulated, while PAM was inconsistently regulated between siRNA a and siRNA b.

Furthermore, I found that SERPINE1 was down-regulated in RNA-seq and siRNA b in Microarray but not siRNA a in Microarray. Thus, *SERPINE1* might be implicated in the downstream of NPNT siRNA treatment in HBECs.

Of these 5 DEGs between RNA-seq and Microarray, Vang-like 2 (VANGL2) is a four-pass transmembrane protein and involves in the embryonic development (215). VANGL protein regulates tumour cell migration and invasion (216). Studies have reported that total VANGL2 protein levels in vitro and zebrafish have been reduced following integrin αv knockdown and also regulated by fibronectin ECM substrate not vitronectin (217), and integrin αv interacts with *VANGL2* to regulate matrix metalloprroteinase activity and cell adhesion (218). Furthermore, I found that *VANGL2* is significantly upregulated in *NPNT* knockdown epithelial cells, which means ECM nephronectin protein could potentially regulate the expression of VANGL2.

Chromobox (CBX) 1 is a member of the CBX family which control the transcription of target genes via chromatic modification. Xie et al reported CBX1 and CBX3 are associated with overall survival in non-small-cell lung cancer (219). Furthermore, CBX1 mRNA and protein are highly expressed in lung tissue in the proteinatlas website (https://www.proteinatlas.org/). CBX1 is upregulated when NPNT expression is decreased in HBECs. Thus, the role of CBX1 in airway relevant cells need further study.

Cyclin-dependent protein kinase (CDK) 6 is a protein kinase, which plays an important regulator role to promote G1- S transition and mediates the monophosphorylation of Rb in G1 (220). Furthermore, CDK6 is a biomarker and therapeutic target for breast cancer (221). The CDK inhibitors, such as $p21^{Cip1}$ and $p16^{Ink4a}$, regulate of CDK activity and specifically block the transition from G1 to S phase in cell cycle. Tsuji and colleagues reported that $p21^{Cip1}$ and $p16^{Ink4a}$, higher expressed in emphysema when compared to non-symptomatic smoker or non-smokers in type II alveolar cells of lung tissues (222). On the other hand, another down-regulated gene, coiled-coil domain containing (CCDC) 68, is highly expressed in lung tissue (https://www.proteinatlas.org/). CCDC68 is a tumour suppressor in pancreatic adenocarcinoma (223), involved in assembly mechanism of centriole subdistal appendages and microtubule anchoring (224). All the evidence suggests that CDK6 and CCDC68 are involved in cell cycle regulation in epithelial cells of human.

Long non-coding RNA (lncRNA) have been identified to play important roles in the biological and pathological processes of COPD or emphysema. Interestingly, I found several lncRNA were regulated in HBECs using siRNA NPNT treatment analysed by RNA-seq platform. For example, ENST00000597550.1 (CTD-2245F17.3) was identified and was validated in PBMCs from COPD subjects compared with smokers (225), also detected in HBECs when loss of function of NPNT in this chapter. ENST00000597550.1 (CTD-2245F17.3) was predicted to regulate SIGLEC14 expression, which could be detected as a

potentially serum biomarker for exacerbation susceptibility of COPD (226).

NPNT overexpression has been shown recently to increase proliferation in dental epithelial cells (124). NPNT is also involved in kidney development as confirmed by the data previously described on NPNT knockdown mice (103). Furthermore, in a recent study of human, the loss function of NPNT mutation was shown to be associated with an autosomal recessive form of bilateral renal agenesis (227).

Taken together, these differentially expressed genes are involved in the cell cycle, cell proliferation, and migration. This evidence potentially provides insight into the novel function of NPNT gene and protein in human bronchial epithelial cells as well as COPD.

GSEA pathway analysis was carried out using Curated and Hallmark datasets. The most upregulated pathways were the interferon (IFN) signalling, suggesting involvement in inflammatory process. IL6-JAK-STAT3, Notch, TNF-alpha, and P53 pathway were also upregulated which might represent an inflammatory or airway remodelling phenotype. The downregulation gene sets consisted of Mcy targets, E2F/E2F3, protein secretion, G1/S and G2/M cell cycle, senescence TP53 target, and mitotic spindle, which demonstrate NPNT implicated in regulation of cell proliferation, division, senescence and tumorigeneses.

IFNs are involved in the development of COPD via the release of inflammatory mediators, and they are expressed in specific airway cell

types, such as epithelial cells, alveolar macrophages, plasmacytoid dendritic cells (228, 229).Type-I and -III IFN produced during viral infections trigger innate immune responses of asthma and COPD (230). IFNs are suppressed in COPD patients and associated with the severity levels of COPD which indicates that COPD patients have an impairment of lung innate immunity (231). In the present study, the IFN signalling pathway and many cytokines were up-regulated after knocking down NPNT, which might reveal NPNT plays a vital role in the inflammation process of chronic airway disease.

In this NPNT knockdown cell models, CDK6 and the gene sets involved in cell cycle were down-regulated, means loss of function of NPNT (by Microarray and RNA-seq) may decrease epithelial cell proliferation and induce more cell dysfunction. On the other hand, the genes implicate in inflammatory response were upregulated. Loss of function of NPNT, appears to play a role in inducing inflammatory responses in HBECs.

The main aim of this chapter was to determine whether Microarray and RNA-seq would provide unbiased insight into biological function in NPNT knockdown HBECs. The positive correlation results confirmed there are some consistent effects seen with these two platforms of Microarray and RNA-seq platforms when knockdown *NPNT* expression using siRNAs in HBECs. On the other hand, RNA-seq offers more coverage than Microarray. The number of DEGs and affected pathways of RNA-seq platform were significantly larger than Microarray platform. The result shown in this chapter are consistent with studies which

confirmed the higher sensitivity and better dynamic range observed with RNA-seq (232, 233). Interestingly, comparison of Microarray and RNA-seq downstream gene expression and pathways revealed that both platforms identified similar biological functions following NPNT knockdown.

The limitations of this chapter are firstly, the observed differences in the results of RNA-seq and Microarray might be due to use of different donors and numbers of replicates, and the DEGs detected by RNA-seq/ Microarray have not been validated by qRT-PCR. Furthermore, the data analysis of GSEA and DEGs showed that NPNT may be involved in cell proliferation and cell cycle control, while there is no significantly regulation of cell proliferation by MTT and cell migration when assessed by scratch wound healing assay following NPNT knockdown. This is a relatively insensitive method to address cell proliferation and migration and it is possible other methods, such as Ki-67, Transwell assay, might be more sensitive.

Moreover, the efficiency of RNA interference (RNAi) is about 80%~85% in this study, and it may be higher levels of knockdown are needed to fully assess the role of NPNT at a functional level. This could potentially be achieved using CRISPR/Cas9 (clustered regularly interspaced short palindromic repeats/CRISPR-associated), which is a novel powerful method for genome engineering (234). In a recent study, Rapiteanu and colleagues have found CRISPR/Cas9 RNP-based system could achieve highly efficient, cloning- and selection-free, genome editing of

primary HBECs (138). In the next chapter I studied knockdown of NPNT in HBECs by CRISPR-Cas9 to try and provide novel insights into the potential role of NPNT in COPD pathogenesis.

6. Investigation of NPNT downstream pathways using electroporation based CRISPR/Cas9 NPNT knockout in primary human bronchial epithelial cells

6.1 Introduction

As discussed in the previous chapters, numerous GWAS have identified NPNT variants which are associated with lung function variables, including FEV_1 and FVC (10, 44). In chapter 3, I have described the expression of NPNT in human airway structural cells (e.g., HBECs, HASMs) and in a wide range of human tissues, as well as suggesting NPNT might be implicated in human lung development. Furthermore, in chapter 4 and chapter 5, I have demonstrated downstream gene expression changes and suggested possible roles of NPNT in response to inflammation and cell proliferation by Microarray and RNA sequencing in HBECs following NPNT knock down using siRNA. In this chapter, I aim to develop a feasible genome editing method in primary HBECs using CRISPR/Cas9 to further reveal potential biological mechanisms implicated in COPD or cell homeostasis.

As mentioned in chapter 4 and chapter 5, RNA interference (RNAi) has been widely utilised to knockdown (KD) in cells with reduction of gene expression (235). Although RNAi has proven to be a highly useful technology, the unspecific effects “off-target” is a problem (236). The siRNA may not cause complete KD of the target gene which may pose a different phenotype to a complete knockout (KO), which may be desirable.

A target gene can be KO through gene editing, depending on spontaneous homologous recombination (HR). Researchers have

developed nucleases which can recognise a DNA single-strand or double strand break (SSB or DSB, respectively) through binding to specific genome loci (237). These three widely used engineered nucleases are a useful tool for genome editing. These are Zinc finger nucleases (ZFN), transcription activator-like effector nucleases (TALEEN) and RNA-guided endonucleases (RGENs) such as the clustered regulatory interspaced short palindromic repeat/CRISPR-associated protein 9 (CRISPR/Cas9) system (238). The CRISPR/Cas9 system originated from the adaptive immune system in prokaryotes, preventing viruses and plasmids invading (137). This system is formed from a Cas9 endonuclease and an RNA dimer (guide RNA, gRNA) containing a mature CRISPR RNA (crRNA) and transactivating crRNA (tracrRNA). The tracrRNA is required to form a complex with the Cas9 nuclease enzyme, and the crRNA sequence links the gRNA and tracrRNA (239, 240). The RNA complex guides the Cas nuclease to the specific DNA sequence. The Cas9 protein is composed of the HNH domain, which cleaves the complementary strand of the target DNA, and the RuvC-like domain, which cleaves the non-complementary strand (238).

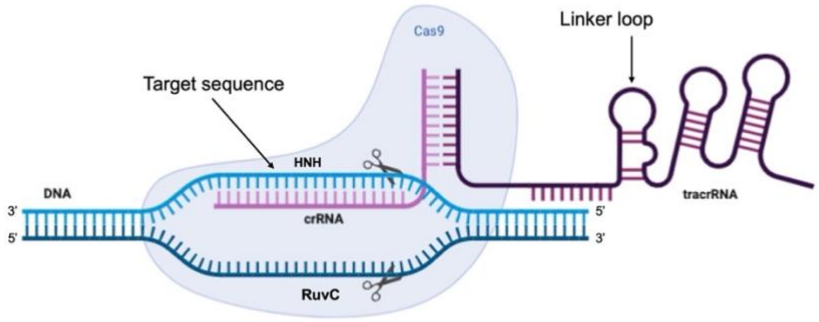


Figure 6-1 Schematic diagram of the CRISPR/Cas9 system. A single guide RNA (sgRNA) matches the DNA code where the cut is to be made. Cas9 protein and gRNA were the two parts of CRISPR/Cas9 system. gRNA is composed of the crRNA and tracrRNA. The 20 nt of the gRNA complementary to the target locus guides the Cas9 endonuclease to introduce double-strand breaks (DSB). The Cas9 nuclease contains a HNH domain and the RuvC-like domain. This figure was sourced from BioRender (https://app.biorender.com/biorender-templates) and modified in Microsoft PowerPoint.

There are many delivery methods for CRISPR/Cas9, including viral delivery methods (e.g. adeno-associated virus, full-sized adenovirus, and lentiviral), non-viral delivery methods (e.g. liposomes, polyplexes, gold particles), and physical delivery method (e.g. microinjection, electroporation) (241). However, it is challenging in primary cells originating from patients or healthy control donors to perform genome

editing experiments. Some primary cells cannot proliferate or are difficult to culture. It is reported for example that the human bronchial epithelial cells may change molecular features and function if they go through more than 4 passages (242). Thus, one solution is to directly use the bulk cell culture for experiment to investigate the biological mechanisms underlying diseases of human being.

To date, there are only a few studies that have utilised CRISPR/Cas9 to KO target genes in human primary airway structural cells, including HBECs, HASMs, and fibroblasts. In 2015, Bellec et al has provided a novel protocol with lentiviral vectors (LVV) encoding shRNA or CRISPR/Cas9 genome editing in the Calu-3 cell line and primary human airway epithelial cells to inactive cystic fibrosis transmembrane conductance regulator (CFTR) gene expression (243). And MUC18 was edited using the lentiviral-delivery of CRISPR/Cas9 genome editing method in human primary nasal airway epithelial cells (AECs) (244). Wu and colleagues have successfully established a CRISPR/Cas9 system to KO growth differentiation factor 15 (GDF15) in primary human tracheobronchial epithelial cells with a mixed cell population (245). In 2017, Voets and colleagues have described a method using adenoviral (AdV) CRISPR/Cas9 vectors for high efficiency gene inactivation in normal human lung fibroblasts and human bronchial epithelial cells (246). They demonstrated that AdV CRISPR/Cas9 was a useful and efficient tool for protein KO in these human primary cells. In 2018, Tollip gene knockout using lentivirus CRISPR/Cas9 was

described in passage 2 primary human tracheobronchial epithelial cells (247).

Recently, Rapiteanu and colleagues have developed a novel editing protocol using an electroporation-based genome editing method in primary HBECs (138) and primary human lung fibroblasts (248). They have edited the B2M gene in primary basal HBECs from two healthy donors and demonstrated a high gene editing efficiency at 94% when assessed by MiSeq genotyping. In particular, the genome editing efficiencies could be maintain throughout the 28 days ALI differentiation culture. Furthermore, they also targeted transcription factor Forkhead box J1 (FOXJ1) which has been suggested to regulate ciliation in the airway. Two types of FOXJ1 crRNA were designed and utilised in the experiment, and decreased the proportion of ciliated cells in bronchial cells. They concluded the strength of the phenotype was dependent on genome editing efficiency rather than other factors, such as CRISPR/Cas9 off-target or culture conditions (138). This was therefore the first report to describe gene knockout by CRISPR/Cas9 RNP-based highly efficient, cloning-free, genome editing in primary human bronchial epithelial cells (bulk cells) with subsequently successfully differentiation at ALI (138). Furthermore, Koh and colleagues also reported that targeting IL-13-STAT6-SAM pointed domain containing E26 transformation-specific transcription factor (SPDEF) abolished IL-13-induced MUC5AC expression and goblet cell differentiation by electroporation genome editing method (249). Given this background I

aimed to knockout NPNT expression in an HBECs model using electroporation CRISPR/Cas9 genome editing technology.

Because HBECs lose their ability to differentiate after three or four passages, a cell model with enhanced lifespan whilst retaining plasticity and phenotype of primary HBECs is need for research (160). B-cell-specific Moloney murine leukaemia virus integration site 1 (BMI-1) is a polycomb protein and plays an important role in cell senescence via downregulation of the cycle-dependent kinase inhibitor p16(Ink4a) (250). With high expression BMI-1 can suppress p16(Ink4a), a tumour suppressor that induces G1 cell cycle arrest, and delays cell senescence (251). Professor Ian Sayers and colleagues have assessed BMI-1-engineered cells (BMI-1 cells) at early (p5), mid (p10), and late (p15) passage, and confirmed BMI-1 cells delay HBEC senescence whilst retaining their plasticity and ability to differentiate at ALI (125). Furthermore, O'Loughlin and colleagues have further studied BMI-1-engineered HBECs by RNA sequencing and demonstrated that there are only a small number of differentially expressed genes (293) between BMI-1-engineered HBECs to normal HBECs (160). These differentially expressed genes were enriched in epithelial mesenchymal transition (EMT) and cell cycle pathways (160). Importantly, in our previously study (160), the expression of NPNT mRNA was not up- or down-regulated in BMI-1 cells compared to normal primary HBECs using RNA sequencing (Figure 6-2). Thus, using BMI-1-engineered HBECs can delay HBEC senescence by promoting cell cycle

progression whilst retaining the key features for CRISPR/Cas9 genome editing for studying NPNT in human bronchial epithelial cells.

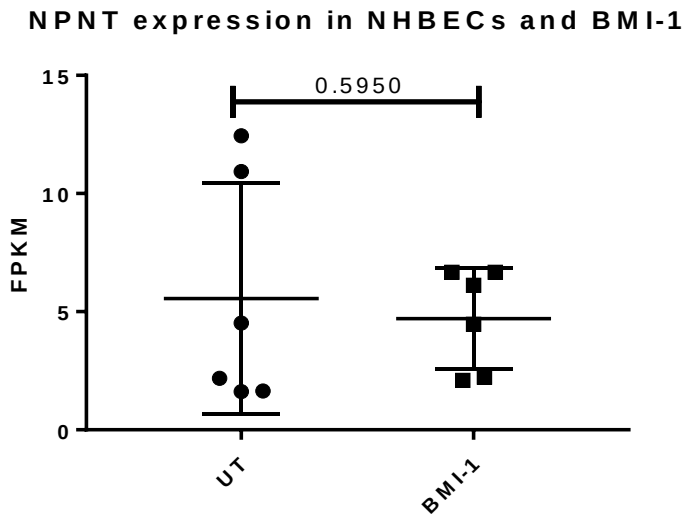


Figure 6-2 NPNT mRNA expression value in untransfected primary HBECs (UT) and BMI-1 cells by RNA-seq. Fragments per kilobase of exon per million fragments mapped (FPKM) (Data from Dr. Robert Hall, University of Nottingham) (160). NHBECs, normal human bronchial epithelial cells.

6.2 Aims

In this chapter, the specific aims were to 1) establish a feasible protocol of electroporation based CRISPR/Cas9 genome editing of NPNT in BMI-1-engineered HBECs; and 2) detect the downstream genes following NPNT knockout to provide insight into the potential biological mechanisms implicated in human bronchial epithelial cells homeostasis.

6.3 Methods

6.3.1 Cell culture

6.3.1.1 Cell culture of human bronchial epithelial cells

The BMI-1-engineered HBECs were genetically modified as described previously in Division of Respiratory Medicine, University of Nottingham (125, 160). For the work described in this chapter, the BMI-1 cells were obtained from Robert Hall (Division of Respiratory Medicine, University of Nottingham). BMI-1 cells were cultured in PneumaCult-Ex Basal Medium (Lonza) as described in chapter 2. Cells were kept in a humidified 10% CO_2 atmosphere at 37℃.

6.3.2 Gene knockout using CRISPR/Cas-9

Full details of the procedures for CRISPR/Cas9 are detailed in Chapter 2. A brief summary is included below of key details.

6.3.2.1 NPNT expression in BMI-1 cells

As discussed above, BMI-1 cells were engineered to over-express BMI-1 using viral vector systems (125). Global mRNA expression was assessed by RNA-sequencing (160) and NPNT expression was found to be not changed by BMI-1 over-expression (Figure 6-2).

6.3.2.2 CRISPR/Cas9 procedure

Three types of crRNA were designed using IDT (https://sg.idtdna.com/pages) as described in detail in chapter 2 (Table 6-1). Briefly, the 3 types of gRNA were hybridized crRNA and tracrRNA at 50μM final concentration. RNP assembly used 90pmoles of Cas9

with 450pmoles crRNA-tracrRNA hybrid. 0.9µl of electroporation enhancer was added to give a total volume of 11.40µl BMI-1 cells were cultured, typsinised, centrifuged and then resuspended with 200,000 cells in 8.6μl of P3 Primary cell nucleofector solution (Lonza, V4XP-3032). The CM-113 program on the Amaxa 4D Nucleofactor was used. The BMI-1 cells were electroporated and then incubated in 2mL of pre-warned media in incubator.

Type	Sequence	PAM	Exon	Activity	Without Off-target Predicted	cut site	GC %
crRNA 1	TCCTCCCACCATAACGACAT	AGG	2	72	99	chr4 [+105,897,936 : -105,897,936]	50
crRNA 2	CATTCACCATGTTTGCATCG	TGG	3	65	93	chr4 [+105,927,352 : -105,927,352]	45
crRNA 3	GGATATATGCTCATGCCGGA	TGG	4	62	100	chr4 [+105,937,111 : -105,937,111]	50

Table 6-1 Three types of crRNA was designed and provided from IDT as list above. The table shows details of sequence, PAM, Exon number, activity, predicted off-target, cut site, and GC%.

6.3.3 Molecular Biology Methods

6.3.3.1 RNA extraction

RNA was extracted from cell lysates using the Sigma GenElute^{TM} Mammalian Total RNA Miniprep Kit (RTN70- 1KT, Sigma Aldrich), cDNA synthesised by DNase I, Amplification Grade (18068015, Thermo Fisher Scientific) and then the RNA or cDNA samples were stored at -

80℃ freezer for RT-PCR or RNA-seq. The RNA extraction is described in detail in chapter 2.

6.3.3.2 RNA quality control

RNA was detected for quality and quantity with a Nanodrop ND-2000 full spectrum spectrophotometer and RNA integrity number (RIN) scores were measured by Agilent 4200 TapeStation as explained in Chapter 2.

6.3.3.3 TaqMan PCR

qRT-PCR was performed as previously described in chapter 2 for NPNT by NPNT PDAR (Hs01568131_m1, Thermo Fisher Scientific).

6.3.3.4 RNA sequencing

6.3.3.4.1 Library preparation and RNA-sequencing

Following RNA isolation, the next procedure was to create an RNA-seq library. This was carried out by Oxford genomics to generate paired end (150bp) reads at a depth of 30 million reads per sample on Novaseq 6000.

6.3.3.4.2 Trimming and alignment of fastq files

Processing of RNA-seq data was performed using the Unix operating system. Quality of fastq files was assessed using fastqc (252). Adaptor and quality trimming of fastq files was performed using Trimmomatic (253). Trimmed reads were aligned to the CRCh37 build using Star

producing count files that were used for differential expression analysis (145).

6.3.3.4.3 Pre-processing of count files for differential gene expression

Raw counts were converted to counts per million (CPM) and low expressed genes across all samples were filtered out of the analysis using the **filterByExpr** function in **EdgeR** (254). Normalisation of data was performed using the **calcNormFactors** in **EdgeR** using the trimmed mean of M-values (TMM) method.

6.3.3.4.4 Analysis of differentially expressed genes

Gene count values generated from fastq files in Star were used in independent linear regression models (empirical bayes method) including within subject correlation for matched samples across each experiment using *limma* package (version 3.48.0) to identify differentially expressed genes. Log-CPM mean-variance relationship was accommodated using precision weights calculated by the voom function and specific samples weights were calculated and used by the voomwithQualityWeights function. Contrasts were set up in *limma* using the *makeContrasts* function. Significant genes were classified based on a Benjamini-Hochberg false discovery rate (FDR) less than 0.05.

6.3.3.5 Bioinformatics analysis

6.3.3.5.1 Enrichment analysis in DAVID database

The Database for Annotation, Visualization, and Integrated Discovery (DAVID, https://david.ncifcrf.gov/home.jsp) provides functional annotation of large gene set for biological interpretation (255, 256). I used the DAVID Functional Annotation Tool of DAVID Bioinformatics Resources 6.8 on 25th Dec 2021. Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) databases have been used for retrieving function and pathways terms.

6.3.3.5.2 Construction of PPI

I used the search tool for the retrieval of interaction genes (STRING) database (https://string-db.org/) to predict the interaction networks among DEGs. A combination score of >0.4 was selected as the threshold. The Cytoscape version 3.7.0 was used to illustrate the relationship between each DEGs.

6.4 Results

6.4.1 BMI-1 Cells confluency

To ensure that electroporation had not produced extensive cell death, I used the Olympus CKX53 culture microscope to assess BMI-1 cells confluency after CRISPR/Cas9 procedure. As shown in Figure **6-3**, the confluency was similar after gRNAs treated or electroporated groups at 96h suggesting electroporation does not produce large effects on cell viability

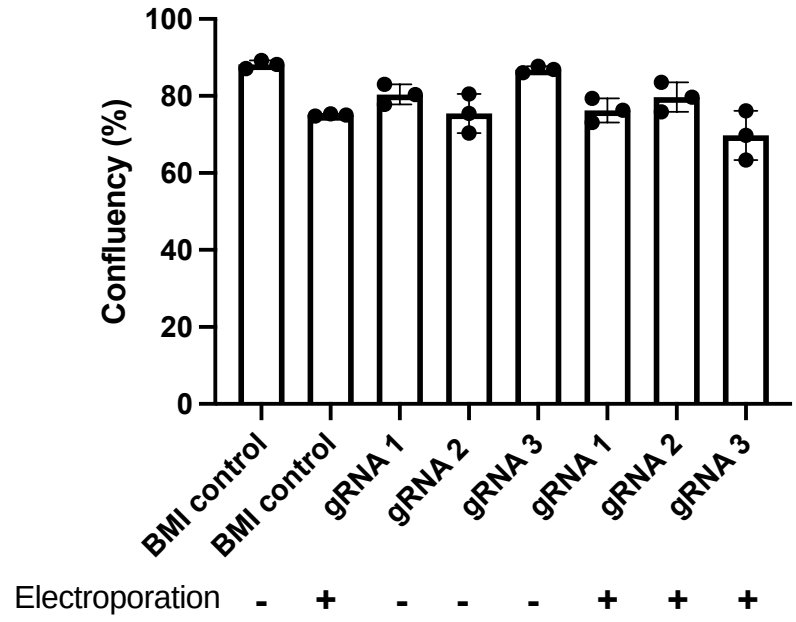


Figure 6-3 Confluency of CRISPR/Cas9 genome editing BMI-1 cells at 96h. The confluency was counted by Olympus CKX53 culture microscope (n=3). BMI, BMI-1-engineered cells.

6.4.2 Quality control of RNA samples for RNA-seq

The RNA samples were extracted and the quality control information for the RNA used for sequencing is described in Table 6-2. The RNA concentration and wavelength ratio (260/280 ratio) were detected by Nanodrop, and RIN scores were determined by Agilent 4200 TapeStation as detailed in chapter 2. All samples (RIN>8) are considered to be of sufficient quality for RNA sequencing by Oxford Genomic Centre.

Sample No.	Treatment	Con.ng/μl	260/280	RIN
E12.1	BMI control NT	156.5	2.05	10
	BMI+E	111.5	2.04	10
	gRNA 1+cas9 NT	115.4	2.09	10
	gRNA 2+cas9 NT	94.1	2.09	10
	gRNA 3+cas9 NT	125.6	2.04	10
	gRNA 1+cas9+E	94.6	2.13	10
	gRNA 2+cas9+E	109.7	2.05	10
	gRNA 3+cas9+E	68.1	2.05	10
E12.2	BMI control NT	33.0	2.02	10
	BMI+E	125.6	1.99	10
	gRNA 1+no cas9+E	36.2	2.06	10
	gRNA 2+no cas9+E	31.6	2.07	10
	gRNA 3+no cas9+E	24.7	2.06	10
	gRNA 1+cas9+E	44.6	2.05	10
	gRNA 2+cas9+E	40.1	2.00	10
	gRNA 3+cas9+E	34.6	2.02	10
E12.3	BMI control NT	117.3	2.07	>8
	BMI+E	31.5	2.24	>8
	gRNA 1+cas9+E	52.1	2.15	>8

	gRNA 2+cas9+E	39.6	2.19	>8
	gRNA 3+cas9+E	35.5	2.11	>8
E12.4	BMI control NT	176.2	2.00	>8
	BMI+Elect	101.4	1.98	>8
	gRNA 1+cas9+E	139.9	2.02	>8
	gRNA 2+cas9+E	130.7	2.05	>8
	gRNA 3+cas9+E	107.2	2.01	>8

Table 6-2 Information of RNA samples quality control for RNA-seq using CRISPR/Cas9 to knockout *NPNT* gene expression in BMI-1-engineered HBECs. BMI, BMI-1 human bronchial epithelial cells. NT, no-electroporated control. E, electroporation by Lonza Nucleofector^{TM} 4D.

6.4.3 RT-PCR analysis using CRISPR-Cas9 to knockout NPNT mRNA in HBECs

NPNT mRNA expression was analysed by RT-PCR to confirm the KO efficiency in BMI-1 cells. The NPNT mRNA expression levels were 30%, 5% and 17% with gRNA1, gRNA 2 and gRNA 3, respectively, when compared to the BMI-1 cell control group without electroporation. And the levels were 58.44%, 9.09% and 33.77% when compared to the BMI-1 cell with electroporation of gRNA1, gRNA 2 and gRNA 3, respectively.

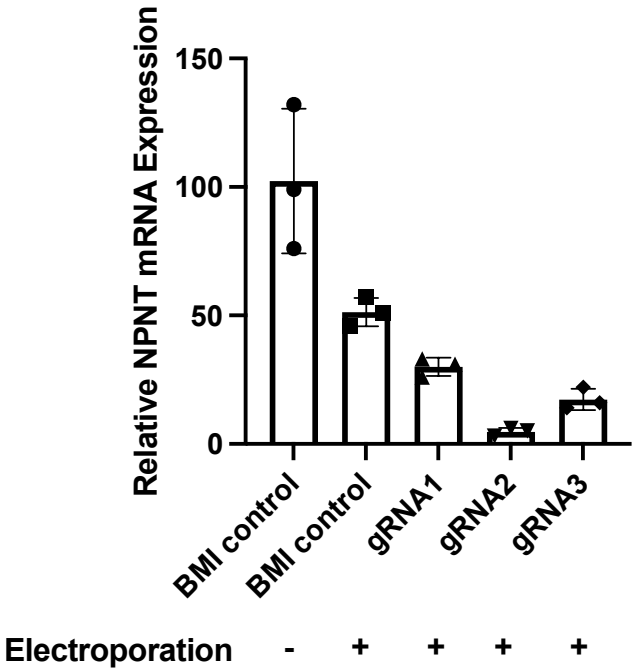


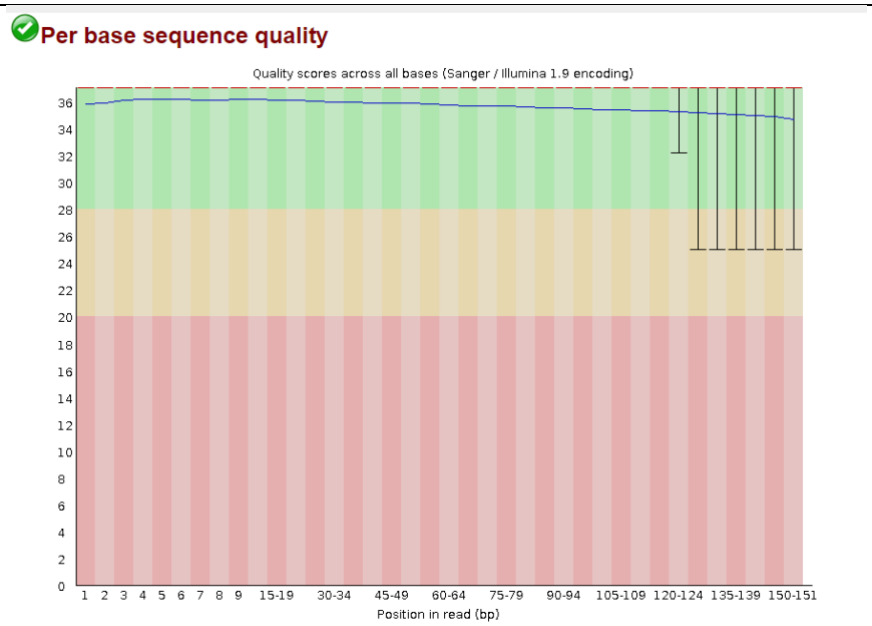
Figure 6-4 NPNT mRNA expression in electroporation based CRISPR/Cas9 genome editing system for 96h. The NPNT mRNA expression was detected by TaqMan PCR (mean±SD, n=3).

6.4.4 RNA-seq analysis using CRISPR/Cas9 to knockout NPNT in HBECs

6.4.4.1 Sequencing data quality control

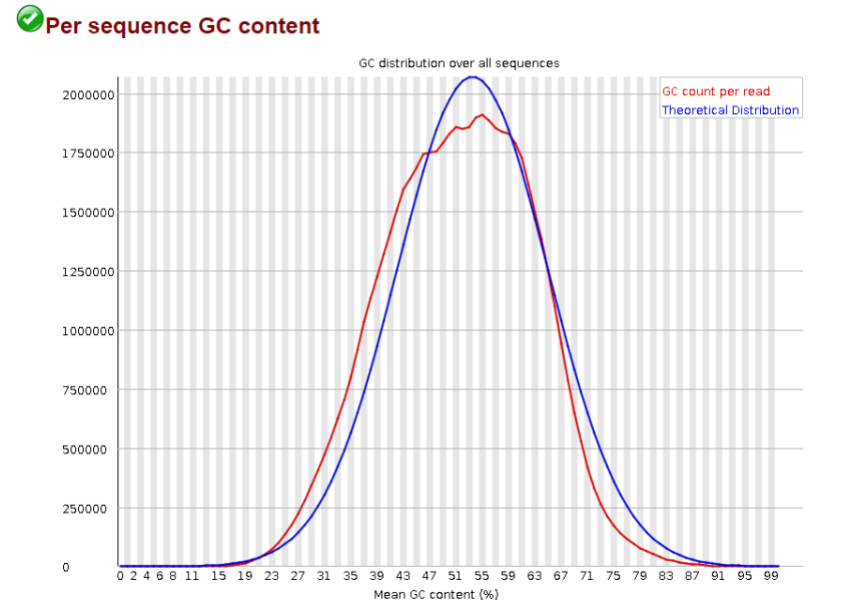
The quality of sequencing data was assessed and interpreted by FastQC tool. This tool was applied as described in Chapter 2. The outputs are shown in Figure 6-5, including a) per base sequence quality, b) per sequence GC content, c) per base N content.

a



An overview of the range of quality values across all bases at each position in the FastQ file. In the above figure a, per base sequence quality assessment showed that all reads met q score >25.

b



This detected GC content across the whole length of each sequence and compare it a normal distribution of GC content. As shown in above figure b, it is a normal distribution of the GC content.

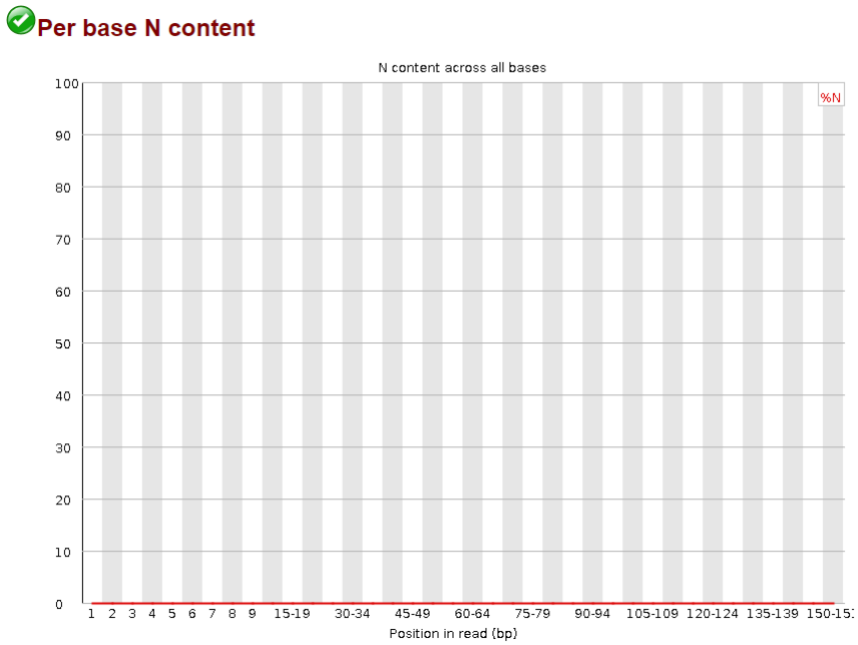
c	
	Per base N content =0 (failure ≥20) in this above figure c. N content >5% is a warning and >20% is a failure.

Figure 6-5 Results of quality control by FastQC. a) per base sequence quality assessment showed that all reads met a q score >25. b) normal distribution of the GC content. C) Per base N content =0.

6.4.4.2 Differential Gene Expression Analysis

6.4.4.2.1 Removing genes that are lowly expressed

Some genes are lowly or not expressed across all samples and were therefore filtered out of the analysis because they are unlikely to be biologically relevant. Furthermore, keeping low not expressed genes increases the number of statistical tests that are performed therefore removing them provides greater statistical power to detect differentially expressed genes (184). The ***filterByExpr*** function in the ***edgeR***

package was used to provide an automatic way to filter genes to keep as many genes as possible with worthwhile counts (Figure **6**-**6**).

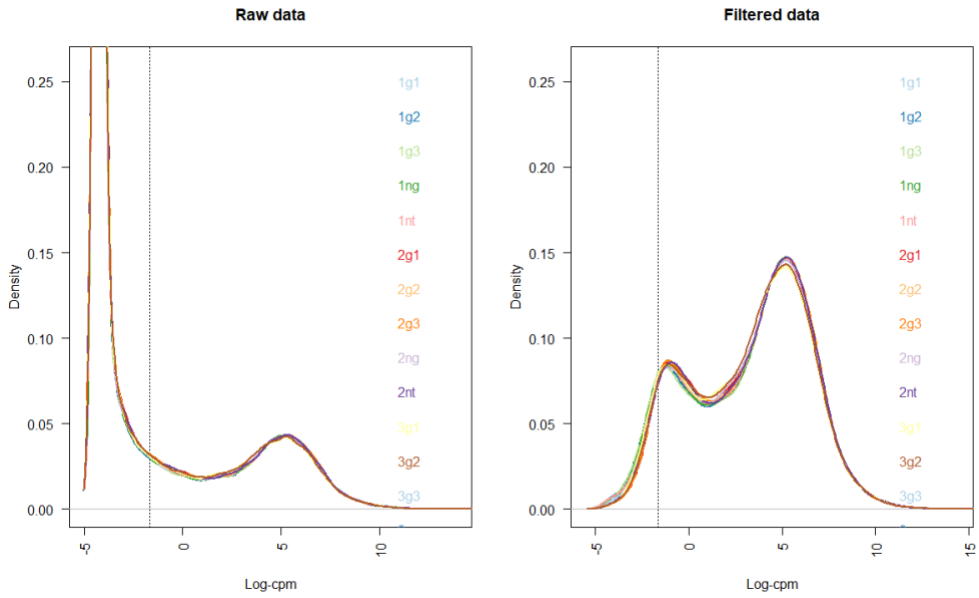


Figure 6-6 The density of log-CPM values for raw data and filtered data. CPM, counts per million.

6.4.4.2.2 Clustering of variations in replicates quality

Variations in sample quality can affect the expression of differentially expressed genes (257). In this section, the ***plotMDS*** function of ***limma*** package in R studio was utilised to assess the replicate quality of RNA-seq data sets as shown in Figure **6**-**7**.

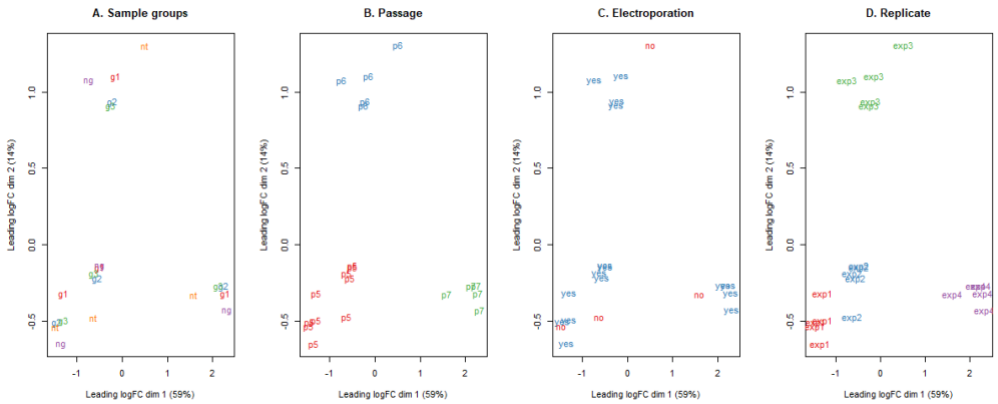


Figure 6-7 Variations of RNA-seq data sets. Multi-dimensional scaling (MDS) plots of logFC values over dimensions with samples coloured labelled by (A) sample groups, (B) passage, (C) electroporation, and (D) replicate. There was some evidence of clustering by sample group/passage number but no evidence that electroporation produced a systematic effect.

6.4.4.2.3 Applying sample specific weights improves power to detect differentially expressed genes

Variations in sample quality produce significant source of noise in RNA-seq datasets which can make detection of differentially expressed genes difficult. One way to remove this variation is to remove samples with higher variability but this reduces the power to detect DEGs. I used *voomWithQualityWeights* to assess sample-specific quality weights to be used in the voom linear model which resulted in more genes with significant p values (Figure **6-8** and Figure **6-9**).

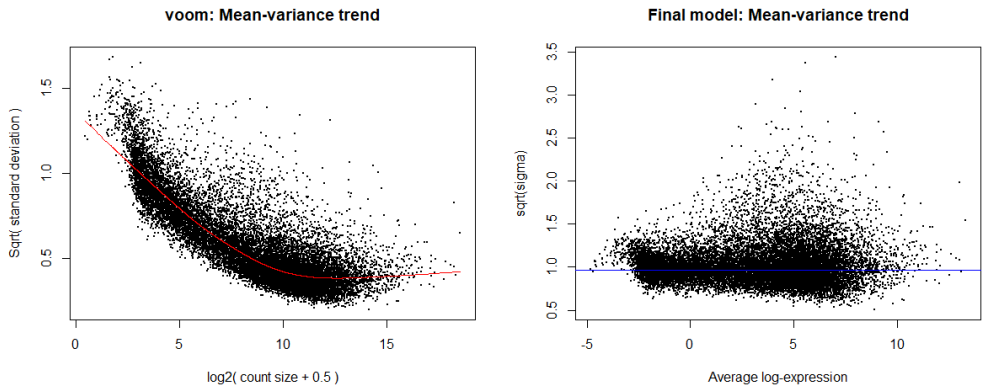


Figure 6-8 The “voom” standard model. The final model of Mean-variance trend estimated from the normalised log2CPM using the voom method.

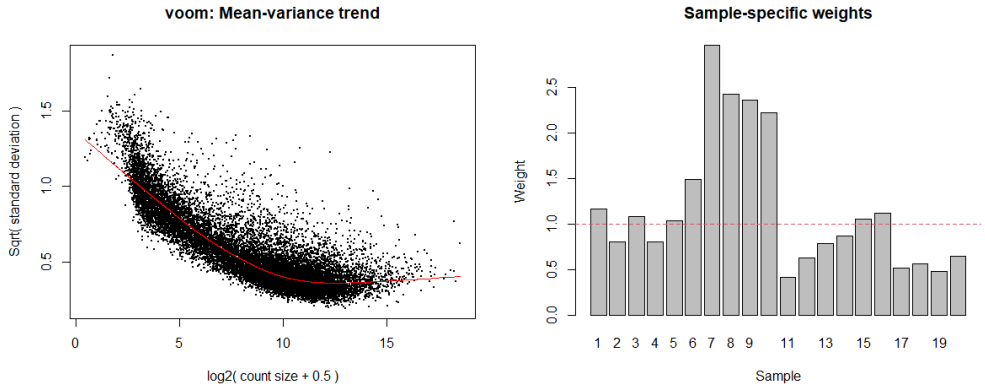


Figure 6-9 Plot showing mean-variance trend before *voomWithQualityWeights* and sample specific weights used by *voomWithQualityWeights*. The sample-specific parameters analysed from a variance model parameterised using weight strategy.

6.4.4.2.4 P value distribution by quality weights

In this section, I compared the p value distribution of data after running voom to *voomWithQualityWeights*. After incorporating sample-level weights there are far more significant p values (Figure **6**-**10**).

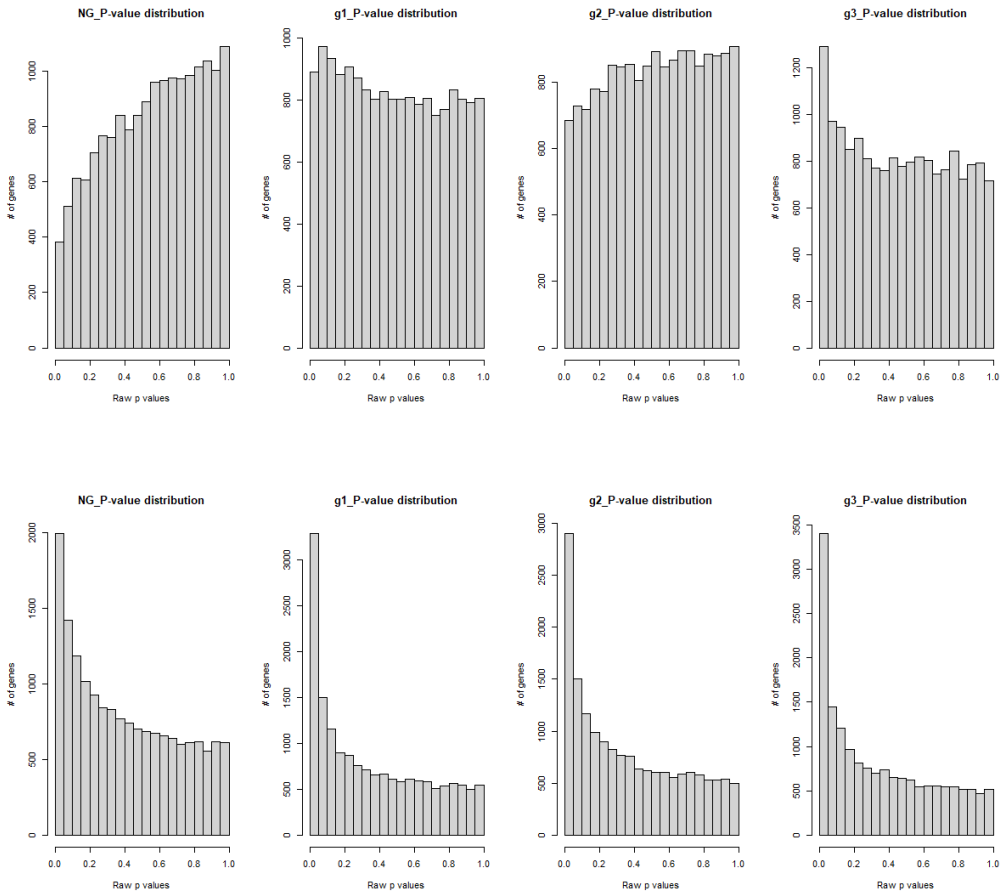


Figure 6-10 p value distribution by quality weights. Top panel shows raw p value distributions using voom alone. Bottom panel shows p value distributions after using *voomWithQualityWeights*.

6.4.4.3 Differentially expressed genes identified by RNA-seq

6.4.4.3.1 DEGs summary of KO NPNT in BMI-1 cells by CRISPR/Cas9

The DEGs were analysed in BMI-1 cells which KO NPNT gene expression by CRISPR/Cas9 and analysed by RNA-seq. The number of significantly up- or down-regulated genes are summarised in Table **6-3**. There were no DEGs between no gRNA and no electroporation groups, 22 upregulated DEGs and 27 downregulated DEGs between gRNA1 versus no gRNA, only 4 upregulated and 7 downregulated between gRNA2 and no gRNA control, and 155 upregulated and 264 downregulated DEGs between gRNA3 and no gRNA control groups.

	ngvsnt	g1vsng	g2vsng	g3vsng
Down	0	27	7	264
NotSig	16687	16638	16676	16268
Up	0	22	4	155

Table 6-3 Summary of the number of DEGs from CRISPR/Cas9 KO NPNT gene in BMI-1 cells using RNA sequencing. The significant DEGs were identified by adjusted p value <0.05 and |logFC|>0.5. ng, no gRNA treatment. g1, gRNA1; g2, gRNA2; g3, gRNA3. NotSig, not significant statistics.

Table **6**-**4** shows that NPNT could be suppressed by the CRISPR/Cas9 genome editing system and the average expression level were similar from three gRNA. However, only G3 gRNA had significantly downregulated (adjusted p value) NPNT expression compared to no gRNA treatment control group. Thus, I only analysed the DEGs compared G3 gRNA treatment to control group.

gRNA	logFC	AveExpr	Adjusted p value
G1 (crRNA1)	-1.3118	-0.6360	0.0621
G2 (crRNA2)	-0.2818	-0.6360	0.6008
G3 (crRNA3)	**-1.1851**	**-0.6360**	**0.0011**

Table 6-4 RNA-seq data of NPNT expression in BMI-1 cells which KO *NPNT* gene by different gRNA electroporation-based CRISPR/Cas9 genome editing system. logFC, log fold change; AveExpr, average expression.

6.4.4.3.2 Heatmap of 3 gRNAs KO NPNT in BMI-1 cells

The most significant DEGs are illustrated in the heatmap, which provides insight into the DEGs across individuals and samples. The top 20 DEGs were plotted by logCPM values from 3 gRNAs versus other groups using the ***heatmap.2*** function from the ***gplots*** package (Figure **6**-**11**).

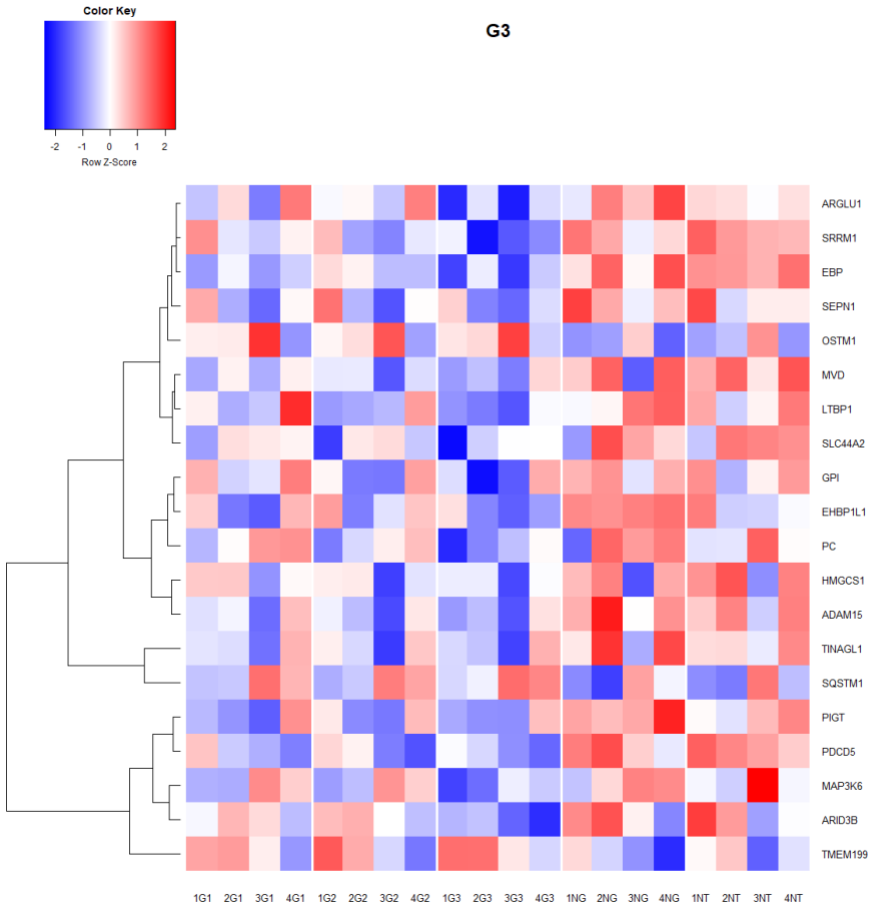


Figure 6-11 Heatmap of logCPM value for top 20 DEGs from gRNA3 (G3) compared to gRNA1 (G1), gRNA2 (G2) and control groups. The red marked samples are those with relatively high gene expression and the blue marked samples with relatively low gene expression. The variability seen is perhaps not surprising given the lack of a significant effect on NPNT expression.

6.4.4.3.3 volcano plot of G3 gRNA KO NPNT in BMI-1 cells

In the RNA-seq results, only the G3 gRNA (crRNA 3) showed significant NPNT KO and therefore I only used results from this guide RNA in the following analysis. I produced a volcano plot using the *Enhancedvolcano* package. The top up- or down-regulated 50 DEGs

(selected by logFC > 0.5 or <-0.5) are labelled in red and blue, respectively (Figure **6-12**). NPNT was significantly decreased and listed as top downregulated DEGs after KO NPNT gene by CRISPR/Cas9 technology.

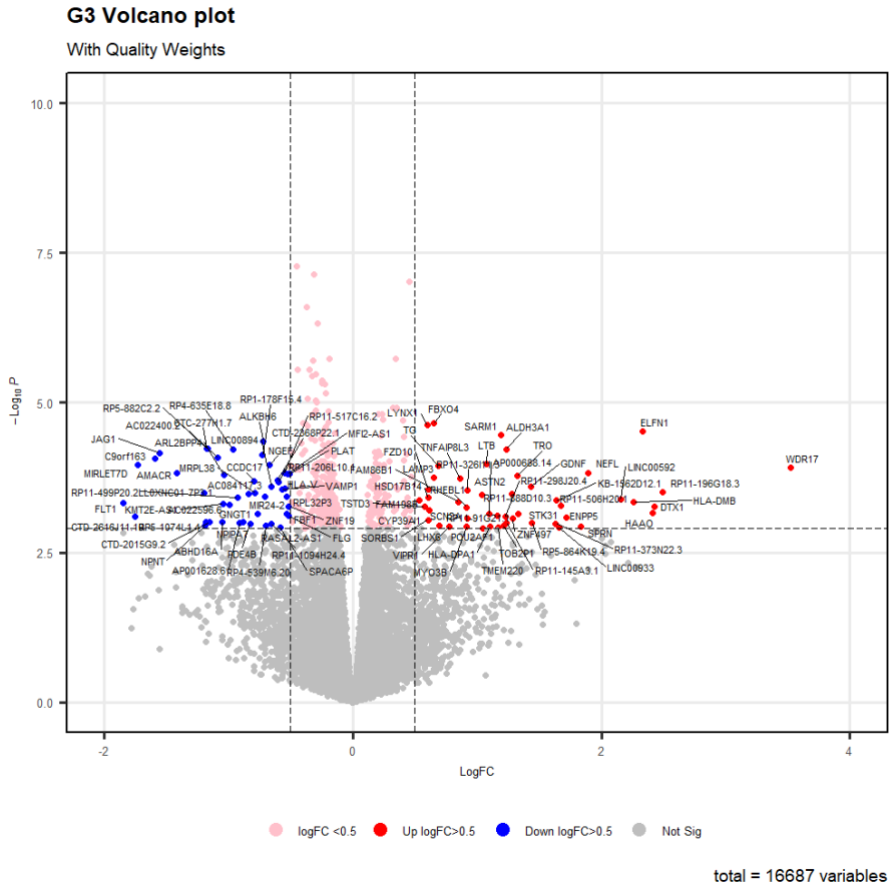


Figure 6-12 Volcano plot of gRNA3 (G3) KO NPNT expression in BMI-1 cells for 96h. Each point represents a transcript. Only significant genes in (adjusted BH p<0.05) and log FC > 0.5 or < -0.5 are labelled with gene symbol name. Up-regulated DEGs are labelled with red, down-regulated genes labelled with blue. P values were computed with adjusted BH p<0.05.

6.4.4.3.4 Top differentially expressed genes functions and biology

I ranked the top DEGs according to adjusted p value <0.05 firstly and then by logFC >1.0 or <-1.0. There were 32 DEGs up-regulated (Table **6**-**5**) and 16 DEGs down-regulated (Table **6**-**6**) in BMI-1 cells when NPNT gene knockout was performed by electroporation-based CRISPR/Cas9 genome editing system.

Gene Symbol	logFC	Gene functions and biology
WDR17	3.52	WDR17 (WD Repeat Domain 17) is associated with WDR17 include Retinitis Pigmentosa 29 and Basosquamous Carcinoma.
RP11-196G18.3	2.50	Non-coding RNA
DTX1	2.43	DTX1 (Deltex E3 Ubiquitin Ligase 1) may act as a positive regulator of Notch pathway. Involved in neurogenesis, lymphogenesis and myogenesis, and may also be involved in MZB (Marginal zone B) cell differentiation.
HAAO	2.41	HAAO (3-Hydroxyanthranilate 3,4-Dioxygenase). Gene Ontology (GO) annotations HAAO related to this gene include iron ion binding and oxygen binding.
ELFN1	2.34	ELFN1 (Extracellular leucine rich repeat and fibronectin type III domain containing 1) is broad expression in liver, kidney and brain. It is predicted to be active in ECM and extracellular space.

HLA-DMB	2.26	HLA-DMB (Major Histocompatibility Complex, Class II, DM Beta) plays a critical role in catalyzing the release of class II-associated invariant chain peptide (CLIP). Its expression in tumor epithelium is associated with increased CTL infiltration (258).
LINC00592	2.16	long intergenic non-protein coding RNA 592.
NEFL	1.89	NEFL (Neurofilament Light Chain) may also play a role in intracellular transport to axons and dendrites.
SPRN	1.84	SPRN (Shadow of Prion Protein). Gene Ontology (GO) annotations SPRN related to this gene include nucleic acid binding.
ENPP5	1.72	ENPP5 (Ectonucleotide Pyrophosphatase/Phosphodiesterase Family Member 5) encodes a type-I transmembrane glycoprotein which implicates in neuronal cell communications.
RP11-506H20.1	1.67	Non-coding RNA.
LINC00933	1.66	long intergenic non-protein coding RNA 933
KB-1562D12.1	1.63	Non-coding RNA.
RP11-373N22.3	1.63	Non-coding RNA.
RP5-864K19.4	1.44	Non-coding RNA.
GDNF	1.44	GDNF (Glial Cell Derived Neurotrophic Factor) encodes a secreted ligand of the TGF-beta (transforming growth factor-

		beta) superfamily of proteins and may be as a potential candidate gene with obstructive sleep apnoea (259) and decreased in Npnt knockout mouse model (103).
STK31	1.33	STK31 (Serine/Threonine Kinase 31) is high expressed in testis, and its upregulation is associated with gastric cancer (260) and lung cancer (261).
TRO	1.32	TRO (Trophinin) plays a vital role as cell adhesion molecule complex on the cell surface (262). Enhanced expression of TRO promotes invasive of human gallbladder cancer cells (263).
ZNF497	1.29	ZNF497 (Zinc Finger Protein 497) is involved in transcriptional regulation.
RP11-888D10.3	1.28	Non-coding RNA
RP11-145A3.1	1.24	Non-coding RNA
ALDH3A1	1.24	ALDH3A1 (Aldehyde Dehydrogenase 3 Family Member A1) is involved in the metabolism of corticosteroids, biogenic amines, neurotransmitters, and lipid peroxidation (Probable).
RP11-298J20.4	1.23	Non-coding RNA
TOB2P1	1.21	TOB2P1 (Transducer of ERBB2, 2 Pseudogene 1) is a Pseudogene.
SARM1	1.19	SARM1 (Sterile Alpha and TIR Motif Containing 1) plays a key role in axonal degeneration following injury by regulating NAD (+) metabolism.

TMEM220	1.17	TMEM220 (Transmembrane Protein 220) is a protein coding gene.
RP11-91G21.1	1.16	Non-coding RNA
HLA-DPA1	1.10	HLA-DPA1 (Major Histocompatibility Complex, Class II, DP Alpha 1) related pathways are IL-2 Pathway and TCR signalling (Qiagen), and also associated with pulmonary arterial hypertension and lung cancer.
AP000688.14	1.10	Non-coding RNA
LTB	1.08	LTB (Lymphotoxin Beta) is a type II membrane protein of the TNF family.
POU2AF1	1.04	POU2AF1 (POU Class 2 Homeobox Associating Factor 1) is related to transcription coactivator activity and transcription coregulator activity genes.
ASTN2	1.04	ASTN2 (Astrotactin 2) is expressed in the brain and may function in neuronal migration.

Table 6-5 The 32 up-regulated DEGs were ranked by adjusted p value <0.05 and then LogFC >1.0. Gene functions or biology were obtained from GeneCards (187) and Entrez Gene (188).

Gene Symbol	**logFC**	**Gene functions and biology**
FLT1	-1.84	FLT1 (Fms Related Receptor Tyrosine Kinase 1) encodes a member of the vascular endothelial growth factor receptor (VEGFR) family. It contributes to endothelial dysfunction (264) and as a

		protective factor in sepsis-associated ARDS (265).
CTD-2616J11.10	-1.75	Non-coding RNA
MIRLET7D	-1.73	MIRLET7D (MicroRNA Let-7d) non-coding RNAs that are involved in post-transcriptional regulation of gene expression, involved in DNA damage response, and regulates proliferation, migration and tubulogenesis in endothelial cells (266).
C9orf163	-1.59	C9orf163 (Chromosome 9 Putative Open Reading Frame 163)
JAG1	-1.56	JAG1 (Jagged Canonical Notch Ligand 1). Among its related pathways are NOTCH2 Activation and Transmission of Signal to the Nucleus and Signalling by NOTCH1.
AMACR	-1.42	AMACR (Alpha-Methylacyl-CoA Racemase) related pathways are Synthesis of bile acids and bile salts and Peroxisomal lipid metabolism.
LL0XNC01-7P3.1	-1.19	Non-coding RNA
NPNT	-1.19	NPNT (Nephronectin) is a functional ligand of integrin alpha-8/beta-1 in kidney development. May also play a role in the development and function of various tissues, regulating cell adhesion, spreading and survival through the binding of several integrins.
RP5-1074L1.4	-1.18	Non-coding RNA
RP5-882C2.2	-1.17	Non-coding RNA

CTC-277H1.7	-1.17	Non-coding RNA
CTD-2015G9.2	-1.15	Non-coding RNA
AC022400.2	-1.09	Non-coding RNA
ABHD16A	-1.05	ABHD16A (Abhydrolase Domain Containing 16A, Phospholipase) is involved in some aspects of immunity.
KMT2E-AS1	-1.04	KMT2E-AS1 (KMT2E Antisense RNA 1) is affiliated with the lncRNA class.
ARL2BPP4	-1.04	ARL2BPP4 (ADP Ribosylation Factor Like GTPase 2 Binding Protein Pseudogene 4) is a Pseudogene

Table 6-6 There were 16 top down-regulated DEGs were ranked by adjusted p value <0.05 firstly and then by logFC<-1.0. Gene functions or biology were obtained from GeneCards (187) and Entrez Gene (188).

6.4.4.4 KEGG pathways analyses identified many pathways regulated by NPNT are involved in adherens junction, metabolism pathways, focal adhesion

To further investigate the functions of DEGs after KO NPNT in HBECs, I performed online biological analysis using the DAVID tool (https://david.ncifcrf.gov/) to retrieve functional annotations GO and KEGG for the DEGs. There were 419 DEGs screened according to |logFC|>0.5 and p<0.05 in this study. The KEGG pathways were ranked as in Table **6**-**7**, including the details of KEGG ID, term, p value, count, and gene symbol.

KEGG ID	Term	P Value	Count	Gene symbol
hsa05205	Proteoglycans in cancer	0.0044	12	TGFB1, SDC4, CASP3, FAS, CBLC, MSN, FRS2, FZD10, THBS1, TLR4, ACTB, ACTG1
hsa05416	Viral myocarditis	0.0080	6	HLA-DMB, CASP3, CD55, ACTB, ACTG1, HLA-DPA1
hsa05146	Amoebiasis	0.0085	8	ARG2, PLCB3, TGFB1, CASP3, ACTN1, ACTN4, TLR4, VCL
hsa00900	Terpenoid backbone biosynthesis	0.0118	4	MVK, HMGCS1, MVD, HMGCR
hsa01200	Carbon metabolism	0.0119	8	CS, GPI, HADHA, MCEE, PFKL, PC, ALDOA, HK2
hsa05323	Rheumatoid arthritis	0.0125	7	TGFB1, FLT1, HLA-DMB, LTB, TLR4, ATP6V1C1, HLA-DPA1
hsa05164	Influenza A	0.0141	10	SOCS3, HLA-DMB, IRF3, DDX39B, OAS2, FAS, TLR4, ACTB, ACTG1, HLA-DPA1

hsa01130	Biosynthesis of antibiotics	0.0179	11	CS, GPI, HADHA, ARG2, PFKL, MVK, HMGCS1, MVD, HMGCR, ALDOA, HK2
hsa04122	Sulfur relay system	0.0192	3	CTU1, MOCS3, TST
hsa04520	Adherens junction	0.0195	6	ACTN1, ACTN4, SORBS1, VCL, ACTB, ACTG1
hsa01100	Metabolic pathways	0.0209	38	GALNT12, PANK4, GPI, PIGT, MVK, RPN2, HAAO, GBA, HMGCR, HK2, MCEE, EXTL2, GANAB, EBP, IDS, ATP6V1C1, POLR2J3, ARG2, HMGCS1, CSGALNACT2, ALG5, NDUFC2, ALG1, ATP5J2, CS, HADHA, ALDH3A1, PLCB3, PFKL, PC, AMACR, CYP2U1, TST, LPCAT4, MVD, DHCR7, ALDOA, ALG10B
hsa00510	N-Glycan biosynthesis	0.0220	5	GANAB, RPN2, ALG5, ALG1, ALG10B
hsa05110	Vibrio cholerae infection	0.0267	5	KDELR3, ATP6V1C1, ACTB, PDIA4, ACTG1
hsa04510	Focal adhesion	0.0371	10	FLT1, SHC1, ITGB4, ACTN1, ZYX, ACTN4, THBS1, VCL, ACTB, ACTG1
hsa04670	Leukocyte transendothelial migration	0.0404	7	ACTN1, PTK2B, MSN, ACTN4, VCL, ACTB, ACTG1
hsa04145	Phagosome	0.0466	8	HLA-DMB, HGS, THBS1, TLR4, ATP6V1C1, ACTB, ACTG1, HLA-DPA1

Table 6-7 There were 16 KEGG pathways involved in CRISPR/Cas9 KO NPNT using RNA-seq analysis. Kyoto Encyclopaedia of Genes and Genomes, KEGG.

6.4.4.5 Gene Oncology (GO) function annotation

6.4.4.5.1 Biological processes (BPs)

GO functional annotation of the 419 DEGs enriched 37 terms of the cell-cell signalling of BP ($p<0.05$). The top 10 terms are shown in Table **6**-**8**, and include processes such as protein localization, epidermal growth factor receptor signalling pathway, movement of cell or subcellular component and positive regulation of gene expression.

Term	Count	P Value	Genes
GO:0006695~cholesterol biosynthetic process	6	0.0007	EBP, MVK, HMGCS1, MVD, HMGCR, DHCR7
GO:0008104~protein localization	7	0.0011	GNGT1, CD81, MALL, RAB35, PLOD3, MAPK8IP3, SQSTM1
GO:0009749~response to glucose	7	0.0019	PFKL, TGFB1, CASP3, SARM1, PTK2B, THBS1, PPARD
GO:0008299~isoprenoid biosynthetic process	4	0.0022	MVK, HMGCS1, MVD, HMGCR
GO:0007173~epidermal growth factor receptor signalling pathway	6	0.0043	TGFB1, CAMLG, SHC1, PTK2B, TGFA, SOX9
GO:0006928~movement of cell or subcellular component	7	0.0061	IGSF8, CD9, MSN, MST1R, VCL, ACTB, ACTG1

GO:0006401~RNA catabolic process	4	0.0071	STK31, OAS2, PABPC4, SND1
GO:0010628~positive regulation of gene expression	12	0.0123	CNN2, TFAP2A, EPB41L4B, TGFB1, HGS, LAMP3, GBA, PDCD5, MSN, TLR4, LDLR, PPARD
GO:0061621~canonical glycolysis	4	0.0130	GPI, PFKL, ALDOA, HK2
GO:0002576~platelet degranulation	7	0.0142	TGFB1, ACTN1, CD9, ACTN4, ALDOA, THBS1, VCL

Table 6-8 GO functional annotation of the 419 DEGs enriched: the top 10 cell-cell signalling of biological processes (BPs).

6.4.4.5.2 Cell components (CCs)

GO functional annotation of the 419 DEGs were enriched for 31 terms of cell-cell signalling of CCs ($p<0.05$). The top 10 of the CCs pathways were shown in Table **6**-**9**, and include membrane, endoplasmic reticulum, focal adhesion, cell surface, and extracellular exosome components.

Term	Count	PValue	Genes
GO:0016020~membrane	69	0.0000	GPI, PIGT, HSP90AB1, CD81, PREP, AP2A1, IL1RAP, HK2, ACTB, ACTG1, PNN, AP1G2, LAMP3, PDE4B, PDE4A, PHLDA2, CDK5RAP3, CAST, CHID1, CAMLG, CSGALNACT2, ALG5, EMP1, FRS2, ALG1, **NPNT**, SND1, SREBF2, TAPBP, ATXN2L, PLCB3, OAS2, EHBP1L1, HNRNPH2, PABPC1, ALDOA, PPIB, ERGIC1, NGEF, IGSF8, SLC26A2, RPN2, RRBP1, ADIPOR1, KLC1, HSP90B1, CNN2, GANAB, KLC2, MLEC, LRRC8A, LDLR, RHBDL1, ACBD5, JAG1, PCDHGC3, SMARCA4, CLCN5, PFKL, LPCAT4, DNAJA2, CD9, NUCB1, FAS, GOPC, CLPTM1L, DHCR7, RHEBL1, SLC26A6
GO:0005783~endoplasmic reticulum	33	0.0000	RPN2, PLOD3, RRBP1, HMGCR, THBS1, HSP90B1, CLN5, EXTL2, EBP, SPAST, MLEC, TPD52, CAST, CRTAP, CAMLG, SPINK5, CNPY2, ALG1, SREBF2, TMCO1, TAPBP, SUMF1, ALDH3A1, TMX4, LPCAT4, OAS2, TRAPPC6B, DHCR7, CRELD2, PPIB, SQSTM1, RNF185, SLC26A6
GO:0005925~focal adhesion	20	0.0001	CYFIP1, FLT1, SDC4, CD81, FLII, ACTN1, MSN, ACTN4, SORBS1, ACTB, ACTG1, HSP90B1, CNN2, EHD3, ZYX,

			CD9, PTK2B, PABPC1, PPIB, VCL
GO:0009986~cell surface	22	0.0010	TGFB1, CLDND1, HSP90AB1, SDC4, ITGB4, ENTPD6, TGFA, PLAT, MST1R, FZD10, THBS1, PDIA4, ADAM15, FAS, VAMP1, ADAM8, LRRC8A, LDLR, RAMP1, TLR4, CD55, HLA-DPA1
GO:0005765~lysosomal membrane	14	0.0016	HSP90AB1, SLC44A2, GBA, CLN5, TM9SF1, CLCN5, HLA-DMB, OSTM1, LAMP3, GOPC, UBA1, LAMTOR2, ATP6V1C1, HLA-DPA1
GO:0045177~apical part of cell	7	0.0024	EPB41L4B, CLCN5, IFIT5, MSN, PLAT, LDLR, ATP6V1C1
GO:0001725~stress fiber	6	0.0029	CNN2, ACTN1, ZYX, ACTN4, SORBS1, MST1R
GO:0097433~dense body	3	0.0065	SND1, ACTB, ACTG1
GO:0070062~extracellular exosome	69	0.0072	GPI, CYFIP1, TINAGL1, HSP90AB1, SLC44A2, CD81, ITGB4, PDCD5, PLOD3, CBLC, PLAT, PTPN23, ELFN1, ACTB, ACTG1, CAPN7, SH3BGRL3, POLG, CHID1, LYNX1, ACTN1, ENTPD6, SPINK5, ACTN4, **NPNT**, SDCBP2, SND1, ATP5J2, ANO1, ADAM15, TST, HGS, RAB35, PABPC1, ALDOA, PPIB, SQSTM1, PLBD2, VCL, IGSF8, SLC26A2, MVK, SEMA3C, SDC4, HAAO, GBA, LTBP4, LTBP3, THBS1, HSP90B1, CLN5, CNN2, GANAB, EXTL2, SPAST, ATP6V1C1, CD55, PCDHGC3, MSN, CS, PFKL, MAN2B2, DNAJA2, CD9, NUCB1, FAS, UBA1, LAMTOR2, SERINC2

GO:0005829~cytosol	79	0.0078	GPI, CYFIP1, MIDN, MOCS3, HSP90AB1, FAM13B, VARS, PDCD5, AP2A1, ZDHHC8, HK2, ACTB, ACTG1, CAPNS2, CASP3, PDE4B, FBXO4, NEFL, PDE4A, CAST, HMGCS1, ACTN1, SPINK5, HAUS3, CSNK1E, SREBF2, SRRM1, ALDH3A1, PLCB3, CTU1, IRF3, TRAPPC6B, HGS, OAS2, ATG4C, VAMP1, TNFRSF25, MVD, PABPC1, ALDOA, SQSTM1, SFI1, NGEF, VCL, MVK, SHC1, HAAO, CIPC, FBF1, DTX1, HSD17B14, KLC1, HDAC6, HSP90B1, CLN5, SOCS3, KLC2, BLOC1S3, HSF1, PTK2B, IGF2BP2, IP6K1, ATP6V1C1, RBM38, UBE2A, DDHD1, SORBS1, WASH3P, PFKL, EHD3, PC, OSTM1, CAPN10, DNAJA2, SARM1, UBE2O, FAS, UBA1, SPRN

Table 6-9 GO functional annotation of analyses of the 419 DEGs enriched components: the top 10 cell-cell signalling of cell components (CCs).

6.4.4.5.3 Molecular functions (MFs)

GO functional annotation of the 419 DEGs identified 12 terms of cell-cell signalling of Molecular Functions (MFs) ($p<0.05$). The top 10 of the MFs pathways are shown in Table **6**-**10**, and include protein binding, transforming growth factor beta binding, glycoprotein binding, and protein complex binding.

Term	Count	P Value	Genes

GO:0005515~protein binding	205	0.0000	CYFIP1, VIPR1, VARS, PLEKHB2, ZMYND8, PREP, FZD10, ALKBH6, ACTB, SCAMP5, FAM122A, FBXO4, SOX9, CCNL2, JUNB, CDK5RAP3, CHID1, JMJD4, ACTN1, FRS2, ACTN4, CSNK1E, MAPK8IP3, EML2, CDC40, SND1, SRRM1, ARMC1, DDX39B, MALL, SF1, FLG, IGSF8, RWDD2B, RPN2, SDC4, FLII, FBF1, TGFA, MST1R, SPAST, BLOC1S3, ARGLU1, ATP6V1C1, JAG1, SMARCA4, TMEM199, EHD3, CDK10, TINAGL1, MOCS3, PIGT, AP2A1, PHF6, EDA2R, GNGT1, MCEE, NEFL, GABPA, UTP14C, CAMLG, PABPC4, EMP1, IL17RE, SREBF2, MED29, CNKSR1, ANO1, HADHA, CTU1, VAMP1, HNRNPH2, PABPC1, ALDOA, CRELD2, SFI1, TLR4, ERGIC1, STK40, PPARD, FYTTD1, GRAMD1A, HAAO, CIPC, GBA, HSF1, SHOC2, LDLR, CD55, MPZL2, GABBR1, EGLN2, MBD1, CRIPT, SORBS1, SARM1, ZYX, NUCB1, FAS, LAMTOR2, GOPC, RAMP1, COL16A1, CD81, PDCD5, PLOD3, PLAT, CMIP, ACTG1, CCDC28A, CAPN7, PIM1, SLC25A46, CAST, TPD52, CSGALNACT2, CNPY2, ZBTB33, SDCBP2, INIP, PDIA4, DISP1, ATXN2L, ALDH3A1, ADAM15, HGS, RAB35, ADAM8, PPIB, MZF1, SQSTM1, VCL, INO80E, SHC1, LTBP4, DTX1, KLC1, LTBP1, HSP90B1, SOCS3, KLC2, ZKSCAN5, IP6K1, TFAP2A, ZBTB17, CCDC17, TGFB1, MSN, UBE2A, CLK2, GDNF, APPBP2, UBXN11, COPS2, CAPRIN2, LHX6, UBE2O, UBA1, RNF185, NFKBID, HSP90AB1, FLT1, ITGB4, PTPN23, HK2, TRO, AP1G2, ABHD16A, SNRNP70, CASP3, PDE4A, NPIPA7, TNPO2, POLG, RBM6, NAA30, AHSA2, HAUS3, TAPBP, IRF3, TRAPPC6B, OAS2, MVK, LENG8, HSD17B14, HMGCR, THBS1, HDAC6, CLN5, HDAC7, SH3BP1, HECTD3, PTK2B, LRRC8A, IGF2BP2, TAF12, HIP1R, ARID3B, TTC17, PFKL, PC, POU2AF1, DNAJA2, CD9, RHEBL1

GO:0050431~transforming growth factor beta binding	4	0.0032	LTBP4, LTBP3, THBS1, LTBP1
GO:0005178~integrin binding	8	0.0040	COL16A1, ADAM15, ACTN1, LTBP4, CD9, ACTN4, NPNT, THBS1
GO:0044822~poly(A) RNA binding	35	0.0058	FYTTD1, HSP90AB1, IFIT5, RRBP1, HTATSF1, PHF6, YBX3, PNN, GANAB, RAVER1, SNRNP70, IGF2BP2, RBM6, RBM38, CAST, DIMT1, PABPC4, ACTN4, CSNK1E, CDC40, SND1, PDIA4, SRRM1, ATXN2L, CS, DDX39B, ZYX, UBE2O, UBA1, HNRNPH2, PABPC1, ALDOA, PPIB, SUGP2, SF1
GO:0001948~glycoprotein binding	6	0.0080	TGFB1, HSP90AB1, TGFA, PLAT, THBS1, LDLR
GO:0005509~calcium ion binding	22	0.0356	FLG, TPD52, FKBP10, JAG1, ACTN1, PCDHGC3, LTBP4, CBLC, ACTN4, LTBP3, NPNT, THBS1, LTBP1, HSP90B1, PLCB3, EHD3, CAPNS2, NUCB1, ADAM8, CRELD2, ASTN2, LDLR
GO:0023026~MHC class II protein complex binding	3	0.0367	HLA-DMB, HSP90AB1, CD81
GO:0000247~C-8 sterol isomerase activity	2	0.0379	EBP, SREBF2
GO:0004871~signal transducer activity	9	0.0431	GNGT1, PLCB3, SLC44A2, COPS2, FAS, CBLC, PTK2B, IL1RAP, GULP1
GO:0032403~protein complex binding	9	0.0452	HADHA, CYFIP1, CRTAP, DDX39B, CASP3, CRIPT, PTK2B, AP2A1, PPIB

Table 6-10 GO functional annotation of the 419 DEGs enriched the top 10 cell-cell signalling of molecular functions (MFs).

6.4.4.6 Protein-protein interaction

The PPI network was performed in Cytoscape (Figure **6-13**). The top 10 target genes were selected, including ACTB, CASP3, ACTG1, HSP90B1, HSP90AB1, SMARCA4, TLR4, TGFB1, VCL, and DDX39B (Figure **6-14** and Table **6-11**). These top 10 DEGs were significantly up- or down-regulated following KO by CRISPR/Cas9 in HBECs.

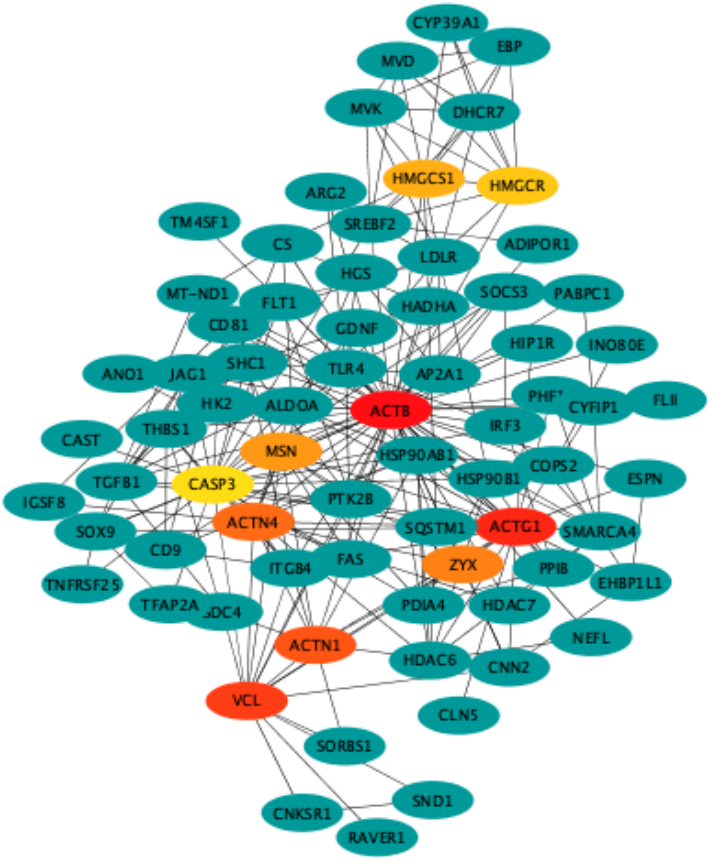


Figure 6-13 PPI network based on 419 DEGs after CRISPR/Cas9 KO of NPNT. The ACTB, CASP3, ACTG1, HSP90B1, HSP90AB1, SMARCA4, TLR4, TGFB1, VCL, DDX39B were ranked as the top 10 nodes by Degree in Cytoscape (coloured from dark red to yellow light).

The lines between the nodes represent the relationship among different target genes. PPI, protein-protein interaction.

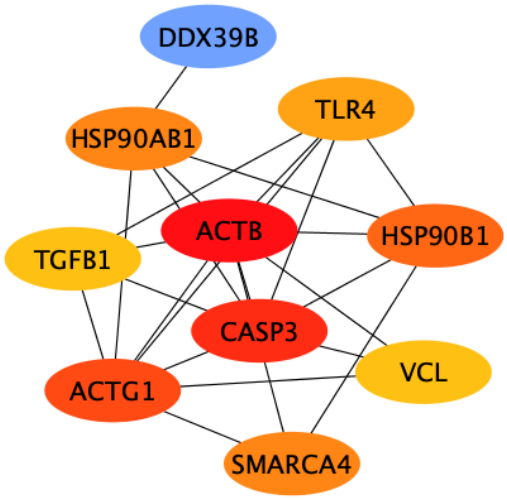


Figure 6-14 The relationship among the top 10 target genes screened by Cytoscape. The lines represent the association between the target genes.

Rank	Gene symbol	Score
1	ACTB	51.0
2	CASP3	25.0
3	ACTG1	24.0
4	HSP90B1	20.0
5	HSP90AB1	17.0
5	SMARCA4	17.0
7	TLR4	16.0
8	VCL	15.0
8	TGFB1	15.0
10	DDX39B	14.0

Table 6-11 Top 10 in network ranked by Degree method in Cytoscape. The higher value of score indicates the stronger association.

6.4.4.7 Overlap of DEGs between CRISPR/Cas9 and siRNA experiments

As shown in Figure **6-15** and Figure **6-16**, there are 41 and 70 DEGs which overlap between experiments using CRISPR/Cas9 and siRNA a or siRNA b, respectively. All DEGs were selected using an adjusted p value <0.05.

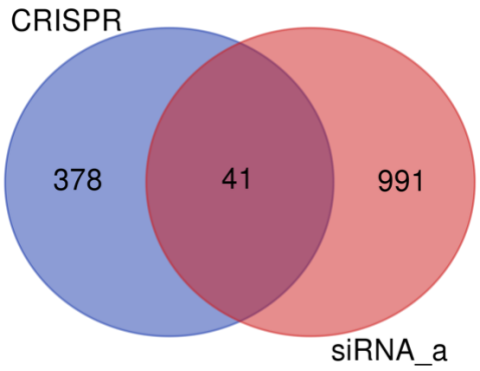


Figure 6-15 Venn diagram of DEGs expressed genes in common both of CRISPR/Cas9 and siRNA a experiments. There are 41 DEGs overlap in common. All DEGs were selected by adjusted p value <0.05.

Figure 6-16 Venn diagram of DEGs expressed genes in common both of CRISPR/Cas9 and siRNA b experiments. There are 70 DEGs overlap in common. All DEGs were selected by adjusted p value <0.05.

There are 20 DEGs in common among of 3 experiments including CRISPR/Cas9, siRNA a and siRNA b (Figure 6-17). However, there are only 7 differentially expressed genes which were consistent up- or down-regulation after NPNT knockdown or knockout by siRNA or CRISPR/Cas9, respectively, one of which was not surprisingly NPNT itself.

Of the DEGs which were consistently altered in regulation in all 3 data sets, and therefore in which one would have the greatest confidence that they are regulated by NPNT expression, there were 3 DEGs up-regulated. Thiosulfate sulfurtransferase like domain containing 3 (TSTD3) is highly expressed in a wide range of tissues, including lung, but the functions still need further study (188). Chondroitin sulfate N-acetylgalactosaminyltransferase 2 (CSGALNACT2) is expressed in Zebrafish early development (267). Family with sequence similarity 43 member A (FAM43A) is expressed in triple negative breast cancer (268) and SNPs located within FAM43A associated with autism and learning difficulties (269).

There were 3 down-regulation DEGs. Polymorphisms of gene Filaggrin (FLG) are associated with atopic dermatitis in women (270). Also studies demonstrate FLG is potentially involved in various cancers, including lung cancer (271). Thrombospondin 1 (THBS1) encodes a protein which can bind to fibrinogen, fibronectin, laminin, type V collagen and integrin $\alpha V\beta 1$ (188) and it has also been reported that THBS1 is a promising therapeutic target of lung cancer (272). It is also

and expressed in systemic sclerosis fibroblasts (273). Actin gamma 1 (ACTG1) is involved in epithelial mesenchymal transition via the MAPK/ERK signal pathway in prostate cancer (274) and is upregulated in lung cancer cell lines (275).

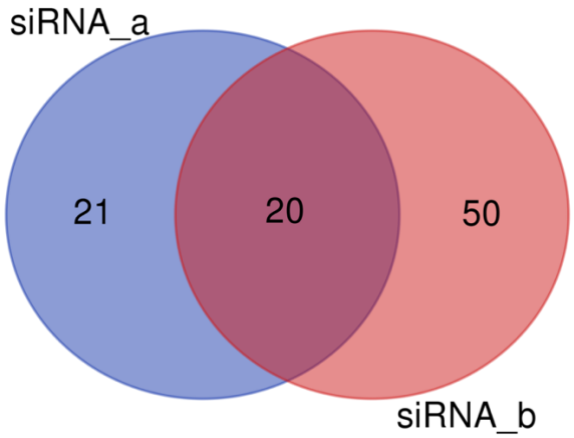


Figure 6-17 The overlap of DEGs in common between siRNA a and siRNA b and overlap with CRISPR/Cas9. There are 20 DEGs in common among all 3 experimental approaches.

Gene symbol	**CRISPR/Cas9**	**siRNA a**	**siRNA b**
GABBR1	-0.30	-0.26	0.22
PVRL1	-0.20	0.37	0.33
PLBD2	-0.16	0.30	0.29
NPNT	**-1.19**	**-2.60**	**-3.19**
TSTD3	**0.58**	**0.62**	**0.78**
CSGALNACT2	**0.28**	**0.26**	**0.32**
SEMA3C	0.40	-0.66	-0.45
UBE2O	-0.12	0.21	0.27
FAM43A	**0.27**	**0.93**	**1.01**

HTATSF1	0.18	-0.29	-0.20
COL16A1	-0.18	0.26	0.29
FLG	**-0.52**	**-0.91**	**-1.04**
THBS1	**-0.25**	**-1.28**	**-0.70**
LRRC8A	-0.14	0.15	0.22
SOCS3	-0.34	0.42	0.56
SPOCD1	-0.28	0.33	0.38
IGSF8	-0.22	0.26	0.25
ACTG1	**-0.19**	**-0.62**	**-0.30**
CLDND1	-0.21	0.27	0.39
ENTPD6	-0.17	0.31	0.22

Table 6-12 The logFC values of 20 overlap DEGs among 3 experiments. There are only 7 genes which show consistent up- or down-regulation after NPNT knockdown or knockout by NPNT RNAi or CRISPR/Cas9.

6.5 Discussion

In this chapter, I employed the protocol of a fast, robust, cloning-free and high efficiency electroporation-based CRISPR/Cas9 genome editing system which was established by Rapiteanu and colleagues (138) to further explore gene expression by NPNT. I investigated of NPNT downstream DEGs and pathways KO by CRISPR/Cas9 using transcriptomics in BMI-1 cells.

The main finding in this chapter was that I successfully used the electroporation-based CRISPR/Cas9 genome editing system to KO NPNT expression in human BMI-1-engineered bronchial epithelial cells. The NPNT expression was detected and confirmed by RT-PCR and RNA sequencing. In the data shown in this chapter, there was only one crRNA which produced significant knockdown of NPNT in BMI-1 cells using RNA-seq, although all three crRNA could significantly reduce NPNT expression in cells when measured by RT-PCR. This is probably due to statistical issues. In RT-PCR was looking specifically at one outcome, in RNA-seq analyses was looking at multiple outcomes. Thus, the statistical tests have to be more rigid to exclude false positives. This result suggests that I could establish a stable NPNT Knockout BMI-1 cell line for further research, such as air-liquid interface (ALI) culture.

HBECs are often only cultured for two or three passages due to losing their ability to differentiate, or due to senescence after further passages (276). Furthermore, the bronchial epithelial cells from different donors have different genetic backgrounds, and may have diverse morphology

in ALI culture (277). The KO efficiency of CRISPR/Cas9 may be different in different donors. The BMI-1-engineered HBECs enables the generation of a large number of cells which should allow selection of cells for ALI culture. In the future this could facilitate the development of techniques to provide stably manipulated cell populations from a single donor (276). In this thesis, I have used the same donor HBECs for BMI-1 overexpression and for each replicate in the experiment. The use of a single donor could avoid bias from genetic background differences.

In addition, the findings from the CRISPR/Cas9 genome editing of NPNT demonstrate that NPNT regulated downstream gene sets are implicated in many pathways, including the inflammation pathway, protein binding and cell adhesion. Consistent with the results described in chapter 4 and chapter 5 on NPNT knockdown by anti-NPNT siRNA, NPNT appears to play an important role in the inflammation pathways and cell proliferation in HBECs.

There were 155 upregulated and 264 downregulated DEGs in comparisons between G3 (crRNA3) and the no gRNA control group in this study. DTX1, GDNF, STK31, TRO, ALDH3A1, HLA-DPA1 and POU2AF1 were ranked as top upregulated DEGs. DTX1 is implicated in Notch signalling pathways via encoding a negative regulator, and the SNP located in DTX1 was associated with worse chemotherapy response and lover survival in lung cancer (278). GDNF was significantly up-regulated when NPNT was KO in HBECs in this thesis. In contrast, Linton and colleagues have demonstrated that the Gdnf expression was significantly decreased in the kidneys of Npnt knockout

mice (103). Thus, there may be an interaction between Npnt and Gdnf, which deserves further research in HBECs or lung. Furthermore, many long non-coding RNA were regulated after KO NPNT expression in HBECs. STK31 is highly expressed in lung cancer and regulates cell proliferation and cell cycle via the Wnt/β-catenin singal pathway (261). TRO (Trophinin) plays a vital role in the cell adhesion molecule complex on the cell surface (262). ALDH3A1 is highly expressed in lung, and has been reported driven metastasis of lung cancer via p53/BAG1 (279). Genetic variants at the HLA-DPA1 locus are associated with pulmonary arterial hypertension (280) and copy number variation (CNV) at the HLA-DPA1 gene in chromosome 6 (p=0.02) in current smokers was also associated with non-small cell lung cancer (NSCLC) risk (281). POU2AF1 has been reported to regulate host defence genes in human airway epithelial cells (282) and is be upregulated in IL-38 induced asthma models (283).

On the other hand, FLT1, MIRLET7D and JAG1 ranked as top downregulated DEGs. The genetic variant rs9508032 at the FLT1 locus was associated with sepsis-associated ARDS by GWAS and might be a potential therapeutic target in ARDS (265, 284). MIRLET7D is a non-coding RNA which involves in proliferation, migration and tubulogenesis in endothelial cells (266). JAG1 encodes jagged 1 protein which is a ligand for the receptor Notch 1 (188) and is involved in alveolar repair and fibrosis (285). Furthermore, it is reported that JAG1 is associated with poor survival of lung cancer through inducing metastases (286).

From the PPI network, I used Cytoscape to select the top 10 target genes, including ACTB, CASP3, ACTG1, HSP90B1, HSP90AB1, SMARCA4, TLR4, TGFB1, VCL, and DDX39B. ACTB was identified in the COPD exhaled breath condensate and might be a novel biomarker of airway of inflammation in COPD (287). ACTG1 was increased with increasing COPD exacerbation frequency (288). Furthermore, genetic variants in the HSP90B1 and HSP90AB1 have been identified association with decreased COPD risk in subjects exposed to biomass-burning smoke (289). TLR4 is a member of Toll-like receptor (TLR) family which plays an important role in innate immunity. Genetic variant of TLR4 was associated with COPD phenotype (290). A recent study showed that TLR4 was lower expressed in current smokers, but TLR4 did not associate with neutrophilic inflammation or airway hyperresponsiveness (291). TGF-β1 protein had lower expression in COPD bronchiolar epithelial cells compared to smoker controls using immunohistochemistry (292). And Celedon et al has revealed genetic variant of TGFB1 is associated with COPD (293). However, a meta-analysis demonstrated that the rs1800469 T allele was not significantly association with the risk of COPD in Asian population. These genes were potentially involved in the biological mechanisms of COPD when NPNT knockout in HBECs by CRISPR/Cas9 genome editing method.

KEGG pathways analyses identified many pathways regulated by NPNT are involved in adherens junction (AJ), metabolism pathways, and adhesion. Adherens junctions plays an important role in maintaining the permeability and integrity of the epithelial and

endothelial barriers (294). KEGG showed that AJ was significantly regulated after genome editing NPNT in epithelial cell, indicating that AJ were also impaired after NPNT KO and might involve in COPD. A COPD GWAS gene, FAM13A promoted fatty acid oxidation and implicated in cellular metabolic pathways when cigarette smoking exposure (295). In the analysis results of KEGG also demonstrated that NPNT potentially shaped the metabolic pathways after NPNT KO in epithelium and need further investigation. Interestingly, Wijnhoven et al reported that the focal adhesion complex was no differences in the diaphragm biopsies between COPD and non-COPD patients (296). However, I found focal adhesion pathways have regulated after NPNT knockout by genome editing, indicating these pathways might implicated in the cellular proliferation and migration in epithelial cell.

There were 7 DEGs (including NPNT itself, and TSTD3, CSGALNACT2, FAM43A, FLG, THBS1, ACTG1) which were common in all 3 data sets from knockdown experiments (siRNA a, siRNA b, and CRISPR/Cas9). This could suggest that there are likely to be quite a few false positives in each different experiment, but one would have strong confidence that the 7 DEGs which were present in the three data sets using two different approaches (RNAi and CRISPR/Cas9), and which showed a consistent direction of effect are real.

There are some limitations of the work presented in this chapter. Firstly, I detected the NPNT mRNA expression from the bulk cells, not by analysing just the proportion of KO cells by MiSeq genotyping. Secondly, my data showed that NPNT regulated many genes and

pathways in BMI-1 engineered HBECs, but this work ideally needs to be followed up with experiments in cells differentiated at ALI culture. Finally, as discussed above, whilst there were a number of genes where consistent effects were seen using multiple approaches to knock down NPNT, there were genes seen which were only differentially regulated using one or two rather than all 3 approaches. Increasing the number of replicates might allow detection of genes which are differentially regulated following NPNT knockdown.

7. Serum Nephronectin as a promising biomarker for chronic obstructive pulmonary disease

7.1 Introduction

A biomarker is an objectively measured marker, for example a circulating protein, which can potentially provide an indication of disease activity (297). Whilst COPD is usually defined using spirometry to assess airway obstructive, it can also produce a systemic inflammation. Blood samples are the most widely used source of biomarkers in clinical research (298). According to recent studies, four types of blood biomarkers including transcriptomics, DNA methylation, proteomics, and metabolomic profiling have been applied in COPD (298). These biomarkers could lead to novel ways to help with diagnosis, assessing disease progression, exacerbation prediction, identification of different sub phenotypes of disease, and may provide novel therapeutic targets.

There has been previous work on developing biomarkers to predict progression and enable precision medicine with COPD. For example, Sun and colleagues have measured 88 blood proteins in patients with COPD and explored genetic factors which might affect biomarker levels (299). However, to date it has been extremely challenging to translate research to clinical application due to poor replication of findings or non-consistent effects in different cohorts. Up until now, sRAGE is probably the most promising biomarker for COPD, and also showed strong correlation with FEV_1/FVC and COPD risk in a recent study which integrated INTERVAL cohort (n=3,301) and UK Biobank cohorts (n=400,102) (300). sRAGE will be the first blood biomarker of

emphysema to be submitted to the Food and Drug Administration (FDA) (301). Moreover, plasma fibrinogen has been suggested as a drug development tool for COPD by the COPD Foundation Biomarker Qualification Consortium (302). Other plasma/serum biomarkers of COPD and emphysema include: MMPs (168), interleukins (303), adiponectin (304), and surfactant protein D (SPD) (298).

COPD exacerbation is defined as “an increase in dyspnoea, cough, or sputum purulence with or without symptoms of upper respiratory infection” (305). Exacerbations of COPD result in declines of lung function, disease progression, higher cost, and higher mortality. It is therefore important to facilitate early recognition and prediction of COPD exacerbations in clinical practice. Jones et at suggested 14 items which could quantitatively identify exacerbation risk (306, 307). Some studies have reported that the white blood cell count (308), C-reactive protein (CRP), fibrinogen (309), adiponectin (304), and Fetuin-A (310) could be implicated as predictive of COPD exacerbations. In a recent study, Keene et al investigated 90 plasma or serum candidate proteins from the subjects from SPIROMICS (n=1544) and COPDGene (n=602) cohorts (311). The results overall showed poor reproducibility of specific biomarkers between two cohorts. However, they identified decorin and α2-macroglobulin as showing predictive value for future severe exacerbations. Furthermore, Tong and colleagues reported in a meta-analysis that serum YKL-40 levels protein was higher during exacerbation when compared with the stable stage, and suggested that

the serum YKL-40 expression was correlated with clinical severity and associated with reduced lung function (312).

Nephronectin (NPNT), as a novel extracellular matrix present in the blood, has the potential to be a biomarker for disease. There have been some preliminary studies exploring expression in serum/plasma of patients with kidney disease, silicosis and autoimmune mouse models (99, 105, 169). Watany and colleagues investigated the expression levels of NPNT and assessed NPNT as a candidate non-invasive biomarker of glomerular regeneration and to identify steroid treatment response (105). In this retrospective study, they collected 80 nephrotic syndrome patients and 40 healthy controls, measured NPNT concentration in serum protein and mRNA using ELISA and real time PCR, respectively. They found that *NPNT* mRNA and protein were significantly overexpressed in steroid sensitive nephrotic syndrome (SSNS) (10.82$\pm$7.39 ng/mL) than in steroid-resistant nephrotic syndrome (SRNS) (1.19$\pm$0.94 ng/mL). Thus, they suggested that NPNT could be a potential biomarker of glomerular regeneration (105).

Lee et al reported NPNT was expressed at higher levels in the blood of patients with silicosis (12.77$\pm$1.50 ng/mL) compared to healthy controls (9.86$\pm$3.16 ng/mL) (99). NPNT expression was not correlated with age or FVC, but correlation was observed between NPNT level and profusion rate (radiological classification), FEV_1 percentage, MCP-1 and MCP-1α. The results suggest that NPNT plays a role in the initiation and progression of silica-induced lung fibrosis, and NPNT

levels are affected by the worsening of inflammatory responses. Thus, NPNT could be suggested as a potential biomarker and therapeutic target of silica-induced lung fibrosis (99). In a recent study by Kon also reported that nephronectin protein levels were elevated in the serum of autoimmune mouse models when measured by ELISA, including concanavalin A (ConA)-induced hepatitis, collagen antibody-induced arthritis (CAIA), and EAE mouse models (169).

7.2 Aims

In this chapter, it was hypothesised that serum nephronectin levels might be a potential biomarker for COPD or the exacerbation stage of COPD, contributing to the airway remodelling of chronic airway diseases. I also explored correlations between serum nephronectin levels and lung function, smoking status, age, or kidney function (eGFR).

7.3 Methods

7.3.1 Enzyme Linked Immunosorbent Assay

NPNT was measured using enzyme-linked immunosorbent assay (ELISA) as detailed description in Chapter 2. The kit provided microtiter plate was pre-coated with an antibody specific to NPNT.

In brief, 100μL each of diluted standards or serum samples were assayed in 96 well ELISA plates. After incubation for 1h at 37℃ the liquid from each well was removed and 100μL of detection reagent A working solution (biotin-conjugated anti-NPNT) added and incubated for 1h at 37℃. Before discarding all fluid, 100μl of detection reagent B working solution (Avidin conjugated to Horseradish Peroxidase, HRP) was added, and incubated for 30 min at 37℃. The wash process was repeated 5 times and then 90μL of substrate solution was added to each well, and incubated for 10-20 min at 37℃. Finally, the sulphuric acid stop solution (50μL) was added to each well, and NPNT protein was measured at 450nm immediately using a plate reader. The selected minimax of detection of the assay was 10 pg/ml.

7.3.2 Patient recruitment

Study participants from two separate cohorts took part in the study, the UK Nottingham Cohort and China Ningbo Cohort. In the UK Nottingham cohort, all subjects were recruited from City Hospital of University of Nottingham. All subjects were seen in clinic, and the data was collected by Professor Charlotte Bolton (City Hospital of Nottingham, University

of Nottingham) as part of an ongoing research study on COPD. In the China Ningbo Cohort, the serum of acute exacerbation of chronic obstructive pulmonary disease (AECOPD) patients were collected from Feb 2017 to Sep 2020 within 48 h of hospital admission, and the healthy controls were enrolled from the outpatient clinic. All participants provided written informed consent and the Ningbo study was approved by the Ethical Committee of Ningbo First Hospital, Ningbo City, Zhejiang Province, China (2017-R015-01), and the project of Lung function and Cardiovascular risk was approved by the University of Nottingham, Nottingham, UK (10/H0406/65).

In both cohorts, the diagnostic criteria for COPD were based on forced expiratory volume in 1 s (FEV_1)/forced vital capacity (FVC) ratio of < 0.7 and post-bronchodilator FEV_1 < 80% predicted. Severity of COPD was graded according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria (15).

Healthy controls: healthy controls were collected and had no history of co-existing diseases, such as pulmonary diseases (interstitial lung fibrosis, bronchiectasis, granulomatous lung diseases), renal or hepatic dysfunction, infection, malignancy or autoimmune diseases.

Stable COPD patients: The stable COPD patients were defined as follows: age > 40 years; diagnosed with COPD as GOLD criteria; without an acute exacerbation in the last 6 months.

COPD Exacerbation patients: The diagnostic approach to exacerbation of COPD was based on the diagnosis of physicians and

clinical setting. All patients were of age > 40 years. Any patients with asthma, lung cancer, or severe heart failure were exclusive.

7.3.3 Samples collection and biochemical analysis

Blood samples were collected in tubes and serum samples were separated by centrifugation at 10,000×g for 10 min. The laboratory results, including electrolyte, renal function, liver function, blood routine test, triglyceride and cholesterol, were measured by automated chemistry analyser in the hospital clinical laboratory. Serum nephronectin concentration was detected by ELISA as described above and in Chapter 2.

7.3.4 Statistical analyses

7.3.4.1 Population statistics

The serum nephronectin levels were compared in both cohorts between healthy controls and COPD stable or exacerbation groups using the Kolomogorov-Smirnov tests.

7.3.4.2 Correlation analyses

The distribution of NPNT levels in both cohorts were determined to be non-normalised. All the correlation analyses used Person's correlation. Correlation was performed on the potential co-variates for nephronectin expression in serum, such as age, height, weight, and smoking pack/years, creatinine blood, urea nitrogen, eGFR, urinary albumin creatinine ratio (UACR). Moreover, correlation analyses were also performed on serum NPNT levels with the lung function parameters,

including FEV_1 (L), FEV_1 (%pred), FEV_1/FVC. Analyses were carried out separately in each cohort.

7.3.4.3 Linear regression

Prior to further analyses, data were nalysed to confirm they were normally distributed. Simple linear regression analysis was used to evaluate the relationship between variables and NPNT.

7.4 Results

7.4.1 Serum NPNT levels are no different between healthy controls and stable COPD in the UK Nottingham Cohort

There were 75 healthy control subjects and 112 stable COPD patients recruited into the UK Nottingham cohort. The features of the subjects are outlined in Table 7-1. The results show, the stable COPD group were older than the healthy controls, and more likely to be male but no different in body mass index (BMI). Furthermore, the proportion of smokers was higher in stable COPD than healthy controls and also had a higher smoking pack/year history. The history of ischemic heart disease (IHD), congestive heart failure (CHF), hypertension and diabetes were higher frequency in COPD group than in healthy controls, while there were no significant different of electrolyte, urea nitrogen, creatinine blood between these two groups. Serum NPNT levels were not significantly different between healthy controls (median 527.1 pg/mL) and stable COPD patients (median 621.7 pg/mL) (Figure 7-1).

	Healthy controls	**Stable COPD patients**	***P* value**
Subject n	75	112	-
Age, yrs	67.51(8.92)	71.55(6.66)	<0.05
Males (%)	39 (52.00%)	65 (58.04%)	<0.001
Height, m	1.68(0.09)	1.68(0.09)	>0.05
Weight, kg	80.79(16.06)	79.74(18.78)	>0.05
BMI	28.41(4.78)	28.1(5.8)	>0.05
Lung function			
FVC, L	3.59(0.89)	2.92(0.77)	<0.001
FVC % pred	110.37(17.22)	92.31(17.99)	<0.001
FEV_1, L	2.63(0.61)	1.5(0.6)	<0.001
FEV_1 % pred	101.80(13.79)	58.89(19.38)	<0.001
FEV_1/FVC	73.75(5.44)	50.06(13.47)	<0.001
Smoking and questionnaires			
Smoker n	65	93	<0.001
Smoking pack/years	26.92(13.59)	41.53(26.76)	<0.001
CAT [IQR]	4[2-8]	17[12-23]	<0.001
mMRC [IQR]	1[1-2]	2[2-3]	<0.05
Comorbidities			
IHD n (%)	4 (5.33%)	11 (8.04%)	<0.001
CHF n (%)	1 (1.33%)	4 (3.57%)	<0.001
Hypertension, n (%)	27 (36.00%)	55 (49.11%)	<0.001
Diabetes, n (%)	11 (14.67%)	21 (18.75%)	<0.05
Renal disease, n (%)	0	2 (1.79%)	>0.05
Laboratory results			
Na^+, mmol/L	138.68(2.30)	138.71(2.83)	>0.05
K^+, mmol/L	4.49(0.45)	4.50(0.38)	>0.05
Urea nitrogen, mmol/L	6.46(8.58)	6.07(2.48)	>0.05
Creatinine blood, μmol/L	75.93(18.53)	76.94(19.34)	>0.05
eGFR, mL/min	77.88(12.30)	75.99(15.26)	>0.05
UACR	1.91(2.93)	4.71(13.81)	>0.05
NPNT, pg/mL	527.13 [321.19-895.55]	621.68[391.96-905.17]	>0.05

Table 7-1 Summary of clinical features and laboratory results of the UK Nottingham Cohort. Data are presented as average (SD) or median [interquartile range], unless otherwise stated. Body mass index (BMI), ischemic heart disease (IHD), congestive heart failure (CHF).

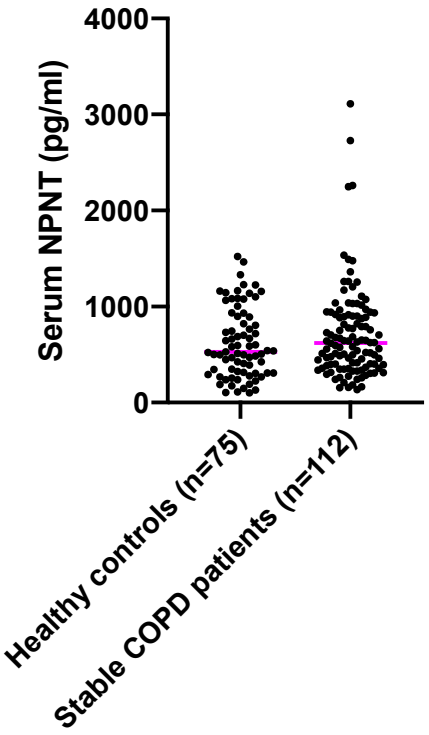


Figure 7-1 Serum NPNT levels in healthy controls and stable COPD patients. A scatter plot diagram identifies that serum NPNT levels in healthy controls (median 527.13 pg/ml) were not statistically significant when compared to COPD stable patients (median 621.68 pg/ml).

7.4.2 Identifying co-variates that influence NPNT levels of UK Nottingham cohort

In the UK Nottingham cohort study, the patient age and creatinine blood were positively correlated with serum NPNT levels, while eGFR was negatively correlated with serum NPNT levels (Table **7**-**2**). The population smoking pack/years, height, weight, urea nitrogen, UACR were not associated with the levels of serum NPNT (Figure **7**-**2** and Figure **7**-**3**).

Factor	Pearson Correlation (r)	Sig. P value	Correlated
Age	0.163	0.026	Yes
Height	-0.050	0.493	-
Weight	0.039	0.600	-
Smoking pack/years	0.063	0.392	-
Urea nitrogen	0.042	0.574	-
Creatinine blood	0.290	<0.0001	Yes
eGFR	-0.338	<0.001	Yes
UACR	0.139	0.063	-

Table 7-2 Age, height, weight, smoking pack/years, Urea nitrogen, creatinine blood, eGFR, UACR were tested as covariates in the UK Nottingham Cohort. UACR, urinary albumin creatinine ratio.

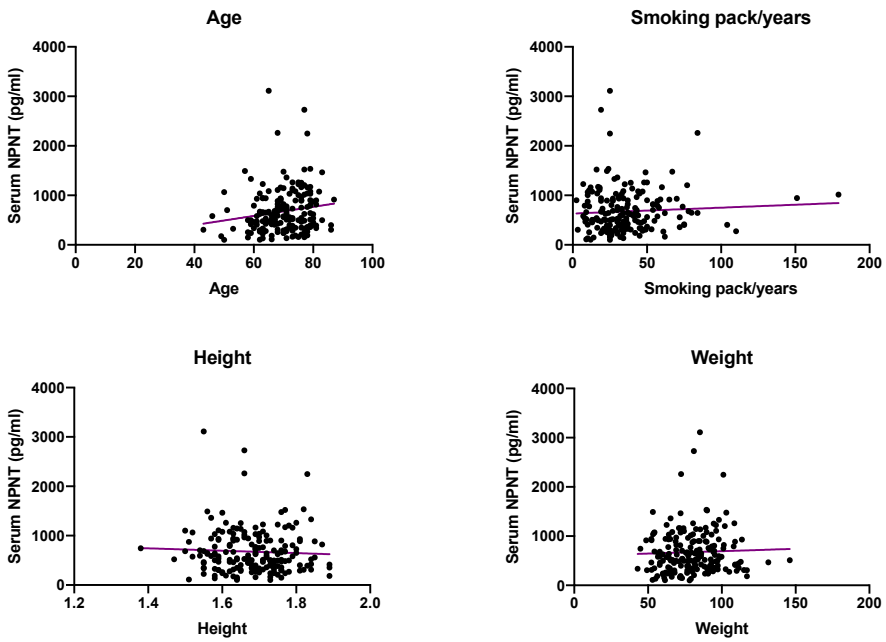


Figure 7-2 Dot plots for the linear variants tested for correlation. In the UK Nottingham Cohort, the serum NPNT expression levels were positively correlated for age (r=0.163, p<0.05), while not significantly correlated for smoking pack/years, height and weight.

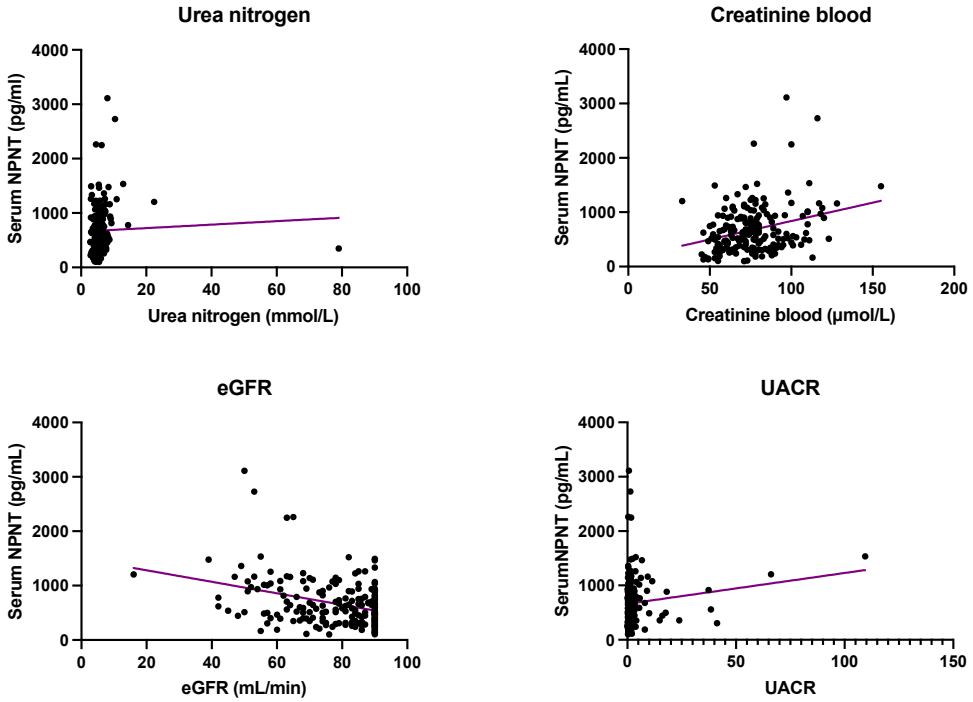


Figure 7-3 Dot plots for the linear variants tested for correlation. In the UK Nottingham Cohort, the expression levels of serum NPNT were positively correlated for creatinine blood (r=0.290, p<0.0001), negatively correlated with eGFR (r=-0.338, p<0.001), while not significantly correlated for urea nitrogen and UACR. eGFR, estimated glomerular filtration rate, UACR, urinary albumin creatinine ratio.

7.4.3 Serum NPNT levels are not associated with COPD severity in UK Nottingham cohort

I next assessed whether or not serum NPNT expression levels were regulated by COPD severity using airflow limitation according to the GOLD criteria (313). The serum NPNT levels were not significantly statistic different between the 4 grades of COPD stages (*P*=0.88) (Figure **7**-**4** A). As the numbers in the most severe group were small, I

also performed an analysis of mild/moderate v severe/very severe disease. This showed the expression of serum NPNT levels was not different between the combined mild+moderate group and severe+very severe group (Figure **7-4** B).

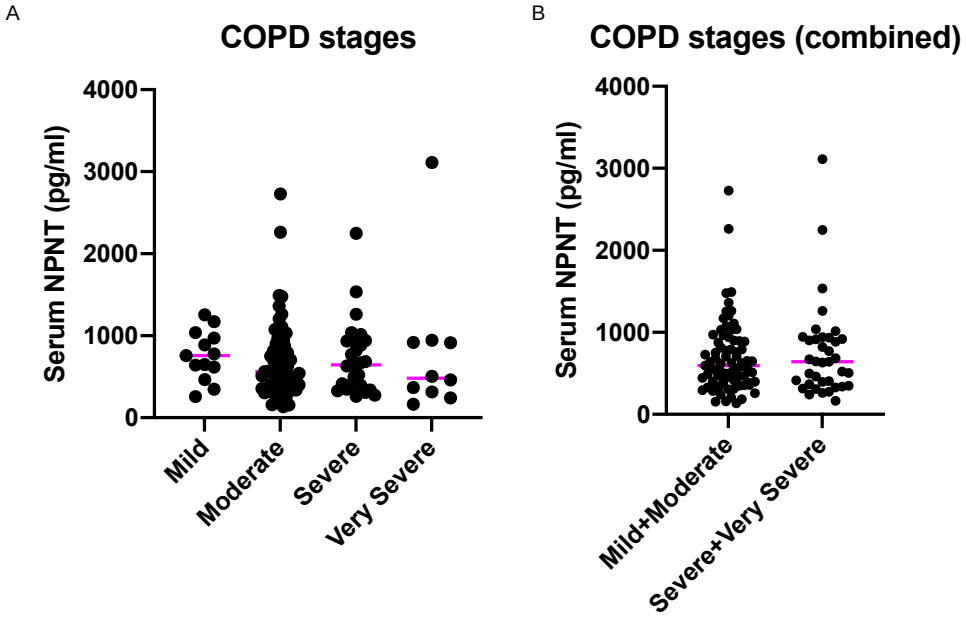


Figure 7-4 A scatter plot of serum NPNT levels in the different COPD stages (A), combined COPD stages (B). The COPD severity was stratified according to the GOLD criteria.

7.4.4 Serum NPNT levels decreased in AECOPD from China Ningbo Cohort

As shown in Table **7-3**, 64 healthy controls and 50 COPD exacerbation patients were recruited from the China Ningbo Cohort. The proportion of males were higher in the AECOPD groups (82.00%) than in healthy subjects (62.50%) and they were older than control subjects. Furthermore, the AECOPD patients had more smoking pack/years and

had lower measurements of lung function, including FEV_1, FEV_1/FVC, MEF75%, MEF50%, and MEF25% than the healthy controls. The urea nitrogen and creatinine levels were lower in healthy controls than AECOPD patients, while the serum albumin was higher in healthy subjects. The WBC, neutrophils, and NLR was significantly higher in COPD exacerbation patients than in healthy control subjects as shown in Table **7**-**3**. Importantly, the serum NPNT levels were decreased in AECOPD patients (median 1428.67 pg/mL) when compared to healthy controls (median 2728.31 pg/mL).

	Healthy controls	**AECOPD**	***P* value**
Subject, n	64	50	-
Male, n (%)	40 (62.50%)	41 (82.00%)	<0.05
Age, yrs	47.38(5.57)	70.93(11.88)	<0.001
Height, cm	164.44(7.74)	164.05(6.64)	>0.05
Weight, kg	59.00(11.75)	57.66(10.12)	>0.05
smoking pack/years	20.32(17.79)	37.52(14.61)	<0.05
CAT	-	15[8-19]	-
mMRC	-	1[2-3]	-
Lung function			
FEV_1, L	2.39(0.66)	1.33(0.82)	<0.001
FVC, L	2.82(0.83)	2.49(0.99)	>0.05
FEV_1/FVC	85.81(8.93)	48.85(11.66)	<0.001
MEF75%	4.23(1.77)	11.51+11.68	<0.05
MEF50%	3.25(1.19)	7.37(7.79)	<0.05
MEF25%	2.32(4.71)	14.73(14.90)	<0.001
Laboratory results			

Urea nitrogen, mmol/L	62.12(12.07)	89.48(69.89)	<0.001
Creatinine blood, μmol/L	4.43(1.23)	6.31(4.40)	<0.05
eGFR, mL/min	119.38(12.86)	88.91(26.98)	<0.0001
Serum albumin	48.39(3.86)	35.29(5.99)	<0.0001
Triglyceride, mmol/L	0.78(0.53)	1.00(0.48)	<0.05
Cholesterol, mmol/L	4.48(0.71)	4.24(1.24)	<0.05
C-reactive protein, ng/mL	-	46.50(62.83)	-
WBC, 10^9/L	5.92(1.37)	9.63(6.06)	<0.0001
Neutrophils, 10^9/L	3.48(1.08)	7.65(6.09)	<0.0001
Lymphocytes, 10^9/L	1.98(0.54)	1.18(0.68)	<0.0001
NLR	1.75(1.99)	6.5(8.91)	<0.0001
Eosinophils, 10^9/L	0.14(0.11)	0.16(0.19)	>0.05
NPNT serum level, pg/ml	2728.31 [2377.45-3220.12]	1428.67 [1156.09-1855.37]	<0.001

Table 7-3 Summary of the clinical features and laboratory results of healthy controls and AECOPD in China Ningbo Cohort. Data are presented as average (SD) or median [interquartile range], unless otherwise stated. NLR, neutrophil to lymphocyte ratio.

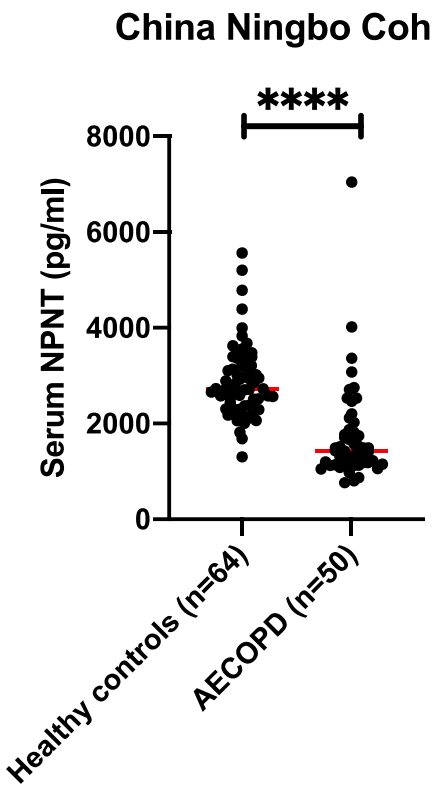


Figure 7-5 Serum NPNT levels in healthy controls and acute exacerbation of COPD patients. A scatter plot diagram identifies that serum NPNT levels in healthy controls (median 2728.31 pg/ml) were significantly higher when compared to exacerbation COPD patients (median 1428.67 pg/ml). **** $p<0.0001$

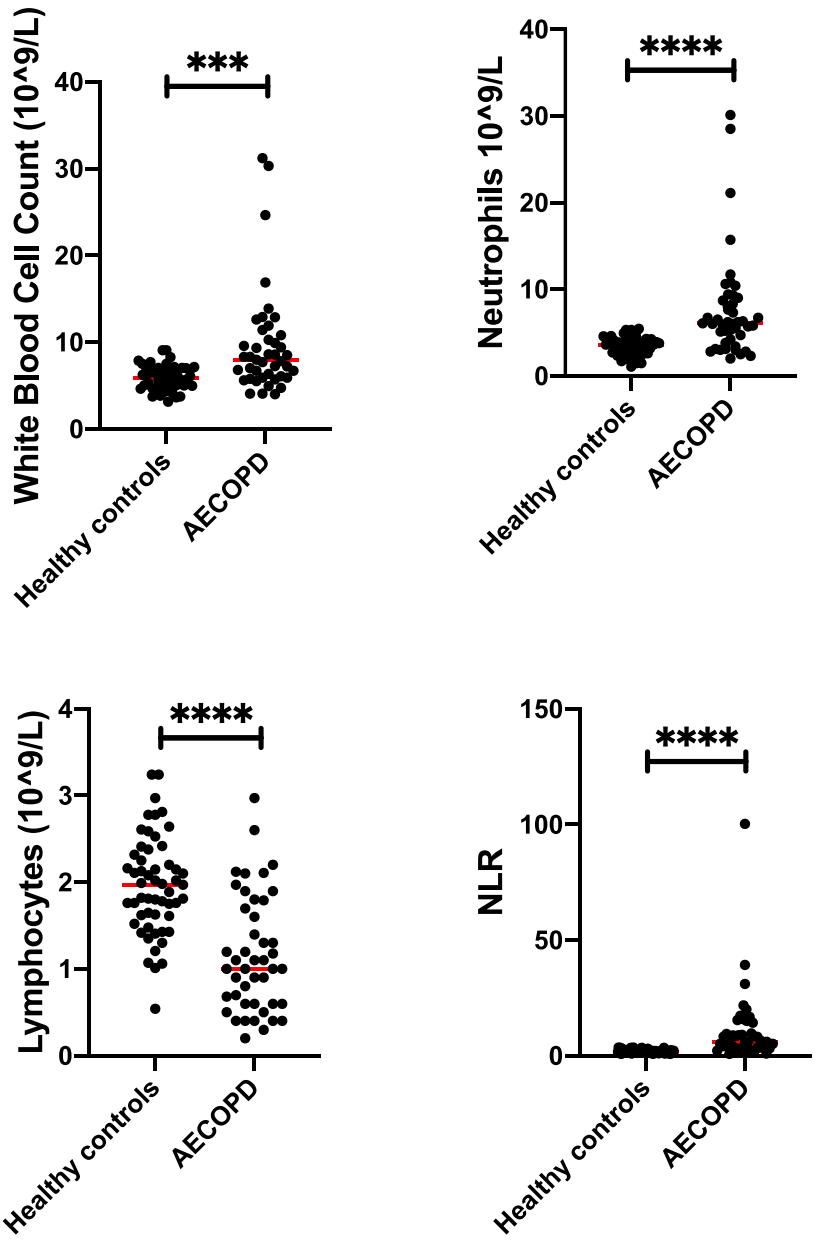


Figure 7-6 Comparison of blood tests between healthy controls and AECOPD from China Ningbo Cohort. The blood tests including white blood cell, neutrophils, lymphocytes, and NLR. ***<0.001, ****<0.0001.

7.4.5 Correlation between serum NPNT levels and lung function

There was no correlation between serum nephronectin level and lung function, including FEV_1 and FEV_1/FVC in the UK Nottingham Cohort (Figure **7**-**7** A and B). However, a significant correlation was seen between serum NPNT levels and FEV_1 (r^2=0.118 and p=0.0022) and FEV_1/FVC (r^2=0.262 and p<0.0001) in the Ningbo cohort as shown in Figure **7**-**7** C and D. There is a positive correlation between serum NPNT levels and the FEV_1 of healthy controls and COPD exacerbation patients.

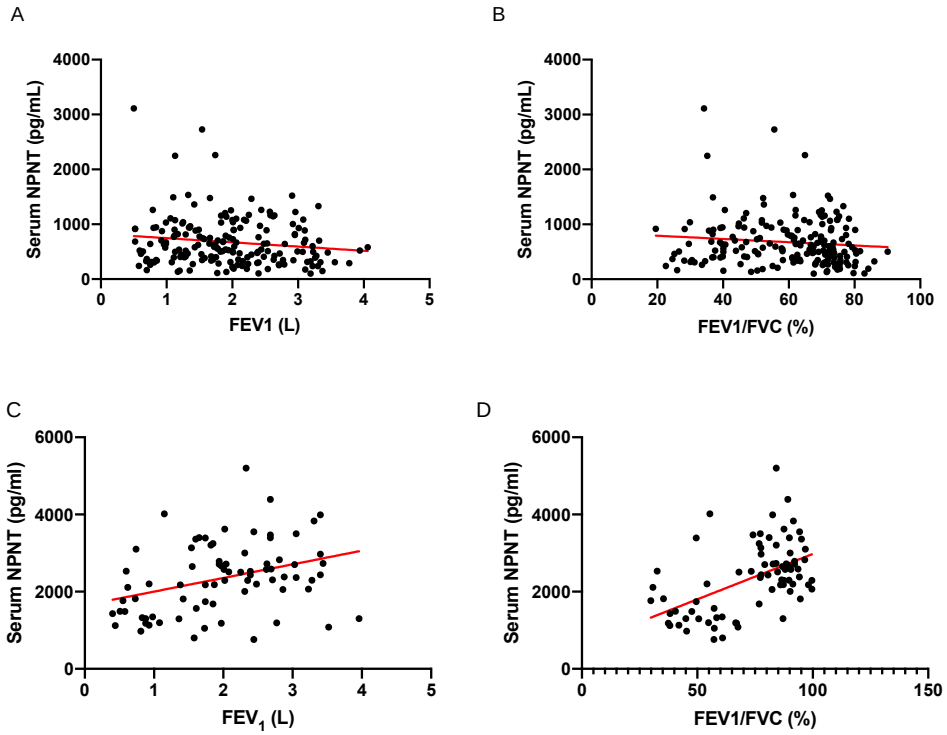


Figure 7-7 Correlation between serum NPNT levels and lung function, including forced expiratory volume in 1s (FEV_1) and FEV_1/FVC(%). A) and B) The serum NPNT levels were not associated with lung function in the UK cohort. C) and D) In the China Ningbo cohort, serum NPNT

levels was associated with FEV_1 (r^2=0.118, p=0.0022) and FEV_1/FVC(%) (r^2=0.2617, p<0.0001).

7.4.6 Identifying co-variates that influence NPNT levels of China Ningbo cohort

In China Ningbo Cohort, the age and height were negatively correlated with the levels of NPNT, while smoking pack/years and weight were not associated with serum NPNT protein expression Figure 7-8. Furthermore, the level of urea nitrogen was positively associated with serum NPNT level, and the level of creatinine blood and eGFR were not associated the expression of NPNT levels in serum (Table **7-4**, Figure 7-8, and Figure 7-9).

Factor	R square	Sig. P value	Correlated
Age	0.06048	0.0132	Yes
Height	0.04475	0.0453	Yes
Weight	0.00077	0.7937	-
Smoking pack/years	0.03221	0.2616	-
Urea nitrogen	0.04292	0.0428	Yes
Creatinine blood	0.00518	0.4859	-
eGFR	0.01003	0.3315	-

Table 7-4 Age, height, weight, smoking pack/years, Urea nitrogen, creatinine blood, and eGFR were analysed as covariate for China Ningbo Cohort.

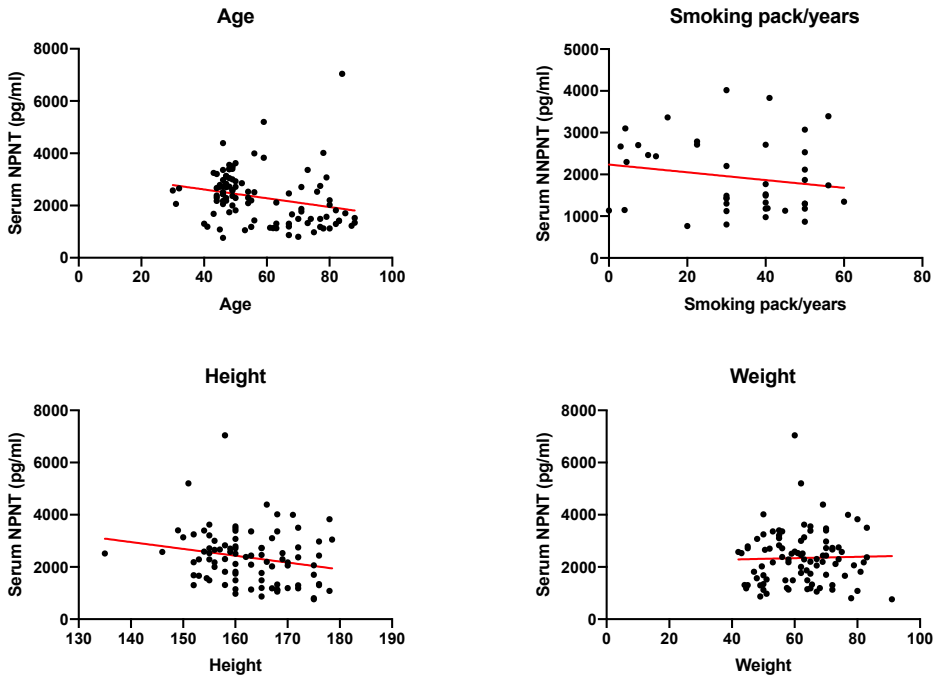


Figure 7-8 Dot plots for the linear variants tested for correlation. In the China Ningbo Cohort, the serum NPNT expression levels were negatively correlated for age (r^2=-0.06048, p<0.05), height (r^2=0.04475, p<0.05), while not significantly correlated for smoking pack/years and weight.

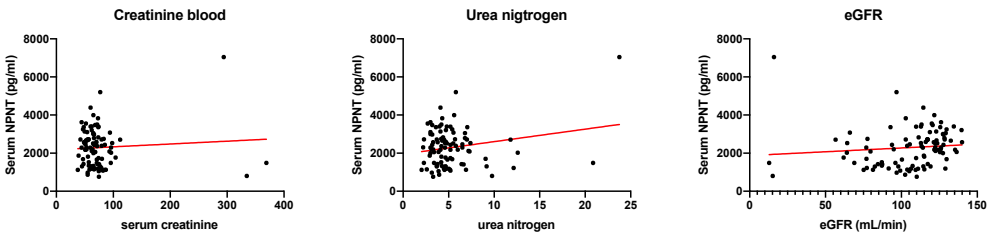


Figure 7-9 Dot plots for the linear variants tested for correlation. In the China Ningbo Cohort, the expression levels of serum NPNT were positively correlated for urea nitrogen (r^2=0. 04292, p<0.05), while not significantly correlated for creatinine blood and eGFR. eGFR, estimated glomerular filtration rate**.**

7.4.7 Serum NPNT levels are not associated with smoking

As shown in Figure **7**-**10**, serum NPNT expression were not correlated with smoking pack/years in either the China Ningbo cohort or the UK Nottingham cohort.

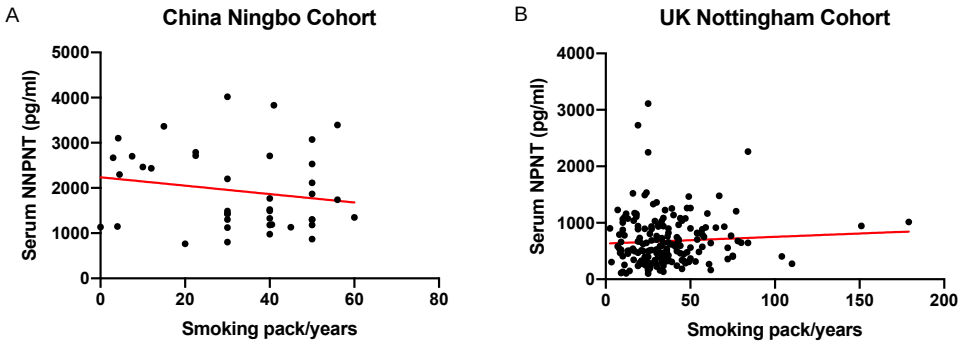


Figure 7-10 Dot plots for the test of correlation between serum NPNT levels and smoking pack/years. The serum NPNT levels were not correlated with smoking in both cohorts. A) China Ningbo cohort with r^2=-0.032, p=0.262, B) UK Nottingham cohort with r^2=0.004, p=0.395.

7.5 Discussion

In this chapter I aimed to identify whether serum NPNT levels is regulated in stable or exacerbation COPD patients and whether it is correlated with measures of lung function, such as FEV_1, FEV_1/FVC. The main findings identified that serum NPNT levels of stable COPD patients were not significantly different when compared to healthy controls, while they were decreased in COPD exacerbation patients. NPNT, as a novel lung relevant ECM, plays a vital role in the development of organs (314). ECM is involved in cell migration, proliferation, adhesion, and differentiation in the healthy lung. On the other hand, ECM is degraded by ECM enzymes (such as elastase, hyaluronidases, chondroitinases and metalloproteinases) which could be elevated in exacerbations of COPD, hence potentially leading to a fall in NPNT levels (315). Moreover, recent studies demonstrated MMP-2, MMP-9, and TIMP-1 were considerably elevated in AECOPD (316) and increased levels of MMP-9 are associated with COPD exacerbation risk (175).

AECOPD is a common clinical problem and frequent exacerbations lead to declines in lung function and accelerated deterioration in patients with COPD. Therefore, AECOPD might contribute to either lung destruction or airway remodelling, for example through loss of elasticity in lung tissue. It would therefore be helpful to find an effective early biomarker of exacerbation to permit more personalised interventions. The initial data presented here showing that the serum level of NPNT in AECOPD groups was significantly decreased

compared to the healthy control group in the China Ningbo cohort, means NPNT appears to be involved in AECOPD and could potentially be used as a biomarker.

In the UK Nottingham cohort, the correlation analysis showed that the serum NPNT levels positively associated with age, although the age of the stable COPD group was older than healthy control subjects. However, Lee et al reported that NPNT levels in healthy volunteers were not dependent on age although there was a difference in age between the silicosis patients and healthy volunteers in their study (99). On the other hand, the patients' age appeared to be negatively associated with NPNT protein levels in the China Ningbo Cohort. This could however be because the older patients had more severe COPD and the decreased circulating NPNT protein seen in COPD exacerbation patients might explain this.

In the UK Nottingham Cohort, however, the serum NPNT levels were not associated with FEV1/FVC, and also not significantly statistically different between the 4 grades of COPD stages. Furthermore, the expression of serum NPNT levels is no different between combined mild+moderate group and severe+very severe. In contrast, there was a positive correlation between serum NPNT levels and FEV_1 in the China Ningbo Cohort in analyses including health individuals and COPD exacerbation patients. The reason for these results might be that the AECOPD patients were older than healthy controls, as well as the NPNT downregulated in COPD exacerbation patients. In a recent study, Lee et al found that the serum nephronectin negatively correlated with

%FEV_1, but had no significant correlation with FVC (99). Thus, the relationship between serum NPNT levels and lung function overall remains unclear and still need further study.

Interestingly, the most recent GWAS analyses have suggested important roles for the ECM pathway and elastic fibre pathways in determining lung function. Wain and colleagues integrated the 234 causal genetic variants identified in one study with expression quantitative trait loci (eQTL) and identified enrichment for several ECM-related genes, including *NPNT*, *TGFB2*, *LTBP4*, *MFAP2*, and *MMP15* (44). More recently, another study found enrichment of pathways of GWAS of genetic variants with elastic fibers and ECM organization (86). Thus, disturbed ECM development and homoeostasis appears to be important in the genetic control of COPD remodelling.

In this chapter, I have analysed the data using a multiple linear regression analysis which showed there was not any significantly correlation with age, height, weight, and smoking pack/years from 2 cohort centres. In the Nottingham cohort there was a relationship between eGFR and NPNT levels, but as this was not seen in the Ningbo cohort. I did not include eGFR as a variable in the analysis.

The serum levels of China Ningbo Cohort was approximately 2-fold higher than in the UK Nottingham Cohort subjects, which might due to the different stages of COPD, or difference due to ethnicity in subjects from China or the UK. Furthermore, the serum level of NPNT was lower in this study than reported by Watany (105), while nearly 10-old higher than levels seen in a self-developed ELISA kit by Kon et al. using

serum samples from silicosis patients (99). As mentioned above, there are significant challenges of studying blood biomarker due to poor replication of results from different cohorts or samples.

There are several limitations to the work described in this chapter. Firstly, the sample numbers are small and recruited from 2 single centre studies of COPD stable and exacerbation patients, respectively. Secondly, we have not collected the matched serum from the stable stage of COPD (or at convalescence stage after exacerbation) to compare healthy or exacerbation individuals. Thirdly, it would be helpful to look at longitudinal change in NPNT levels in individual patients based on their clinical condition.

8. General Discussion

8.1 Hypothesis and Aims

The aim of the work described in this thesis was to gain further understanding of the potential role of nephronectin (NPNT) in lung pathophysiology. Nephronectin is an ECM protein, and has been shown to be involved in a range of processes in development, particularly in kidney and bone homeostasis (93, 106), and it has a range of functions in the kidney, bone, tooth, muscle and tumour (95, 114, 317, 318). More recent studies reported that NPNT can be regulated by factors such as TGF-β1, oncostatin M (OSM), tumour necrosis factor (TNF)-α, interleukin (IL)-1β and fibroblast growth factors (FGF)-2 (108-113, 318).

The main evidence suggesting a role for NPNT in the lung comes from GWAS studies. Studies from 2010 and later identified that genetic variation at the chromosome 4 (NPNT) locus is associated with both lung function variability and the risk of developing COPD (10, 79). The credible variant set implicates NPNT itself as the likely causal gene underlying this association.

Given the above evidence, the main hypothesis explored in this thesis is that NPNT is implicated in the developmental and homeostasis of lung and/or structurally relevant cells. Furthermore, I hypothesised that altered NPNT expression might contribute to the airway inflammation response and/or cell proliferation. I also hypothesised that serum NPNT levels might vary in exacerbation COPD compared with healthy controls.

Therefore, the aims of the work described in this thesis were:

a) To investigate the profile of NPNT mRNA and protein in different airway structural cell lines and lung tissue from individuals with COPD and healthy controls.

b) To identify the contribution of targeting NPNT (through loss of function or reduced expression) to downstream gene expression and potential pathways involved in human bronchial epithelial cell homeostasis.

c) To explore the expression of NPNT protein in serum of stable or acute exacerbation patients with COPD and healthy controls.

8.2 Main Findings

The data reviewed in this thesis identified that SNP rs34712979 located in *NPNT* locus is the most significant genetic variant which is associated with the FEV_1/FVC signal at this locus. There is a second variant, rs17331332 which is in close linkage disequilibrium with the sentinel variant. SNP rs17331332 and rs34712979 are both eQTL for NPNT. The effect allele G of variant rs34712979, which is associated with lower expression of NPNT, is also associated with COPD risk, and positively associated with weight in PheWAS (159). Taken together, the above evidence suggests that altered expression of NPNT may contribute to the COPD risk or lung function decline.

NPNT is highly expressed in lung tissue, and was detected in basal bronchial epithelial cells, differentiated HBECs, HASMs and the BEAS-2B cell line. The high expression of NPNT in human foetal and adult lung was confirmed by RNA-seq and antibody-based detection using the human protein atlas database, and immunohistochemistry. I found that *NPNT* mRNA expression intensity increased steadily with increasing foetal lung development across the Pseudoglandular and into the Canalicular stages of growth. Furthermore, the expression of NPNT protein in lung tissues from healthy control donors and COPD patients was detected by immunohistochemistry. NPNT protein is expressed in the alveolar and fibro-elastic layers, and especially in pneumocytes, fibroblasts from healthy and COPD donors. *NPNT* mRNA *was* decreased in individuals with COPD who smoked compared with

smokers who had no history of airway obstruction in an analysis performed in the GEO dataset.

In order to try and identify a functional role for NPNT in lung relevant cells, NPNT KD in primary bronchial epithelial cells was successfully established for Microarray and RNA sequencing studies. This work identified that NPNT is functional in HEBCs and can influence many pathways and genes expression relevant to epithelial cell homeostasis. Because Microarray experiments appeared to have greater variability, and crucially did not find significant NPNT KD following siRNA treatment, I based the majority of my analyses on the RNA-seq data set where NPNT knockdown was confirmed. There were 5 DEGs in common between RNA-seq and Microarray analysis. Genes upregulated by NPNT specific RNAi KD included VANGL2 and CBX1, and downregulated genes included CDK6 and CCDC68, which play a crucial role in cell cycle regulation in HBECs.

VANGL2 is involved in embryonic development (215) and protein levels are regulated by integrin αv and fibronectin, but not vitronectin (217), and integrin αv interacts with *VANGL2* to regulate matrix metalloproteinase activity and cell adhesion (218). CBX1 was significantly upregulated after NPNT knockdown, and CBX1 mRNA and protein are highly expressed in lung tissue. Furthermore, CDK6 is a key regulator to promote transition from G1 to S phase in the cell cycle and CCDC68 is involved in assembly of the centriole, a key component in cell cycle progression (220). Overall, the findings presented here

suggest that NPNT probably plays a key role in airway remodelling and lung homeostasis.

Gene Set Enrichment Analysis (GSEA) revealed that the interferon alpha/gamma response, IL6-JAK-STAT3 signalling, Notch signalling and P53 signalling pathway were upregulated following NPNT knockdown suggesting NPNT can regulate the inflammation response in HBECs. Moreover, GSEA demonstrated that E2F targets, Myc targets, G2M checkpoint, and mitorc1 signalling were downregulated by NPNT knockdown. This shows that downstream gene sets/ pathways involved in cell cycle or cell proliferation are downregulated by NPNT knockdown by siRNA. These findings therefore support the hypothesis that NPNT contributes to the inflammation response and airway remodelling in human lung or cell homeostasis.

In addition, I extended these studies by assessing the results of NPNT knockdown in BMI-1-engineered HBECs by electroporation-based CRISPR/Cas9 genome editing technology. I successfully established the protocol of a fast, robust, cloning-free and high efficiency electroporation-based CRISPR/Cas9 genome editing system (138). NPNT knockout regulated many downstream genes and implicated many pathways, including the inflammation pathway, protein binding, membrane and cell adhesion.

Interestingly, 7 DEGs (including NPNT) were consistent up- or down-regulated after NPNT knockdown by RNAi or CRISPR/Cas9, respectively. There were 3 up-regulation DEGs. TSTD3 is highly

expressed in lung, but the functions in COPD still need further study (188). CSGALNACT2 is expressed in Zebrafish early development (267) and identified susceptibility loci for Hirschsprung disease (319). FAM43A is expressed in triple negative breast cancer (268) and SNPs located within FAM43A associated with autism and learning difficulties (269). On the other hand, there were also 3 down-regulated DEGs. FLG was associated with lung cancer (271). THBS1 encodes a protein which can bind to fibrinogen, fibronectin, laminin, type V collagen and integrin $\alpha V\beta 1$ (188) and THBS1 is a promising therapeutic target of lung cancer (272). ACTG1 involves in epithelial mesenchymal transition via the MAPK/ERK pathway in prostate cancer (274) and increases in lung cancer cell lines (275). Given the current lack of evidence about how these 6 genes might be associated with COPD or emphysema this is an obvious area for further study.

Next, I assessed the potential value of measuring serum NPNT protein levels from two cohorts, the UK Nottingham cohort and the China Ningbo cohort. In the UK Nottingham Cohort, there was no difference in NPNT levels between healthy control and stable COPD patients. The serum NPNT expression levels were weakly positively correlated with age but not significantly correlated for smoking pack/years, height, weight or lung function. Levels of serum NPNT were also positively correlated for creatinine blood, and therefore not surprisingly negatively correlated with eGFR, but not significantly correlated for urea nitrogen and UACR. Interestingly, serum NPNT levels were decreased in COPD exacerbation patients from the China Ningbo Cohort. In this cohort, a

significant correlation has found between serum NPNT levels and FEV_1 and FEV_1/FVC. As above evidence supports that dysregulated NPNT levels in serum of COPD might be associated with pathological conditions and could exacerbate disease progression. Interestingly, baseline NPNT levels were found to be higher in the Chinese population than in the Nottingham cohort suggesting there may be racial differences in control of serum NPNT.

8.3 Study Limitations

In these studies, the SNPs analysed were those present on the GWAS platform, but without further studies it is difficult to be certain that the key SNPs studied are the main causal variants at this locus. There are also two other genes at this locus, GSTCD and INTS12, both of which are expressed in the lung, and it is possible that these genes may also contribute to abnormal lung function as a consequence of genetic variation at this locus. The role of GSTCD has been investigated by targeted knockdown (using siRNA) in HBECs and these studies revealed it may be involved in airway biology (165). Furthermore, GSTCD was found to have a potential function in mediating inflammatory signals in studies using precision cut lung slice from GSTCD knockout mice (166). INTS12 appears to regulate key pathways affecting translation through affecting the expression of genes belonging to protein synthesis pathways (130). Homozygous INTS12 knockout mice had a lethal in utero phenotype.

To study NPNT, I used 3 different approaches to NPNT knockdown, and this has highlighted potential limitations of the different techniques. In particular, the variability seen using Microarray data and the lack of significant NPNT knockdown when assessed by Microarray rather than RT-PCR suggests this approach is less reliable, although there were common genes identified between the different approaches.

I have successfully achieved to KO NPNT in BMI-1 cells by CRISPR/Cas9, and identified many DEGs and pathways implicated in

cell homeostasis. However, a limitation of the current study is that experiment just detected NPNT mRNA by qRT-PCR in bulk cells, rather than identified genotypes of cells by MiSeq. From the RNA-seq data analyses, NPNT expression was not significantly different following in experiments using two gRNAs (G1 and G2) but was reduced using the third gRNA (G3). This demonstrates that care has to be used when using a bulk approach to CRISPR/Cas9. A second limitation is that I assessed NPNT function in basal epithelial cells rather than in ALI cultures which better represent the reality and complex environment of the airway. Whether or not it is possible to generate gene edited NPNT KO BMI-1 cells successfully differential at ALI remains to be explored. If feasible, this would for example facilitate experiments assessing epithelial morphology imaging, barrier integrity (e.g., transepithelial electrical resistance, TEER) and ciliation ability.

The serum NPNT levels from two cohorts related to COPD/AECOPD was limited by a number of factors. Firstly, the sample number is small and recruited from single centres for the stable COPD subjects and exacerbation patients, respectively. Secondly, I have not collected the matched serum from the stable stage of COPD (or at convalescence stage after exacerbation) to compare healthy or exacerbation individuals. Thirdly, it would be helpful to look at longitudinal change in NPNT levels in individual patients based on their clinical condition.

8.4 Future Work

Future work should focus on addressing the key topic of clarifying the functional role for NPNT in human airway bronchial epithelial cells and how the genes or pathways implicated contribute to the underlying mechanism of COPD. As discussed above, functional studies using BMI-1 cells or primary HBECs cultured at ALI could be used to assess epithelium morphology eg by imaging, barrier integrity assays, or ciliation ability. This work will also require investigation of the interaction between NPNT and other ECM proteins including matrix metalloproteinases (MMPs), or integrin receptors. Clinically, the relationship between serum NPNT levels and lung functions needs further validation in larger cohorts.

Finally, it is worth considering whether or not NPNT is a viable therapeutic target in the airway. The fact that it is regulated during exacerbations of COPD suggest it may play a role in remodelling in the context of COPD exacerbations, but it remains unclear if inhibition of NPNT would be likely to be beneficial in this clinical setting. In addition, the fact that NPNT is expressed outside the lung and plays an important role in other organs (for example the kidney) means that any therapeutic intervention targeting NPNT in patients with lung disease would need to be carefully assessed for unwanted side effects in other organ systems.

9. Appendix

9.1 List of reagents

Material/Reagent	Supplier	Catalogue Code
16-well Nucleocuvette Strip	Lonza	V4XP-3032
18s PDAR	Thermo Fisher Scientific	4310893E
1 Kb Plus DNA Ladder	Thermo Fisher Scientific	10787018
Alt-R® Cas9 Electroporation Enhancer, 10 nmol	IDT	1075916
Alt-R® CRISPR-Cas9 tracrRNA	IDT	1072534
Alt-R® CRISPR-Cas9 crRNA (custom made oligo)	IDT	222880247
Alt-R® S.p. Cas9 Nuclease V3	IDT	1081059
Anti-beta actin antibody	Abcam	Ab8227
Anti-NPNT antibody produced in rabbit	Sigma Aldrich	HPA003711
Anti-Nephronectin antibody	Abcam	Ab64419
Anti-rabbit IgG polyclonal produced in goat	Sigma Aldrich	F0382
DNase I, Amplification Grade	Thermo Fisher Scientific	18068015
Dimethyl sulfoxide (DMSO)	Sigma Aldrich	154938
Dulbecco's Modified Eagle's Medium (DMEM) - high glucose	Sigma Aldrich	D5796
Ethanol	BDH	K3954
GenElute™ Mammalian Total RNA Miniprep Kit	Sigma Aldrich	RTN70- 1KT
Histoclear	Raymond A Lamb	C78-G
Hydrocortisone stock solution	Stem Cell Technologies	7925
MTT (Thiazolyl Blue Tetrazolium Bromide 98%)	Sigma Aldrich	R8755

Nephronectin (NPNT) ELISA kit	Cloud-Clone Corp	SEH522Hu
NPNT PDAR (Hs01568131_m1)	Thermo Fisher Scientific	4351368
NPNT (Human)- unique 27mer siRNA duplexes	ORIGENE	SR316838
Nuclease Free Duplex Buffer	IDT	11-01-03-01
P3 Primary Cell 4D-NucleofectorTM X Kit S	Lonza	V4XP-3032
Phosphate buffered saline (Dulbecco A) PBS tablets	Oxoid	BR0014
Phosphate buffered Saline (PBS) with Calcium and Magnesium	Sigma Aldrich	D8662
Peroxidase-Blocking Solution	Dako	S202386-2
Pneuma cult-ex media	Stem Cell Technologies	5009
siRNA and miRNA transfection reagent, INTERFERin®	VWR international, UK	409-10
Sterile Water(bottled)	Baxter	UKF7114
Superscript First Strand Synthesis System for RT-PCT kit	Invitrogen	11904018
TaqMan® Gene Expression Master Mix	Thermo Fisher Scientific	4369016
Trypan Blue	Sigma Aldrich	T-8154
Trypsin-EDTA solution (1x)	Sigma Aldrich	T3924
Trypsin Inhibitor	Invitrogen	17075-029
Water, for molecular biology, DEPC-treated and sterile filtered	Sigma Aldrich	95284-100ML

9.2 R script used for Microarray data

9.2.1 Array pre-processing

Bioconductor contains a number of packages, such as *limma* and *oligo*. The *oligo* package helps to read and organise information from Affymetrix Human Gene 2.1 ST Array Strip Microarray CEL files*. The package limma* is built primarily for spotted array data. Quality control (QC) is a first step. The R environment provides packages for visualization of distribution statistical QC, such as box plots.

This is illustrated by the following command, which generates image of the box plots.

The next step is normalization in analysing spotted array. The command is:

```
normData <- rma(rawData)
normData
dim(normData)
boxplot(normData)
```

```
Library(oligo)
untar("C:/Users/msxgq/OneDrive - The University of
Nottingham/R/Genechip_CEL/genechip.tar",exdir="genechip/CEL")
list.files("genechip/CEL")
celfiles <-list.files("genechip/CEL", full =TRUE)
```

celfiles

```
rawData <- read.celfiles(celfiles)
```

```
rawplot <- boxplot(rawData,target="core")
```

rawplot

Further analysis could be done using the *limma* package, which has advantages for designing matrices for spotted array experiments. The function *lmFit*() applies design matrix information together and estimates of relative mRNA abundance in each sample. The function *eBayers*() uses estimates of probability of differential expression for each gene.

Codes as following:

```
suppressMessages(library(limma))
group_list=c(rep("Un",3),rep("Scr",3),rep("siRNA_a",3),rep("siRNA_b",3
))
group_list
design <- model.matrix(~0+factor(group_list))
colnames(design)=levels(factor(group_list))
rownames(design)=colnames(exprSet)
design
```

step 1

```
fit <- lmFit(y2_exprSet,design)
contrast.matrix <- makeContrasts(siRNA_b-Scr, levels = design)
contrast.matrix
```

Step 2

```
fit2 <- contrasts.fit(fit,contrast.matrix)
fit2 <- eBayes(fit2)
```

Step 3

```
tempOutput = topTable(fit2,coef = 1,n=Inf)
nrDEG = na.omit(tempOutput)
head(nrDEG)
```

9.2.2 Annotation

Affymetrix Human Gene 2.1 ST Array Strip Microarray CEL files were annotated by R Bioconductor (https://www.bioconductor.org/) package (hugene21sttranscriptcluster.db) and RMA-normalized.

```
ls("package:hugene21sttranscriptcluster.db")
summary(hugene21sttranscriptclusterSYMBOL)
```

```
ids <- toTable(hugene21sttranscriptclusterSYMBOL)
head(ids)
```

```
filter(ids,symbol=="TP53")
filter(ids,symbol=="NPNT")
filter(ids,symbol=="ZNF93")
filter(ids,symbol=="ZNF629")
```

```
length(unique(ids$symbol))
table(sort(table(ids$symbol)))
plot(table(sort(table(ids$symbol))))
```

9.3 RNA-seq trimming and alignment

The code in this chapter was written by **Robert Hall, from Division of Respiratory Medicine, University of Nottingham**.

This RNA-seq pipeline uses Trimmomatic to trim adaptor sequences from fastq files as well as trim poor quality sequences. It then uses Star to align files to the genome and produce BAM files and count files. I then used Edge R in R studio to determine gene expression.

Transfer files

Fastq files were transferred as described into the desired directory on the HPC (drag and drop) using WinSCP. These were then used to generate the **Snakefile**, **config.yaml** and **sub_trim.sh** files generated from Alen's Script by Sam Cox (HPC team).

Concatenate files from multiple lanes of sequencing

The result should be one forward and one reverse fastq.gz file for each sample. The zcat command used to combine multiple lanes of sequencing was:

```
zcat file1_R1.fastq.gz file2_R1.fastq.gz file3_R1.fastq.gz | gzip >
R1.fastq.gz
```

```
rule zcat:
        input:
                "{sample}L1.fastq.gz",
                "{sample}L2.fastq.gz"
```

```
        output:
                "{sample}.fq.gz"
        shell:
                "zcat {input} | gzip>{output}"
```

The “Module load”, “source” and “conda activate” lines of code load the respiratory medicine environment set up in the HPC.

```
#!/usr/bin/bash
#SBATCH --time=72:00:00
#SBATCH --job-name=nepal2concat
#SBATCH --mem=100000
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=28

module load anaconda-uon/3
source /gpfs01/software/miniconda/miniconda3/etc/profile.d/conda.sh
conda activate /gpfs01/software/conda-extra/respmed

cd /gpfs01/home/mszrjh/Nepal2

snakemake -p 12_1B.fq.gz --cores 28
snakemake -p 12_1P.fq.gz --cores 28
snakemake -p 12_1T.fq.gz --cores 28
snakemake -p 12_1L.fq.gz --cores 28
```

```
snakemake -p 12_1I.fq.gz --cores 28

snakemake -p 12_1A.fq.gz --cores 28
```

Download reference genome and primary assemblies

To do this, the terminal was opened in the HPC SOP and the command typed in: (exact link may vary based on build being used) -> this downloaded the primary assembly and reference genome.

wget http://ftp.ensembl.org/pub/grch37/current/gtf/homo_sapiens/Homo_sapiens.GRCh37.87.gtf.gz

wget http://ftp.ensembl.org/pub/grch37/current/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37.dna.primary_assembly.fa.gz

Adaptor sequence file

The sequencing provider provided these sequences.

```
>RNA-SeqBarcodeAdapter1

AGATCGGAAGAG

>RNA-SeqBarcodeAdapter1_rc

CTCTTCCGATCT
```

Config file

fa: and gtf: are the two files downloaded with wget – these were changed this to my directory.

GRC_dir is output – again changed to the directory with GRCh37_STAR as the suffix.

Trim_fa: is the link to the adapter file previously explained.

```
fa:
"/gpfs01/home/mszrjh/Genomes/37/Homo_sapiens.GRCh37.dna.primary_assembly.fa"
gtf: "/gpfs01/home/mszrjh/Genomes/37/Homo_sapiens.GRCh37.87.gtf"
star_genome_ram_limit: 90000000000
GRC_dir: "/gpfs01/home/mszrjh/Nepal2/GRCh37_STAR"
GRC_names: ["Genome",
            "Log.out",
            "SA",
            "SAindex",
            "chrLength.txt",
            "chrName.txt",
            "chrNameLength.txt",
            "chrStart.txt",
            "exonGeTrInfo.tab",
            "exonInfo.tab",
            "geneInfo.tab",
            "genomeParameters.txt",
            "sjdbInfo.txt",
            "sjdbList.fromGTF.out.tab",
            "sjdbList.out.tab",
            "transcriptInfo.tab"]
trim_fa: "/gpfs01/home/mszrjh/Nepal2/contaminant_list_inc_RC.fa"
```

Snakemake

Snakemake contains the commands to run Trimmomatic and star.

Further information can be found in the snakemake tutorial here ->

https://snakemake.readthedocs.io/en/stable/tutorial/tutorial.html

Snakefile script

```
configfile: "config.yaml"
rule all:
  input: config['GRC_dir']+"/Genome"


rule star_genome:
  input:
    fa=config['fa'],
    gtf=config['gtf']
  output:
    expand(config['GRC_dir']+"/{name}", name=config['GRC_names'])
  params:
    out_dir=config['GRC_dir'],
    ram_limit=config['star_genome_ram_limit']
  threads: 28
  shell:
    "STAR "
    "--runThreadN {threads} "
    "--runMode genomeGenerate "
    "--genomeDir {params.out_dir} "
    "--genomeFastaFiles {input.fa} "
```

```
    "--sjdbGTFfile {input.gtf} "
    "--sjdbOverhang 101 "
    "--limitGenomeGenerateRAM {params.ram_limit}"

rule trim:
  input:
    "{input_base}_1.fastq.gz",
    "{input_base}_2.fastq.gz"
  output:
    "{input_base}_trimmed_1.fq.gz",
    "{input_base}_trimmed_1np.fq.gz",
    "{input_base}_trimmed_2.fq.gz",
    "{input_base}_trimmed_2np.fq.gz"
  params:
    trimming_lib=config['trim_fa']
  threads: 28
  shell:
    "trimmomatic PE "
    "-threads {threads} "
    "-phred33 "
    "-trimlog {wildcards.input_base}.trimlog "
    "{input} "
    "{output} "
    "ILLUMINACLIP:{params.trimming_lib}:2:30:10 "
    "LEADING:3 "
```

```
    "TRAILING:3 "
    "SLIDINGWINDOW:4:15 "
    "MINLEN:36 "

rule star_mapping:
  input:
    fq1="{input_base}_trimmed_1.fq.gz",
    fq2="{input_base}_trimmed_2.fq.gz",
    gen=config['GRC_dir']+"/Genome"
  output:
    "{input_base}Aligned.sortedByCoord.out.bam",
    "{input_base}Log.final.out",
    "{input_base}Log.out",
    "{input_base}Log.progress.out",
    "{input_base}ReadsPerGene.out.tab",
    "{input_base}SJ.out.tab"
  params:
    genome_dir=config['GRC_dir']
  threads: 28
  shell:
    "STAR "
    "--runThreadN {threads} "
    "--genomeDir {params.genome_dir} "
    "--readFilesIn {input.fq1} {input.fq2} "
    "--outFileNamePrefix {wildcards.input_base} "
```

```
    "--outSAMstrandField intronMotif "
    "--readFilesCommand zcat "
    "--outSAMmapqUnique 50 "
    "--outSAMtype BAM SortedByCoordinate "
    "--outSAMattrRGline ID:{wildcards.input_base}-RNAseq
SM:{wildcards.input_base} LB:{wildcards.input_base}-3RNAseq
PL:Illumina "
    "--quantMode GeneCounts
```

Shell script

```
#!/usr/bin/bash
#SBATCH --time=72:00:00
#SBATCH --job-name=nepal2
#SBATCH --mem=100000
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=28

module load anaconda-uon/3
source /gpfs01/software/miniconda/miniconda3/etc/profile.d/conda.sh
conda activate /gpfs01/software/conda-extra/respmed

cd /gpfs01/home/mszrjh/Nepal2/fastq
snakemake -p 12_1BAligned.sortedByCoord.out.bam --cores 2
```

```
sbatch myjob.sh
```

Prepping files for analysis

STAR gives a number of output files including a .bam file. However, it also usefully gives a count file for each sample which can be used for analysis in R. This is the **ReadsPerGene.out.tab** file.

mv *ReadsPerGene.out.tab /gpfs01/home/mszrjh/Nepal1/concat/reads

This then needed to be combined with individual count files into a matrix for edgeR.

N_unmapped	449141	449141	449141
N_multimapping	1153330	1153330	1153330
N_noFeature	424001	30556910	620388
N_ambiguous	341304	3 45621	1528870
ENSG00000223972	0	0	0
ENSG00000227232	116	0	117
ENSG00000243485	0	1	0
ENSG00000237613	0	0	0
ENSG00000268020	0	0	0
ENSG00000240361	0	0	0
ENSG00000186092	0	0	0
ENSG00000238009	0	0	2
ENSG00000239945	0	0	0
ENSG00000233750	0	3	0
ENSG00000237683	46	0	46
ENSG00000268903	0	0	0
ENSG00000269981	0	0	0

The first four lines are totals and the columns are ID, total reads, reads mapped to forward strand, and reads mapped to the reverse strand. This script creates a rule for all ReadsPerGene.out.tab files to take the first part of the name and creates a new file called ReadsPerGene.out.tab.counts by removing rows 1-4 and taking row 4, which contains the counts.

```
cd ../    # make sure you're in the dir above reads
mkdir counts
for x in counts/*ReadsPerGene.out.tab; do \
    s=`basename $x | cut -f1`
    echo $s
    cat $x | tail -n +5 | cut -f4 > 04-Counts/$s.count
done
```

Each file will now just be a list of numbers, so I needed to add back in the Ensembl IDs which are in column 1. To do this I made a txt doc of column 1.

```
tail -n +5
pathtoReadsPerGene.out/sampleexample_ReadsPerGene.out.tab | cut
-f1 > geneids.txt
head geneids.txt
```

and pasted the gene IDs to a temporary file tmp.out and combined the counts together by

pasting geneids.txt counts/*.count > tmp.out

I then made a sample file to create a header as follows and called it samples.txt

```
gene
12_1A
12_1B
12_1I
12_1L
12_1P
12_1T
```

This code rotates the samples.txt

```
cat samples.txt | cut -d_ -f1 | paste -s > header.txt
```

Samples were then added to the tmp.out file which contains the matrix to create all_counts.txt

```
cat header.txt tmp.out > all_counts.txt
```

All_counts.txt were then moved to a PC for analysis in R studio.

9.4 R script used for visualisation of RNA sequencing

9.4.1 Data packaging

Set up

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install(version = "3.13")
```

install packages

```
BiocManager::install("limma")

BiocManager::install("Glimma")

BiocManager::install("edgeR")

BiocManager::install("AnnotationDbi")

BiocManager::install("org.Hs.eg.db")

BiocManager::install("EnsDb.Hsapiens.v75")

BiocManager::install("matrixStats")

BiocManager::install("oligoClasses")
```

install packages

```
install.packages("ggplot2")

install.packages("gplots")

install.packages("dplyr")

install.packages("ggrepel")

install.packages("ggfortify")
```

```
BiocManager::install("biomaRt")

BiocManager::install("XML")
```

library

```
library(biomaRt)

library(edgeR)

library(ggplot2)

library(gplots)

library(ggrepel)

library(ggfortify)
```

```
library(Glimma)
library(AnnotationDbi)
library(org.Hs.eg.db)
library(EnsDb.Hsapiens.v75)
library(dplyr)
```

Reading in count-data

all_counts.txt is the matrix of counts from each file produced on the HPC.

The all_counts.txt file was put in the directoryset at the top of the markdown

This code reads the table into R studio and assigns it to "counts"

```
counts <-read.table("all_counts.txt",header=T, row.names = 1)
head(counts)
```

Next I created a DGEL list using **EdgeR** - this contains all the information on the samples and gene names etc

```
DGEL <- DGEList(counts)
```

2) Organising sample information

Here the columns were named

number = experiment number

a = siRNA a

b = siRNA b

S = scrambled

ut = untransfected

then samplenames were assigned to the DGEL list

```
samplenames <- substring(colnames(DGEL), 2, 4)
samplenames
colnames(DGEL) <- samplenames
```

Assign each sample to a group or treatment, this is the order they appear in the columns.

Can also add donor information and passage information.

```
group <-
as.factor(c("A","B","S","UT","A","B","S","UT","A","B","S","UT","A","B","S"
,"UT","A","B","S","UT"))
DGEL$samples$group <- group
donor <- as.factor(c("1", "1","1","1","2","2","2","2","2","2","2","2","1",
"1","1","1","1", "1","1","1"))
DGEL$samples$donor <- donor
DGEL$samples
```

3) Organising gene annotations

Adding gene annotations to the *DGEL* list: our count files only had Ensembl identifiers. We did this using the *mapIds* funciton from the *AnnotationDbi* package from Bioconductor.

```
genes <- mapIds(EnsDb.Hsapiens.v75, keys=rownames(DGEL),
```

```
                keytype="GENEID", column="SYMBOL", multiVals="first")
genelist <- as.data.frame(genes)
DGEL$genes <- data.frame(ENSEMBL=rownames(DGEL),
SYMBOL=genes)
DGEL
```

9.4.2 Data pre-processing

1) Transformations from the raw-scale

I didn't use raw counts because libraries are sequenced at slightly different depths and will therefore have different counts. So, I transformed raw counts onto a scale that accounts for library size differences. Common transformations include CPM (counts per million), log2-CPM, RPKM, FPKM. CPM and Log-CPM were used here.

```
cpm <- cpm(DGEL)
lcpm <- cpm(DGEL, log=TRUE)
```

lookup mean and median library sizes

```
L <- mean(DGEL$samples$lib.size) * 1e-6
M <- median(DGEL$samples$lib.size) * 1e-6
c(L, M)
```

2) Removing genes that are low expressed

Filtering out lowly expressed genes

Looking at genes which have 0 expression across all samples - these can be removed

```
table(rowSums(DGEL$counts==0)==20)
```

I removed lowly expressed samples using the function below

```
keep <- filterByExpr(DGEL, group=group)
DGEL <- DGEL[keep,, keep.lib.sizes=FALSE]
dim(DGEL)
```

The next chunk of code produces a figure which shows unfiltered data vs filtered data.

```
lcpm.cutoff <- log2(10/M + 2/L)
library(RColorBrewer)
nsamples <- ncol(DGEL)
col <- brewer.pal(nsamples, "Paired")
par(mfrow=c(1,2))
plot(density(lcpm[,1]), col=col[1], lwd=2, ylim=c(0,0.26), las=2, main="",
xlab="")
title(main="A. Raw data", xlab="Log-cpm")
abline(v=lcpm.cutoff, lty=3)
for (i in 2:nsamples){
den <- density(lcpm[,i])
lines(den$x, den$y, col=col[i], lwd=2)
}
```

```
legend("topright", samplenames, text.col=col, bty="n")
lcpm <- cpm(DGEL, log=TRUE)
plot(density(lcpm[,1]), col=col[1], lwd=2, ylim=c(0,0.26), las=2, main="",
xlab="")
title(main="B. Filtered data", xlab="Log-cpm")
abline(v=lcpm.cutoff, lty=3)
for (i in 2:nsamples){
den <- density(lcpm[,i])
lines(den$x, den$y, col=col[i], lwd=2)
}
legend("topright", samplenames, text.col=col, bty="n")
```

3) Normalising gene expression distributions

Normalise data were generated using the TMM method

Normalisation by the method of trimmed mean of M-values (TMM) is performed using the *calcNormFactors* function in *edgeR*. When working with DGEList-objects, these normalisation factors are automatically stored in *x$samples$norm.factors*.

```
DGEL <- calcNormFactors(DGEL, method = "TMM")
DGEL$samples$norm.factors
```

4) Unsupervised clustering of samples

Here a multi-dimensional scaling (MDS) plot is generated. The plot shows similarities and differences between samples so we can see the extent to which differential expression can be detected before carrying

out further tests. I used *limma*, specifically the *plotMDS* function, to do this.

```
lcpm <- cpm(DGEL, log=TRUE)
col.group <- group
levels(col.group) <-  brewer.pal(nlevels(col.group), "Set1")
col.group <- as.character(col.group)
plotMDS(lcpm, labels=group, col=col.group)
title(main="A. Sample groups")
```

This can be used to see how the samples cluster on donor.

```
col.donor <- donor
levels(col.donor) <-  brewer.pal(nlevels(col.donor), "Set1")
col.donor <- as.character(col.donor)
plotMDS(lcpm, labels=donor, col=col.donor)
title(main="B. donor")
```

It was also possible to use *Glimma* to produce an interactive *MDSplot* - the output of which is an html file .

```
glMDSPlot(lcpm, labels=paste(group, donor),
          groups=DGEL$samples, launch=TRUE, folder = "MDSplot", html
= "MDSplot")
```

9.4.3 Differential expression analysis

1) I created a design matrix and contrasts:

```
design <- model.matrix(~group -1)
colnames(design) <- gsub("group", "", colnames(design))
design
```

Checked the number of samples in each group (5 in each)

```
colSums(design)
```

Set up contrasts (comparison groups)

```
contr.matrix <- makeContrasts(
  SvsUT = S-UT,
  AvsS = A-S,
  BvsS = B-S,
   levels = colnames(design))
contr.matrix
```

Blocked the donor

```
donor_block <-factor(paste0(donor))
```

In this method I ran voom twice (see -

https://support.bioconductor.org/p/59700/)

```
v <- voom(DGEL, design)
corfit <- duplicateCorrelation(v,design=design, block= donor_block)
v2 <- voom(DGEL, design, block = donor_block, correlation =
```

```
corfit$consensus)
fit <- lmFit(v2, design, block = donor_block, correlation =
corfit$consensus)
fit <- contrasts.fit(fit, contr.matrix)
efit <- eBayes(fit)
```

2）Accessing Differentially expressed genes

```
dg <- decideTests(efit)
summary(dg)
```

```
summary <- summary(dg)
```

```
write.csv(summary, file="summary_of_comparisons_NPNT.csv")
```

All results were placed into one large txt file

-1 = downregulated

1 = upregulated

0 = unchanged

To examine individual DE genes from top to bottom:

View DE genes using the *topTable* function in *limma* using eBayes.

```
Scr <- topTable(efit, coef=1, n=Inf)
A <- topTable(efit, coef=2, n=Inf)
B <- topTable(efit, coef=3, n=Inf)
```

Write results

```
write.csv(Scr, file="Scr_results.csv")
```

```
write.csv(A, file="A_results.csv")
write.csv(B, file="B_results.csv")
```

3） Graphical representations of differential expression results

plotMD (mean difference) plots the log-FC from the linear model fit against the average log-CPM. DE genes are highlighted.

```
plotMD(efit, column=1, status=dg[,1], main=colnames(efit)[1],
       xlim=c(-8,13))
```

```
plotMD(efit, column=2, status=dg[,2], main=colnames(efit)[2],
       xlim=c(-8,13))
```

```
plotMD(efit, column=3, status=dg[,3], main=colnames(efit)[3],
       xlim=c(-8,13))
```

Can also use *Glimma* to produce an interactive plot

```
glMDPlot(efit, coef=1, status=dg, main=colnames(efit)[1],
         side.main="SYMBOL", counts=lcpm, groups=group,
launch=FALSE, folder = "MDplots", html = "Scr")
glMDPlot(efit, coef=2, status=dg, main=colnames(efit)[2],
         side.main="SYMBOL", counts=lcpm, groups=group,
launch=FALSE, folder = "MDplots", html = "A")

glMDPlot(efit, coef=3, status=dg, main=colnames(efit)[3],
```

```
        side.main="SYMBOL", counts=lcpm, groups=group,
launch=FALSE, folder = "MDplots", html = "B")
```

4） Produce a heatmap using the following code (top 20 on p value, change to suit)

A heatmap

```
A.topgenes <- A$SYMBOL[1:20]
A.HM <- which(v$genes$SYMBOL %in% A.topgenes)
mycol <- colorpanel(1000,"blue","white","red")
heatmap.2(lcpm[A.HM,], scale="row",
   labRow=v$genes$SYMBOL[A.HM], labCol=group, cexRow = 0.7,
cexCol = 0.7,
   col=mycol, trace="none", density.info="none",
   dendrogram="column")
```

B heatmap

```
B.topgenes <- B$SYMBOL[1:100]
B.HM <- which(v$genes$SYMBOL %in% B.topgenes)
mycol <- colorpanel(1000,"blue","white","red")
heatmap.2(lcpm[B.HM,], scale="row",
   labRow=v$genes$SYMBOL[B.HM], labCol=group, cexRow = 0.2,
cexCol = 0.7,
   col=mycol, trace="none", density.info="none",
   dendrogram="column")
```

S heatmap

```
S.topgenes <- Scr$SYMBOL[1:100]
S.HM <- which(v$genes$SYMBOL %in% S.topgenes)
mycol <- colorpanel(1000,"blue","white","red")
heatmap.2(lcpm[S.HM,], scale="row",
  labRow=v$genes$SYMBOL[S.HM], labCol=group, cexRow = 0.2,
cexCol = 0.7,
  col=mycol, trace="none", density.info="none",
  dendrogram="column")
```

5） Volcano plots

First I created a file with only significant genes in (5% FDR) and log FC >1 or <-1, these were used to label the volcano plot.

```
A_label=A[which(A$adj.P.Val<0.05&(A$logFC<(-1)|A$logFC>(1))),]
B_label=B[which(B$adj.P.Val<0.05&(B$logFC<(-1)|B$logFC>(1))),]
S_label=Scr[which(Scr$adj.P.Val<0.05&(Scr$logFC<(-
1)|Scr$logFC>(1))),]
```

coloured = significant p value

labelled = significant p value and log FC >1 either direction

A volcano

```
A$Legend <- ifelse(A$adj.P.Val < 0.05&A$logFC < 0,
"Decreased",ifelse(A$adj.P.Val < 0.05&A$logFC > 0, "Increased", "Not
Sig"))
```

```
if(length(unique(A$Legend))>2){
  ggplot(A, aes(x = logFC, y = -log10(P.Value))) + # This line creates the axes plus background of the graph

    geom_point(aes(color = Legend)) + # This line plus previous coding line creates the volcano plot on to the graph with values coloured in three groups

    scale_color_manual(values = c("blue","red", "grey"))+ # Re-colours the plot to only highlight the significantly increased/decreased values

    theme_bw(base_size = 12) + theme(legend.position = "bottom") + # Changes the legend to below the graph

    geom_text_repel(data = subset(A_label, adj.P.Val < 0.05),
                    aes(label =SYMBOL),size = 3, box.padding = unit(0.35, "lines"),
                    point.padding = unit(0.3, "lines") ) }else{

                      ggplot(A, aes(x = logFC, y = -log10(P.Value))) +
                        geom_point(aes(color = Legend)) +
                        scale_color_manual(values = c("blue","red", "grey"))+
                        theme_bw(base_size = 12) + theme(legend.position = "bottom") +
```

```
                    geom_text_repel(data = subset(A_label, adj.P.Val <
0.05),
                                    aes(label =SYMBOL),size = 3, box.padding =
unit(0.35, "lines"),
                                    point.padding = unit(0.3, "lines") )


                }
ggsave("A_Volc.png", dpi = 300, width = 10, height = 5)
```

B volcano

```
B$Legend <- ifelse(B$adj.P.Val < 0.05&B$logFC < 0,
"Decreased",ifelse(B$adj.P.Val < 0.05&B$logFC > 0, "Increased", "Not
Sig"))
if(length(unique(B$Legend))>2){
  ggplot(B, aes(x = logFC, y = -log10(P.Value))) + # This line creates
the axes plus background of the graph

    geom_point(aes(color = Legend)) + # This line plus previous coding
line creates the volcano plot on to the graph with values coloured in
three groups

    scale_color_manual(values = c("blue","red", "grey"))+ # Re-colours
the plot to only highlight the significantly increased/decreased values

    theme_bw(base_size = 12) + theme(legend.position = "bottom") + #
Changes the legend to below the graph

    geom_text_repel(data = subset(B_label, adj.P.Val < 0.05),
                    aes(label =SYMBOL),size = 3, box.padding = unit(0.35,
"lines"),
                    point.padding = unit(0.3, "lines") ) }else{

                      ggplot(B, aes(x = logFC, y = -log10(P.Value))) +
```

```
                    geom_point(aes(color = Legend)) +

                    scale_color_manual(values = c("blue","red", "grey"))+

                    theme_bw(base_size = 12) + theme(legend.position =
"bottom") +

                    geom_text_repel(data = subset(B_label, adj.P.Val <
0.05),
                                    aes(label =SYMBOL),size = 3, box.padding =
unit(0.35, "lines"),
                                    point.padding = unit(0.3, "lines") )

                }
ggsave("B_Volc.png", dpi = 300, width = 10, height = 5)
```

S volcano plot

```
S<-Scr
S$Legend <- ifelse(S$adj.P.Val < 0.05&S$logFC < 0,
"Decreased",ifelse(S$adj.P.Val < 0.05&S$logFC > 0, "Increased", "Not
Sig"))
if(length(unique(A$Legend))>2){
  ggplot(S, aes(x = logFC, y = -log10(P.Value))) + # This line creates
the axes plus background of the graph
```

```
   geom_point(aes(color = Legend)) + # This line plus previous coding
line creates the volcano plot on to the graph with values coloured in
three groups



   scale_color_manual(values = c("blue","red", "grey"))+ # Re-colours
the plot to only highlight the significantly increased/decreased values



   theme_bw(base_size = 12) + theme(legend.position = "bottom") + #
Changes the legend to below the graph

   geom_text_repel(data = subset(S_label, adj.P.Val < 0.05),
                   aes(label =SYMBOL),size = 3, box.padding = unit(0.35,
"lines"),
                   point.padding = unit(0.3, "lines") ) }else{

                     ggplot(S, aes(x = logFC, y = -log10(P.Value))) +

                       geom_point(aes(color = Legend)) +

                       scale_color_manual(values = c("blue","red", "grey"))+

                       theme_bw(base_size = 12) + theme(legend.position =
"bottom") +
```

```
                    geom_text_repel(data = subset(S_label, adj.P.Val <
0.05),
                                    aes(label =SYMBOL),size = 3, box.padding =
unit(0.35, "lines"),
                                    point.padding = unit(0.3, "lines") )

                }
ggsave("S_Volc.png", dpi = 300, width = 10, height = 5)
```

6) To look at genes in common between groups A and B I generated a Venn diagram

[,2] and [,3] are the two comparison groups (AvS and BvS)

442 genes in common

```
ab.common <- which(dg[,2]!=0 & dg[,3]!=0)
length(ab.common)
```

To look at **upregulated** genes in common - here there were 256

```
vennDiagram(dg[,2:3], names = c("A", "B"), include = "up", cex = 1, lwd
= 1, circle.col= c(" salmon", "turquoise")) #only upregulated genes
```

Downregulated genes in common - 152

```
vennDiagram(dg[,2:3], names = c("A", "B"), include = "down", cex = 1,
lwd = 1, circle.col= c(" salmon", "turquoise")) # only down genes
```

Total genes in common - 442

This meant there were a number of genes which go in opposite directions with the two treatment.

```
vennDiagram(dg[,2:3], names = c("A", "B"), include = "both", cex = 1,
lwd = 1, circle.col= c(" salmon", "turquoise")) #genes in any directions
```

```
```Total genes in common - 442
This means we have a number of genes which go in opposite directions with the two treatment

vennDiagram(dg[,2:3], names = c("A", "B"), include = "both", cex = 1,
lwd = 1, circle.col= c(" salmon", "turquoise")) #genes in any directions
```

To generate a list of the genes in common:

```
A_sig=A[which(A$adj.P.Val<0.05),] #select significant genes from A
B_sig=B[which(B$adj.P.Val<0.05),] #select significant genes from B
S_sig=Scr[which(Scr$adj.P.Val<0.05),] #select significant genes from Scr
```

I created lists in order to view the genes in common

```
A_list <- paste(A_sig$SYMBOL) #make list of A significant genes
B_list <- paste(B_sig$SYMBOL) #B significant genes
A_B_genes <- list(A=A_list, B=B_list) #combine together
```

```
A_B_overlap <- calculate.overlap(A_B_genes) #calculate the overlap in these two lists of genes
```
```

To export this list of genes for interest：

```
overlap.table <- as.data.frame(A_B_overlap$a3)
colnames(overlap.table) <- c("SYMBOL")
```

I then added details to this list of genes, including logFC and adjusted p value.

```
A1 <- select(A_sig, ENSEMBL, SYMBOL, logFC, adj.P.Val)
B1 <- select(B_sig, SYMBOL, logFC,adj.P.Val)
```

```
overlap_full <- merge(A1, B1, by = "SYMBOL")
write.csv(overlap_full, file = "A_b_overlap.csv")
```

9.4.4 R script used for GSEA

To parse the files to remove blank gene names, and also separate gene names on the same row separated by ",", and merge duplicates:

Step 1. Install and load required packages.

```
library(dplyr)
library(tidyr)
```

Step 2. Remove the rows which are not gene names from data frame "GSEA". The first column name is "gene_short_name".

```
Parse1 <-GSEA[!GSEA$gene_short_name=="-",]
```

Step 3. Separate the row in which there are more than one gene separated by a comma (,) and give the same expression value to each new row.

```
Parse2 <-Parse1 %>% separate_rows(gene_short_name, sep=",")
```

Step 4. Remove duplicate gene name, average the FPKMs, and combine them to one row.

```
Parse3 <-Parse2 %>%
group_by(gene_short_name) %>%
summarize_each(funs(mean))
```

To save the table as (.txt)

```
write.table(Parse3,"gsea_parsed.txt",sep="\t")
```

References

1. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet* 2012; 380: 2095-2128.
2. Neumeier A, Keith R. Clinical Guideline Highlights for the Hospitalist: The GOLD and NICE Guidelines for the Management of COPD. *J Hosp Med* 2020; 15: 240-241.
3. Probert K, Miller S, Kheirallah AK, Hall IP. Developmental genetics of the COPD lung. *COPD Research and Practice* 2015; 10: 1-10.
4. Lokke A, Lange P, Scharling H, Fabricius P, Vestbo J. Developing COPD: a 25 year follow up study of the general population. *Thorax* 2006; 61: 935-939.
5. Silverman EK, Chapman HA, Drazen JM, Weiss ST, Rosner B, Campbell EJ, O'Donnell WJ, Reilly JJ, Ginns L, Mentzer S, Wain J, Speizer FE. Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. *Am J Respir Crit Care Med* 1998; 157: 1770-1778.
6. Hubert HB, Fabsitz RR, Feinleib M, Gwinn C. Genetic and environmental influences on pulmonary function in adult twins. *American Review Respiratory Disease* 1982; 125: 409-415.
7. Wilk JB, Walter RE, Laramie JM, Gottlieb DJ, T O'Connor G. Framingham Heart Study genome-wide association: results for pulmonary function measures. *BMC medical genetics* 2007; 8: S8.
8. Pillai SG, Ge D, Zhu G, Kong X, Shianna KV, Need AC, Feng S, Hersh CP, Bakke P, Gulsvik A, Ruppert A, Lodrup Carlsen KC, Roses A, Anderson W, Rennard SI, Lomas DA, Silverman EK, Goldstein DB. A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet* 2009; 5: e1000421.
9. Wilk JB, Chen T-h, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, Myers RH, Borecki IB, Silverman EK, Weiss ST. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS genetics* 2009; 5: e1000429.
10. Hancock DB, Eijgelsheim M, Wilk JB, Gharib SA, Loehr LR, Marciante KD, Franceschini N, van Durme YM, Chen T-h, Barr RG. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nature genetics* 2010; 42: 45-52.
11. Repapi E, Sayers I, Wain LV, Burton PR, Johnson T, Obeidat M, Zhao JH, Ramasamy A, Zhai G, Vitart V, Huffman JE, Igl W, Albrecht E, Deloukas P, Henderson J, Granell R, McArdle WL, Rudnicka AR, Wellcome Trust Case Control C, Barroso I, Loos RJ, Wareham NJ, Mustelin L, Rantanen T, Surakka I, Imboden M, Wichmann HE, Grkovic I, Jankovic S, Zgaga L, Hartikainen AL, Peltonen L, Gyllensten U, Johansson A, Zaboli G, Campbell H, Wild SH, Wilson JF, Glaser S, Homuth G, Volzke H, Mangino M, Soranzo N, Spector TD, Polasek O, Rudan I, Wright AF, Heliovaara M, Ripatti S, Pouta A, Naluai AT, Olin AC, Toren K, Cooper MN, James AL, Palmer LJ, Hingorani AD, Wannamethee SG, Whincup PH, Smith GD, Ebrahim S, McKeever TM, Pavord ID, MacLeod AK, Morris AD, Porteous DJ, Cooper C,

Dennison E, Shaheen S, Karrasch S, Schnabel E, Schulz H, Grallert H, Bouatia-Naji N, Delplanque J, Froguel P, Blakey JD, Team NRS, Britton JR, Morris RW, Holloway JW, Lawlor DA, Hui J, Nyberg F, Jarvelin MR, Jackson C, Kahonen M, Kaprio J, Probst-Hensch NM, Koch B, Hayward C, Evans DM, Elliott P, Strachan DP, Hall IP, Tobin MD. Genome-wide association study identifies five loci associated with lung function. *Nature Genetics* 2010; 42: 36-44.

12. Shrine N, Guyatt AL, Erzurumluoglu AM, Jackson VE, Hobbs BD, Melbourne CA, Batini C, Fawcett KA, Song K, Sakornsakolpat P, Li X, Boxall R, Reeve NF, Obeidat M, Zhao JH, Wielscher M, Weiss S, Kentistou KA, Cook JP, Sun BB, Zhou J, Hui J, Karrasch S, Imboden M, Harris SE, Marten J, Enroth S, Kerr SM, Surakka I, Vitart V, Lehtimäki T, Allen RJ, Bakke PS, Beaty TH, Bleecker ER, Bossé Y, Brandsma CA, Chen Z, Crapo JD, Danesh J, DeMeo DL, Dudbridge F, Ewert R, Gieger C, Gulsvik A, Hansell AL, Hao K, Hoffman JD, Hokanson JE, Homuth G, Joshi PK, Joubert P, Langenberg C, Li X, Li L, Lin K, Lind L, Locantore N, Luan J, Mahajan A, Maranville JC, Murray A, Nickle DC, Packer R, Parker MM, Paynton ML, Porteous DJ, Prokopenko D, Qiao D, Rawal R, Runz H, Sayers I, Sin DD, Smith BH, Soler Artigas M, Sparrow D, Tal-Singer R, Timmers P, Van den Berge M, Whittaker JC, Woodruff PG, Yerges-Armstrong LM, Troyanskaya OG, Raitakari OT, Kähönen M, Polašek O, Gyllensten U, Rudan I, Deary IJ, Probst-Hensch NM, Schulz H, James AL, Wilson JF, Stubbe B, Zeggini E, Jarvelin MR, Wareham N, Silverman EK, Hayward C, Morris AP, Butterworth AS, Scott RA, Walters RG, Meyers DA, Cho MH, Strachan DP, Hall IP, Tobin MD, Wain LV. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nature genetics* 2019; 5: 481-493.
13. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, Crapo R, Enright Pv, Van Der Grinten C, Gustafsson P. Standardisation of spirometry. *European respiratory journal* 2005; 26: 319-338.
14. Hughes J. Interpreting pulmonary function tests. *Breathe* 2009; 6: 102-110.
15. GOLD. The Global Initiative for Chronic Obstructive Lung Disease. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease (Updated 2019). 2019; http://goldcopd.org.
16. Nathell L, Nathell M, Malmberg P, Larsson K. COPD diagnosis related to different guidelines and spirometry techniques. *Respiratory research* 2007; 8: 89.
17. Kessler R, Partridge MR, Miravitlles M, Cazzola M, Vogelmeier C, Leynaud D, Ostinelli J. Symptom variability in patients with severe COPD: a pan-European cross-sectional study. *Eur Resp J* 2011; 37: 264-272.
18. Agusti A, Hedner J, Marin J, Barbé F, Cazzola M, Rennard S. Night-time symptoms: a forgotten dimension of COPD. *European Respiratory Review* 2011; 20: 183-194.
19. Miravitlles M, Guerrero T, Mayordomo C, Sánchez-Agudo L, Nicolau F, Segú JL. Factors associated with increased risk of exacerbation and hospital admission in a cohort of ambulatory COPD patients: a multiple logistic regression analysis. *Respiration* 2000; 67: 495-501.
20. Menezes AMB, Perez-Padilla R, Jardim JB, Muiño A, Lopez MV, Valdivia G, De Oca MM, Talamo C, Hallal PC, Victora CG. Chronic obstructive

pulmonary disease in five Latin American cities (the PLATINO study): a prevalence study. *The Lancet* 2005; 366: 1875-1881.
21. Wang C, Xu J, Yang L, Xu Y, Zhang X, Bai C, Kang J, Ran P, Shen H, Wen F. Prevalence and risk factors of chronic obstructive pulmonary disease in China (the China Pulmonary Health [CPH] study): a national cross-sectional study. *The Lancet* 2018; 391: 1706-1717.
22. Raherison C, Girodet P. Epidemiology of COPD. *European Respiratory Review* 2009; 18: 213-221.
23. Postma DS, Bush A, van den Berge M. Risk factors and early origins of chronic obstructive pulmonary disease. *The Lancet* 2015; 385: 899-909.
24. Guarascio AJ, Ray SM, Finch CK, Self TH. The clinical and economic burden of chronic obstructive pulmonary disease in the USA. *ClinicoEconomics and outcomes research: CEOR* 2013; 5: 235.
25. Kelly AM, Holdgate A, Keijzers G, Klim S, Graham CA, Craig S, Kuan WS, Jones P, Lawoko C, Laribi S. Epidemiology, treatment, disposition and outcome of patients with acute exacerbation of COPD presenting to emergency departments in Australia and South East Asia: An AANZDEM study. *Respirology* 2018; 23: 681-686.
26. Collaborators GCRD. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet Respiratory medicine* 2017; 5: 691-706.
27. Berg K, Wright JL. The pathology of chronic obstructive pulmonary disease: progress in the 20th and 21st centuries. *Archives of pathology & laboratory medicine* 2016; 140: 1423-1428.
28. Wright J, Churg A. Advances in the pathology of COPD. *Histopathology* 2006; 49: 1-9.
29. Fischer BM, Pavlisko E, Voynow JA. Pathogenic triad in COPD: oxidative stress, protease–antiprotease imbalance, and inflammation. *International journal of chronic obstructive pulmonary disease* 2011; 6: 413.
30. Kirkham PA, Barnes PJ. Oxidative stress in COPD. *Chest* 2013; 144: 266-273.
31. O'donnell R, Breen D, Wilson S, Djukanovic R. Inflammatory cells in the airways in COPD. *Thorax* 2006; 61: 448-454.
32. Hacievliyagil SS, Gunen H, Mutlu LC, Karabulut AB, Temel İ. Association between cytokines in induced sputum and severity of chronic obstructive pulmonary disease. *Respir Med* 2006; 100: 846-854.
33. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *The Lancet* 2007; 370: 765-773.
34. Kc B, Mahapatra PS, Thakker D, Henry AP, Billington CK, Sayers I, Puppala SP, Hall IP. Proinflammatory Effects in Ex Vivo Human Lung Tissue of Respirable Smoke Extracts from Indoor Cooking in Nepal. *Ann Am Thorac Soc* 2020; 17: 688-698.
35. Stoller JK, Aboussouan LS. α1-antitrypsin deficiency. *The Lancet* 2005; 365: 2225-2236.
36. Bossé Y, Lamontagne M, Gaudreault N, Racine C, Levesque MH, Smith BM, Auger D, Clemenceau A, Paré M, Laviolette L, Tremblay V, Maranda B, Morissette MC, Maltais F. Early-onset emphysema in a large French-Canadian family: a genetic investigation. *The Lancet Respiratory medicine* 2019; 7: 427-436.

37. Hall R, Hall IP, Sayers I. Genetic risk factors for the development of pulmonary disease identified by genome-wide association. *Respirology* 2019; 24: 204-214.
38. Lomas DA, Hurst JR, Gooptu B. Update on alpha-1 antitrypsin deficiency: new therapies. *Journal of hepatology* 2016; 65: 413-424.
39. Pearson M, Wedzicha J. Chronic Obstructive Pulmonary Disease: Critical Debates. John Wiley & Sons; 2008. p. 1-16.
40. Janus E, Phillips N, Carrell R. Smoking, lung function, and α1-antitrypsin deficiency. *The Lancet* 1985; 325: 152-154.
41. Blanco I, Bueno P, Diego I, Perez-Holanda S, Lara B, Casas-Maldonado F, Esquinas C, Miravitlles M. Alpha-1 antitrypsin Pi*SZ genotype: estimated prevalence and number of SZ subjects worldwide. *Int J Chron Obstruct Pulmon Dis* 2017; 12: 1683-1694.
42. Fawcett KA, Song K, Qian G, Farmaki AE, Packer R, John C, Shrine N, Granell R, Ring S, Timpson NJ, Yerges-Armstrong LM, Eastell R, Wain LV, Scott RA, Tobin MD, Hall IP. Pleiotropic associations of heterozygosity for the SERPINA1 Z allele in the UK Biobank. *ERJ open research* 2021; 7: 00049-02021.
43. Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Hersh CP, DeMeo DL, Hunninghake GM, Litonjua AA, Sparrow D. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nature genetics* 2010; 42: 200.
44. Wain LV, Shrine N, Artigas MS, Erzurumluoglu AM, Noyvert B, Bossini-Castillo L, Obeidat M, Henry AP, Portelli MA, Hall RJ, Billington CK, Rimington TL, Fenech AG, John C, Blake T, Jackson VE, Allen RJ, Prins BP, Campbell A, Porteous DJ, Jarvelin MR, Wielscher M, James AL, Hui J, Wareham NJ, Zhao JH, Wilson JF, Joshi PK, Stubbe B, Rawal R, Schulz H, Imboden M, Probst-Hensch NM, Karrasch S, Gieger C, Deary IJ, Harris SE, Marten J, Rudan I, Enroth S, Gyllensten U, Kerr SM, Polasek O, Kahonen M, Surakka I, Vitart V, Hayward C, Lehtimaki T, Raitakari OT, Evans DM, Henderson AJ, Pennell CE, Wang CA, Sly PD, Wan ES, Busch R, Hobbs BD, Litonjua AA, Sparrow DW, Gulsvik A, Bakke PS, Crapo JD, Beaty TH, Hansel NN, Mathias RA, Ruczinski I, Barnes KC, Bosse Y, Joubert P, van den Berge M, Brandsma CA, Pare PD, Sin DD, Nickle DC, Hao K, Gottesman O, Dewey FE, Bruse SE, Carey DJ, Kirchner HL, Jonsson S, Thorleifsson G, Jonsdottir I, Gislason T, Stefansson K, Schurmann C, Nadkarni G, Bottinger EP, Loos RJ, Walters RG, Chen Z, Millwood IY, Vaucher J, Kurmi OP, Li L, Hansell AL, Brightling C, Zeggini E, Cho MH, Silverman EK, Sayers I, Trynka G, Morris AP, Strachan DP, Hall IP, Tobin MD. Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nature Genetics* 2017; 49: 416-425.
45. Kong X, Cho MH, Anderson W, Coxson HO, Muller N, Washko G, Hoffman EA, Bakke P, Gulsvik A, Lomas DA, Silverman EK, Pillai SG. Genome-wide association study identifies BICD1 as a susceptibility gene for emphysema. *American Journal of Respiratory and Critical Care Medicine* 2011; 183: 43-49.
46. Manichaikul A, Hoffman EA, Smolonska J, Gao W, Cho MH, Baumhauer H, Budoff M, Austin JH, Washko GR, Carr JJ. Genome-wide study of percent emphysema on computed tomography in the general population. The Multi-Ethnic Study of Atherosclerosis Lung/SNP Health Association Resource Study. *American Journal of Respiratory and Critical Care Medicine* 2014; 189: 408-418.

47. Cho MH, Castaldi PJ, Hersh CP, Hobbs BD, Barr RG, Tal-Singer R, Bakke P, Gulsvik A, San José Estépar R, Van Beek EJ. A genome-wide association study of emphysema and airway quantitative imaging phenotypes. *American journal of respiratory and critical care medicine* 2015; 192: 559-569.
48. Boueiz A, Lutz SM, Cho MH, Hersh CP, Bowler RP, Washko GR, Halper-Stromberg E, Bakke P, Gulsvik A, Laird NM, Beaty TH, Coxson HO, Crapo JD, Silverman EK, Castaldi PJ, DeMeo DL. Genome-Wide Association Study of the Genetic Determinants of Emphysema Distribution. *Am J Respir Crit Care Med* 2017; 195: 757-771.
49. Zhou X, Baron RM, Hardin M, Cho MH, Zielinski J, Hawrylkiewicz I, Sliwinski P, Hersh CP, Mancini JD, Lu K. Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Human molecular genetics* 2011; 21: 1325-1335.
50. Hobbs BD, Hersh CP. Integrative genomics of chronic obstructive pulmonary disease. *Biochem Biophys Res Commun* 2014; 452: 276-286.
51. Boueiz A, Pham B, Chase R, Lamb A, Lee S, Naing ZZC, Cho MH, Parker MM, Sakornsakolpat P, Hersh CP, Crapo JD, Stergachis AB, Tal-Singer R, DeMeo DL, Silverman EK, Zhou X, Castaldi PJ. Integrative Genomics Analysis Identifies ACVR1B as a Candidate Causal Gene of Emphysema Distribution. *Am J Respir Cell Mol Biol* 2019; 60: 388-398.
52. Mejza F, Nastalek P, Doniec Z, Skucha W. Chronic bronchitis symptoms in subjects without chronic obstructive pulmonary disease: prevalence, burden and risk factors in southern Poland. *Polish archives of internal medicine* 2018; 128: 677-684.
53. Kim V, Han MK, Vance GB, Make BJ, Newell JD, Hokanson JE, Hersh CP, Stinson D, Silverman EK, Criner GJ. The chronic bronchitic phenotype of COPD: an analysis of the COPDGene Study. *Chest* 2011; 140: 626-633.
54. Lee JH, Cho MH, Hersh CP, McDonald M-LN, Crapo JD, Bakke PS, Gulsvik A, Comellas AP, Wendt CH, Lomas DA. Genetic susceptibility for chronic bronchitis in chronic obstructive pulmonary disease. *Respiratory research* 2014; 15: 113.
55. Yeo J, Morales DA, Chen T, Crawford EL, Zhang X, Blomquist TM, Levin AM, Massion PP, Arenberg DA, Midthun DE, Mazzone PJ, Nathan SD, Wainz RJ, Nana-Sinkam P, Willey PFS, Arend TJ, Padda K, Qiu S, Federov A, Hernandez DR, Hammersley JR, Yoon Y, Safi F, Khuder SA, Willey JC. RNAseq analysis of bronchial epithelial cells to identify COPD-associated genes and SNPs. *BMC pulmonary medicine* 2018; 18: 42.
56. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, Sulem P, Rafnar T, Esko T, Walter S. Sequence variants at CHRNB3–CHRNA6 and CYP2A6 affect smoking behavior. *Nature genetics* 2010; 42: 448.
57. Consortium TTaG. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 2010; 42: 441-447.
58. Van Durme YM, Eijgelsheim M, Joos GF, Hofman A, Uitterlinden AG, Brusselle GG, Stricker BH. Hedgehog-interacting protein is a COPD susceptibility gene: the Rotterdam Study. *Eur Respir J* 2010; 36: 89-95.
59. Brehm JM, Hagiwara K, Tesfaigzi Y, Bruse S, Mariani TJ, Bhattacharya S, Boutaoui N, Ziniti JP, Soto-Quiros ME, Avila L. Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease. *Thorax* 2011; 66: 1085-1090.

60. Castaldi PJ, Cho MH, Litonjua AA, Bakke P, Gulsvik A, Lomas DA, Anderson W, Beaty TH, Hokanson JE, Crapo JD. The association of genome-wide significant spirometric loci with chronic obstructive pulmonary disease susceptibility. *American journal of respiratory cell and molecular biology* 2011; 45: 1147-1153.
61. Soler Artigas M, Wain LV, Repapi E, Obeidat M, Sayers I, Burton PR, Johnson T, Zhao JH, Albrecht E, Dominiczak AF, Kerr SM, Smith BH, Cadby G, Hui J, Palmer LJ, Hingorani AD, Wannamethee SG, Whincup PH, Ebrahim S, Smith GD, Barroso I, Loos RJ, Wareham NJ, Cooper C, Dennison E, Shaheen SO, Liu JZ, Marchini J, Medical Research Council National Survey of H, Development Respiratory Study T, Dahgam S, Naluai AT, Olin AC, Karrasch S, Heinrich J, Schulz H, McKeever TM, Pavord ID, Heliovaara M, Ripatti S, Surakka I, Blakey JD, Kahonen M, Britton JR, Nyberg F, Holloway JW, Lawlor DA, Morris RW, James AL, Jackson CM, Hall IP, Tobin MD, SpiroMeta C. Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function. *Am J Respir Crit Care Med* 2011; 184: 786-795.
62. Wan ES, Cho MH, Boutaoui N, Klanderman BJ, Sylvia JS, Ziniti JP, Won S, Lange C, Pillai SG, Anderson WH. Genome-wide association analysis of body mass in chronic obstructive pulmonary disease. *American journal of respiratory cell and molecular biology* 2011; 45: 304-310.
63. Cho MH, Castaldi PJ, Wan ES, Siedlinski M, Hersh CP, Demeo DL, Himes BE, Sylvia JS, Klanderman BJ, Ziniti JP, Lange C, Litonjua AA, Sparrow D, Regan EA, Make BJ, Hokanson JE, Murray T, Hetmanski JB, Pillai SG, Kong X, Anderson WH, Tal-Singer R, Lomas DA, Coxson HO, Edwards LD, MacNee W, Vestbo J, Yates JC, Agusti A, Calverley PM, Celli B, Crim C, Rennard S, Wouters E, Bakke P, Gulsvik A, Crapo JD, Beaty TH, Silverman EK. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Human molecular genetics* 2012; 21: 947-957.
64. Hancock DB, Soler Artigas M, Gharib SA, Henry A, Manichaikul A, Ramasamy A, Loth DW, Imboden M, Koch B, McArdle WL, Smith AV, Smolonska J, Sood A, Tang W, Wilk JB, Zhai G, Zhao JH, Aschard H, Burkart KM, Curjuric I, Eijgelsheim M, Elliott P, Gu X, Harris TB, Janson C, Homuth G, Hysi PG, Liu JZ, Loehr LR, Lohman K, Loos RJ, Manning AK, Marciante KD, Obeidat M, Postma DS, Aldrich MC, Brusselle GG, Chen TH, Eiriksdottir G, Franceschini N, Heinrich J, Rotter JI, Wijmenga C, Williams OD, Bentley AR, Hofman A, Laurie CC, Lumley T, Morrison AC, Joubert BR, Rivadeneira F, Couper DJ, Kritchevsky SB, Liu Y, Wjst M, Wain LV, Vonk JM, Uitterlinden AG, Rochat T, Rich SS, Psaty BM, O'Connor GT, North KE, Mirel DB, Meibohm B, Launer LJ, Khaw KT, Hartikainen AL, Hammond CJ, Glaser S, Marchini J, Kraft P, Wareham NJ, Volzke H, Stricker BH, Spector TD, Probst-Hensch NM, Jarvis D, Jarvelin MR, Heckbert SR, Gudnason V, Boezen HM, Barr RG, Cassano PA, Strachan DP, Fornage M, Hall IP, Dupuis J, Tobin MD, London SJ. Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS Genet* 2012; 8: e1003098.
65. Wilk JB, Shrine NR, Loehr LR, Zhao JH, Manichaikul A, Lopez LM, Smith AV, Heckbert SR, Smolonska J, Tang W, Loth DW, Curjuric I, Hui J, Cho MH, Latourelle JC, Henry AP, Aldrich M, Bakke P, Beaty TH, Bentley AR, Borecki IB, Brusselle GG, Burkart KM, Chen TH, Couper D, Crapo JD, Davies G, Dupuis J, Franceschini N, Gulsvik A, Hancock

DB, Harris TB, Hofman A, Imboden M, James AL, Khaw KT, Lahousse L, Launer LJ, Litonjua A, Liu Y, Lohman KK, Lomas DA, Lumley T, Marciante KD, McArdle WL, Meibohm B, Morrison AC, Musk AW, Myers RH, North KE, Postma DS, Psaty BM, Rich SS, Rivadeneira F, Rochat T, Rotter JI, Soler Artigas M, Starr JM, Uitterlinden AG, Wareham NJ, Wijmenga C, Zanen P, Province MA, Silverman EK, Deary IJ, Palmer LJ, Cassano PA, Gudnason V, Barr RG, Loos RJ, Strachan DP, London SJ, Boezen HM, Probst-Hensch N, Gharib SA, Hall IP, O'Connor GT, Tobin MD, Stricker BH. Genome wide sssociation studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. *Am J Respir Crit Care Med* 2012.

66. Cho MH, McDonald M-LN, Zhou X, Mattheisen M, Castaldi PJ, Hersh CP, DeMeo DL, Sylvia JS, Ziniti J, Laird NM. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The lancet Respiratory medicine* 2014; 2: 214-225.
67. Dijkstra AE, Smolonska J, van den Berge M, Wijmenga C, Zanen P, Luinge MA, Platteel M, Lammers J-W, Dahlback M, Tosh K. Susceptibility to chronic mucus hypersecretion, a genome wide association study. *PloS one* 2014; 9: e91621.
68. Hardin M, Cho M, McDonald M-L, Beaty T, Ramsdell J, Bhatt S, Van Beek EJ, Make BJ, Crapo JD, Silverman EK. The clinical and genetic features of COPD-asthma overlap syndrome. *Eur Resp J* 2014; 44: 341-350.
69. Lee JH, Cho MH, McDonald M-LN, Hersh CP, Castaldi PJ, Crapo JD, Wan ES, Dy JG, Chang Y, Regan EA. Phenotypic and genetic heterogeneity among subjects with mild airflow obstruction in COPDGene. *Respir Med* 2014; 108: 1469-1480.
70. McDonald ML, Cho MH, Sorheim IC, Lutz SM, Castaldi PJ, Lomas DA, Coxson HO, Edwards LD, MacNee W, Vestbo J, Yates JC, Agusti A, Calverley PM, Celli B, Crim C, Rennard SI, Wouters EF, Bakke P, Tal-Singer R, Miller BE, Gulsvik A, Casaburi R, Wells JM, Regan EA, Make BJ, Hokanson JE, Lange C, Crapo JD, Beaty TH, Silverman EK, Hersh CP. Common genetic variants associated with resting oxygenation in chronic obstructive pulmonary disease. *American Journal of Respiratory Cell and Molecular Biology* 2014; 51: 678-687.
71. Loth DW, Soler Artigas M, Gharib SA, Wain LV, Franceschini N, Koch B, Pottinger TD, Smith AV, Duan Q, Oldmeadow C, Lee MK, Strachan DP, James AL, Huffman JE, Vitart V, Ramasamy A, Wareham NJ, Kaprio J, Wang XQ, Trochet H, Kahonen M, Flexeder C, Albrecht E, Lopez LM, de Jong K, Thyagarajan B, Alves AC, Enroth S, Omenaas E, Joshi PK, Fall T, Vinuela A, Launer LJ, Loehr LR, Fornage M, Li G, Wilk JB, Tang W, Manichaikul A, Lahousse L, Harris TB, North KE, Rudnicka AR, Hui J, Gu X, Lumley T, Wright AF, Hastie ND, Campbell S, Kumar R, Pin I, Scott RA, Pietilainen KH, Surakka I, Liu Y, Holliday EG, Schulz H, Heinrich J, Davies G, Vonk JM, Wojczynski M, Pouta A, Johansson A, Wild SH, Ingelsson E, Rivadeneira F, Volzke H, Hysi PG, Eiriksdottir G, Morrison AC, Rotter JI, Gao W, Postma DS, White WB, Rich SS, Hofman A, Aspelund T, Couper D, Smith LJ, Psaty BM, Lohman K, Burchard EG, Uitterlinden AG, Garcia M, Joubert BR, McArdle WL, Musk AB, Hansel N, Heckbert SR, Zgaga L, van Meurs JB, Navarro P, Rudan I, Oh YM, Redline S, Jarvis DL, Zhao JH, Rantanen T, O'Connor GT, Ripatti S, Scott RJ, Karrasch S, Grallert H, Gaddis NC, Starr JM, Wijmenga C, Minster RL, Lederer DJ, Pekkanen J, Gyllensten U, Campbell H, Morris AP, Glaser S, Hammond CJ, Burkart KM, Beilby J, Kritchevsky SB, Gudnason V, Hancock DB,

Williams OD, Polasek O, Zemunik T, Kolcic I, Petrini MF, Wjst M, Kim WJ, Porteous DJ, Scotland G, Smith BH, Viljanen A, Heliovaara M, Attia JR, Sayers I, Hampel R, Gieger C, Deary IJ, Boezen HM, Newman A, Jarvelin MR, Wilson JF, Lind L, Stricker BH, Teumer A, Spector TD, Melen E, Peters MJ, Lange LA, Barr RG, Bracke KR, Verhamme FM, Sung J, Hiemstra PS, Cassano PA, Sood A, Hayward C, Dupuis J, Hall IP, Brusselle GG, Tobin MD, London SJ. Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat Genet* 2014; 46: 669-677.

72. Castaldi PJ, Cho MH, Zhou X, Qiu W, McGeachie M, Celli B, Bakke P, Gulsvik A, Lomas DA, Crapo JD, Beaty TH, Rennard S, Harshfield B, Lange C, Singh D, Tal-Singer R, Riley JH, Quackenbush J, Raby BA, Carey VJ, Silverman EK, Hersh CP. Genetic control of gene expression at novel and established chronic obstructive pulmonary disease loci. *Human molecular genetics* 2015; 24: 1200-1210.
73. Tang W, Kowgier M, Loth DW, Soler Artigas M, Joubert BR, Hodge E, Gharib SA, Smith AV, Ruczinski I, Gudnason V, Mathias RA, Harris TB, Hansel NN, Launer LJ, Barnes KC, Hansen JG, Albrecht E, Aldrich MC, Allerhand M, Barr RG, Brusselle GG, Couper DJ, Curjuric I, Davies G, Deary IJ, Dupuis J, Fall T, Foy M, Franceschini N, Gao W, Gläser S, Gu X, Hancock DB, Heinrich J, Hofman A, Imboden M, Ingelsson E, James A, Karrasch S, Koch B, Kritchevsky SB, Kumar A, Lahousse L, Li G, Lind L, Lindgren C, Liu Y, Lohman K, Lumley T, McArdle WL, Meibohm B, Morris AP, Morrison AC, Musk B, North KE, Palmer LJ, Probst-Hensch NM, Psaty BM, Rivadeneira F, Rotter JI, Schulz H, Smith LJ, Sood A, Starr JM, Strachan DP, Teumer A, Uitterlinden AG, Völzke H, Voorman A, Wain LV, Wells MT, Wilk JB, Williams OD, Heckbert SR, Stricker BH, London SJ, Fornage M, Tobin MD, O'Connor GT, Hall IP, Cassano PA. Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. *PLoS One* 2014; 9: e100776.
74. Chen W, Brehm JM, Manichaikul A, Cho MH, Boutaoui N, Yan Q, Burkart KM, Enright PL, Rotter JI, Petersen H. A Genome-Wide Association Study of chronic obstructive pulmonary disease in Hispanics. *Annals of the American Thoracic Society* 2015.
75. Gharib SA, Loth DW, Soler Artigas M, Birkland TP, Wilk JB, Wain LV, Brody JA, Obeidat M, Hancock DB, Tang W, Rawal R, Boezen HM, Imboden M, Huffman JE, Lahousse L, Alves AC, Manichaikul A, Hui J, Morrison AC, Ramasamy A, Smith AV, Gudnason V, Surakka I, Vitart V, Evans DM, Strachan DP, Deary IJ, Hofman A, Glaser S, Wilson JF, North KE, Zhao JH, Heckbert SR, Jarvis DL, Probst-Hensch N, Schulz H, Barr RG, Jarvelin MR, O'Connor GT, Kahonen M, Cassano PA, Hysi PG, Dupuis J, Hayward C, Psaty BM, Hall IP, Parks WC, Tobin MD, London SJ, Consortium C, SpiroMeta C. Integrative pathway genomics of lung function and airflow obstruction. *Human molecular genetics* 2015; 24: 6836-6848.
76. Hansel NN, Pare PD, Rafaels N, Sin DD, Sandford A, Daley D, Vergara C, Huang L, Elliott WM, Pascoe CD. Genome-wide association study identification of novel loci associated with airway responsiveness in chronic obstructive pulmonary disease. *American journal of respiratory cell and molecular biology* 2015; 53: 226-234.
77. Kim S, Kim H, Cho N, Lee SK, Han BG, Sull JW, Jee SH, Shin C. Identification of FAM13A gene associated with the ratio of FEV1 to FVC in Korean population by genome-wide association studies

including gene-environment interactions. *J Hum Genet* 2015; 60: 139-145.
78. Lutz SM, Cho MH, Young K, Hersh CP, Castaldi PJ, McDonald ML, Regan E, Mattheisen M, DeMeo DL, Parker M, Foreman M, Make BJ, Jensen RL, Casaburi R, Lomas DA, Bhatt SP, Bakke P, Gulsvik A, Crapo JD, Beaty TH, Laird NM, Lange C, Hokanson JE, Silverman EK. A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC genetics* 2015; 16: 138.
79. Soler Artigas M, Wain LV, Miller S, Kheirallah AK, Huffman JE, Ntalla I, Shrine N, Obeidat M, Trochet H, McArdle WL, Alves AC, Hui J, Zhao JH, Joshi PK, Teumer A, Albrecht E, Imboden M, Rawal R, Lopez LM, Marten J, Enroth S, Surakka I, Polasek O, Lyytikainen LP, Granell R, Hysi PG, Flexeder C, Mahajan A, Beilby J, Bosse Y, Brandsma CA, Campbell H, Gieger C, Glaser S, Gonzalez JR, Grallert H, Hammond CJ, Harris SE, Hartikainen AL, Heliovaara M, Henderson J, Hocking L, Horikoshi M, Hutri-Kahonen N, Ingelsson E, Johansson A, Kemp JP, Kolcic I, Kumar A, Lind L, Melen E, Musk AW, Navarro P, Nickle DC, Padmanabhan S, Raitakari OT, Ried JS, Ripatti S, Schulz H, Scott RA, Sin DD, Starr JM, BiLEVE UK, Vinuela A, Volzke H, Wild SH, Wright AF, Zemunik T, Jarvis DL, Spector TD, Evans DM, Lehtimaki T, Vitart V, Kahonen M, Gyllensten U, Rudan I, Deary IJ, Karrasch S, Probst-Hensch NM, Heinrich J, Stubbe B, Wilson JF, Wareham NJ, James AL, Morris AP, Jarvelin MR, Hayward C, Sayers I, Strachan DP, Hall IP, Tobin MD. Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nat Commun* 2015; 6: 8658.
80. Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Artigas MS, Billington CK, Kheirallah AK, Allen R, Cook JP, Probert K, Obeidat Me, Bossé Y, Hao K, Postma DS, Paré PD, Ramasamy A, Mägi R, Mihailov E, Reinmaa E, Melén E, O'Connell J, Frangou E, Delaneau O, Freeman C, Petkova D, McCarthy M, Sayers I, Deloukas P, Hubbard R, Pavord I, Hansell AL, Thomson NC, Zeggini E, Morris AP, Marchini J, Strachan DP, Tobin MD, Hall IP. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory medicine* 2015; 3: 769-781.
81. Hardin M, Cho MH, Sharma S, Glass K, Castaldi PJ, McDonald M-L, Aschard H, Senter-Sylvia J, Tantisira K, Weiss ST. Sex-Based Genetic Association Study Identifies CELSR1 as a Possible Chronic Obstructive Pulmonary Disease Risk Locus among Women. *American journal of respiratory cell and molecular biology* 2017; 56: 332-341.
82. Hobbs BD, de Jong K, Lamontagne M, Bosse Y, Shrine N, Artigas MS, Wain LV, Hall IP, Jackson VE, Wyss AB, London SJ, North KE, Franceschini N, Strachan DP, Beaty TH, Hokanson JE, Crapo JD, Castaldi PJ, Chase RP, Bartz TM, Heckbert SR, Psaty BM, Gharib SA, Zanen P, Lammers JW, Oudkerk M, Groen HJ, Locantore N, Tal-Singer R, Rennard SI, Vestbo J, Timens W, Pare PD, Latourelle JC, Dupuis J, O'Connor GT, Wilk JB, Kim WJ, Lee MK, Oh YM, Vonk JM, de Koning HJ, Leng S, Belinsky SA, Tesfaigzi Y, Manichaikul A, Wang XQ, Rich SS, Barr RG, Sparrow D, Litonjua AA, Bakke P, Gulsvik A, Lahousse L, Brusselle GG, Stricker BH, Uitterlinden AG, Ampleford EJ, Bleecker ER, Woodruff PG, Meyers DA, Qiao D, Lomas DA, Yim JJ, Kim DK, Hawrylkiewicz I, Sliwinski P, Hardin M, Fingerlin TE, Schwartz DA, Postma DS, MacNee W, Tobin MD, Silverman EK,

Boezen HM, Cho MH. Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat Genet* 2017; 49: 426-432.
83. Burkart KM, Sofer T, London SJ, Manichaikul A, Hartwig FP, Yan Q, Soler Artigas M, Avila L, Chen W, Davis Thomas S. A Genome-wide association study in Hispanics/Latinos identifies novel signals for lung function. The hispanic community health Study/Study of latinos. *American journal of respiratory and critical care medicine* 2018; 198: 208-219.
84. Jackson VE, Latourelle JC, Wain LV, Smith AV, Grove ML, Bartz TM, Obeidat Me, Province MA, Gao W, Qaiser B. Meta-analysis of exome array data identifies six novel genetic loci for lung function. *Wellcome open research* 2018; 3.
85. Li X, Ortega VE, Ampleford EJ, Graham Barr R, Christenson SA, Cooper CB, Couper D, Dransfield MT, Han MLK, Hansel NN, Hoffman EA, Kanner RE, Kleerup EC, Martinez FJ, Paine R, 3rd, Woodruff PG, Hawkins GA, Bleecker ER, Meyers DA. Genome-wide association study of lung function and clinical implication in heavy smokers. *BMC Med Genet* 2018; 19: 134.
86. Sakornsakolpat P, Prokopenko D, Lamontagne M, Reeve NF, Guyatt AL, Jackson VE, Shrine N, Qiao D, Bartz TM, Kim DK, Lee MK, Latourelle JC, Li X, Morrow JD, Obeidat M, Wyss AB, Bakke P, Barr RG, Beaty TH, Belinsky SA, Brusselle GG, Crapo JD, de Jong K, DeMeo DL, Fingerlin TE, Gharib SA, Gulsvik A, Hall IP, Hokanson JE, Kim WJ, Lomas DA, London SJ, Meyers DA, O'Connor GT, Rennard SI, Schwartz DA, Sliwinski P, Sparrow D, Strachan DP, Tal-Singer R, Tesfaigzi Y, Vestbo J, Vonk JM, Yim JJ, Zhou X, Bosse Y, Manichaikul A, Lahousse L, Silverman EK, Boezen HM, Wain LV, Tobin MD, Hobbs BD, Cho MH. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet* 2019; 51: 494-505.
87. Hoang LL, Nguyen YP, Aspee R, Bolton SJ, Shen YH, Wang L, Kenyon NJ, Smiley-Jewell S, Pinkerton KE. Temporal and Spatial Expression of Transforming Growth Factor-beta after Airway Remodeling to Tobacco Smoke in Rats. *Am J Respir Cell Mol Biol* 2016; 54: 872-881.
88. Zhou X, Qiu W, Sathirapongsasuti JF, Cho MH, Mancini JD, Lao T, Thibault DM, Litonjua AA, Bakke PS, Gulsvik A. Gene expression analysis uncovers novel hedgehog interacting protein (HHIP) effects in human bronchial epithelial cells. *Genomics* 2013; 101: 263-272.
89. Kim WJ, Lee MK, Shin C, Cho NH, Lee SD, Oh YM, Sung J. Genome-wide association studies identify locus on 6p21 influencing lung function in the Korean population. *Respirology* 2014; 19: 360-368.
90. Milkowska-Dymanowska J, Bialas AJ, Szewczyk K, Kurmanowska Z, Gorski P, Piotrowski WJ. The usefulness of soluble receptor for advanced glycation end-products in the identification of COPD frequent exacerbator phenotype. *Int J Chron Obstruct Pulmon Dis* 2018; 13: 3879-3884.
91. Hautamaki RD, Kobayashi DK, Senior RM, Shapiro SD. Requirement for macrophage elastase for cigarette smoke-induced emphysema in mice. *Science* 1997; 277: 2002-2004.
92. Perez-Rubio G, Lopez-Flores LA, Ramirez-Venegas A, Noe-Diaz V, Garcia-Gomez L, Ambrocio-Ortiz E, Sanchez-Romero C, Hernandez-Zenteno RJ, Sansores RH, Falfan-Valencia R. Genetic polymorphisms in CYP2A6 are associated with a risk of cigarette smoking and predispose to smoking at younger ages. *Gene* 2017; 628: 205-210.

93. Brandenberger R, Schmidt A, Linton J, Wang D, Backus C, Denda S, Müller U, Reichardt LF. Identification and Characterization of a Novel Extracellular Matrix Protein Nephronectin That Is Associated with Integrin α8β1 in the Embryonic Kidney. *J Cell Biol* 2001; 154: 447-458.
94. Morimura N, Tezuka Y, Watanabe N, Yasuda M, Miyatani S, Hozumi N, Tezuka K-i. Molecular Cloning of POEM a novel adhesion molecule that interacts with α8β1 integrin. *Journal of Biological Chemistry* 2001; 276: 42172-42181.
95. Cheng CW, Ka SM, Yang SM, Shui HA, Hung YW, Ho PC, Su YC, Chen A. Nephronectin expression in nephrotoxic acute tubular necrosis. *Nephrol Dial Transplant* 2008; 23: 101-109.
96. Kashani K, Kellum JA. Novel biomarkers indicating repair or progression after acute kidney injury. *Current opinion in nephrology and hypertension* 2015; 24: 21-27.
97. Kuek V, Yang Z, Chim SM, Zhu S, Xu H, Chow ST, Tickner J, Rosen V, Erber W, Li X, An Q, Qian Y, Xu J. NPNT is Expressed by Osteoblasts and Mediates Angiogenesis via the Activation of Extracellular Signal-regulated Kinase. *Sci Rep* 2016; 6: 36210.
98. Ikehata M, Yamada A, Morimura N, Itose M, Suzawa T, Shirota T, Chikazu D, Kamijo R. Wnt/beta-catenin signaling activates nephronectin expression in osteoblasts. *Biochem Biophys Res Commun* 2017; 484: 231-234.
99. Lee S, Honda M, Yamamoto S, Kumagai-Takei N, Yoshitome K, Nishimura Y, Sada N, Kon S, Otsuki T. Role of Nephronectin in Pathophysiology of Silicosis. *International journal of molecular sciences* 2019; 20.
100. Huang JT, Lee V. Identification and characterization of a novel human nephronectin gene in silico. *Int J Mol Med* 2005; 15: 719-724.
101. Lecka-Czernik B, Lumpkin C, Goldstein S. An overexpressed gene transcript in senescent and quiescent human fibroblasts encoding a novel protein in the epidermal growth factor-like repeat family stimulates DNA synthesis. *Molecular and cellular biology* 1995; 15: 120-128.
102. Sánchez-Cortés J, Mrksich M. Using Self-Assembled Monolayers To Understand r8β1-Mediated Cell Adhesion to RGD and FEI Motifs in Nephronectin. *ACS Chemical Biology* 2011: 1078-1086.
103. Linton JM, Martin GR, Reichardt LF. The ECM protein nephronectin promotes kidney development via integrin alpha8beta1-mediated stimulation of Gdnf expression. *Development* 2007; 134: 2501.
104. Nakatani S, Ishimura E, Mori K, Fukumoto S, Yamano S, Wei M, Emoto M, Wanibuchi H, Inaba M. Nephronectin expression in glomeruli of renal biopsy specimens from various kidney diseases: nephronectin is expressed in the mesangial matrix expansion of diabetic nephropathy. *Nephron Clinical practice* 2012; 122: 114-121.
105. Watany MM, El-Horany HE-S. Nephronectin (NPNT) and the prediction of nephrotic syndrome response to steroid treatment. *European Journal of Human Genetics* 2018; 26: 1354-1360.
106. Sun Y, Kuek V, Qiu H, Tickner J, Chen L, Wang H, He W, Xu J. The emerging role of NPNT in tissue injury repair and bone homeostasis. *Journal of cellular physiology* 2018; 233: 1887-1894.
107. Kahai S, Lee SC, Seth A, Yang BB. Nephronectin promotes osteoblast differentiation via the epidermal growth factor-like repeats. *FEBS letters* 2010; 584: 233-238.
108. Kahai S, Vary CP, Gao Y, Seth A. Collagen, type V, α1 (COL5A1) is regulated by TGF-β in osteoblasts. *Matrix biology* 2004; 23: 445-455.

109. Miyazono A, Yamada A, Morimura N, Takami M, Suzuki D, Kobayashi M, Tezuka K, Yamamoto M, Kamijo R. TGF-beta suppresses POEM expression through ERK1/2 and JNK in osteoblasts. *FEBS letters* 2007; 581: 5321-5326.
110. Kurosawa T, Yamada A, Takami M, Suzuki D, Saito Y, Hiranuma K, Enomoto T, Morimura N, Yamamoto M, Iijima T, Shirota T, Itabe H, Kamijo R. Expression of nephronectin is inhibited by oncostatin M via both JAK/STAT and MAPK pathways. *FEBS Open Bio* 2015; 5: 303-307.
111. Tsukasaki M, Yamada A, Suzuki D, Aizawa R, Miyazono A, Miyamoto Y, Suzawa T, Takami M, Yoshimura K, Morimura N, Yamamoto M, Kamijo R. Expression of POEM, a positive regulator of osteoblast differentiation, is suppressed by TNF-alpha. *Biochem Biophys Res Commun* 2011; 410: 766-770.
112. Iezumi Y, Yamada A, Minami E, Ikehata M, Yoshida Y, Kato T, Morimura N, Ogata H, Sakashita A, Iijima T, Chikazu D, Kamijo R. IL-1beta suppresses nephronectin expression in osteoblasts via ERK1/2 and JNK. *Biochem Biophys Res Commun* 2017; 1: 773-775.
113. Kato T, Yamada A, Ikehata M, Yoshida Y, Sasa K, Morimura N, Sakashita A, Iijima T, Chikazu D, Ogata H. FGF‐2 suppresses expression of nephronectin viaJNKandPI3K pathways. *FEBS Open Bio* 2018; 8: 836-842.
114. Fujiwara H, Ferreira M, Donati G, Marciano DK, Linton JM, Sato Y, Hartner A, Sekiguchi K, Reichardt LF, Watt FM. The basement membrane of hair follicle stem cells is a muscle cell niche. *Cell* 2011; 144: 577-589.
115. Patra C, Diehl F, Ferrazzi F, van Amerongen MJ, Novoyatleva T, Schaefer L, Muhlfeld C, Jungblut B, Engel FB. Nephronectin regulates atrioventricular canal differentiation via Bmp4-Has2 signaling in zebrafish. *Development* 2011; 138: 4499-4509.
116. Wu CC, Jeratsch S, Graumann J, Stainier DYR. Modulation of Mammalian Cardiomyocyte Cytokinesis by the Extracellular Matrix. *Circ Res* 2020; 127: 896-907.
117. Meng F, Martin JF. Embryonic ECM Protein SLIT2 and NPNT Promote Postnatal Cardiomyocyte Cytokinesis. *Circ Res* 2020; 127: 908-910.
118. Magnussen SN, Toraskar J, Hadler-Olsen E, Steigedal TS, Svineng G. Nephronectin as a Matrix Effector in Cancer. *Cancers (Basel)* 2021; 13.
119. BL E, BS P, RK vL, CM R, AL N, MD T, KL S, EK S, JM M, RL A. Genomic analysis of a spontaneous model of breast cancer metastasis to bone reveals a role for the extracellular matrix. *Molecular Cancer Research Mcr* 2005; 3: 1.
120. Mei D, Zhao B, Zhang J, Xu H, Huang B. Nephronectin is a prognostic biomarker and promotes gastric cancer cell proliferation, migration and invasion. *Histology and histopathology* 2020: 18260-18260.
121. Miller S, Melen E, Merid SK, Hall IP, Sayers I. Genes associated with polymorphic variants predicting lung function are differentially expressed during human lung development. *Respir Res* 2016; 17: 95.
122. Saferali A, Xu Z, Sheynkman GM, Hersh CP, Cho MH, Silverman EK, Laederach A, Vollmers C, Castaldi PJ. Characterization of a COPD-Associated NPNT Functional Splicing Genetic Variant in Human Lung Tissue via Long-Read Sequencing. *medRxiv* 2020.
123. Kahai S, Lee S-C, Lee DY, Yang J, Li M, Wang C-H, Jiang Z, Zhang Y, Peng C, Yang BB. MicroRNA miR-378 regulates nephronectin

expression modulating osteoblast differentiation by targeting GalNT-7. *PLoS One* 2009; 4: e7535.
124. Arai C, Yoshizaki K, Miyazaki K, Saito K, Yamada A, Han X, Funada K, Fukumoto E, Haruyama N, Iwamoto T, Takahashi I, Fukumoto S. Nephronectin plays critical roles in Sox2 expression and proliferation in dental epithelial stem cells via EGF-like repeat domains. *Sci Rep* 2017; 7: 45181.
125. Torr E, Heath M, Mee M, Shaw D, Sharp TV, Sayers I. Expression of polycomb protein BMI‐1 maintains the plasticity of basal bronchial epithelial cells. *Physiological reports* 2016; 4: e12847.
126. Hall RJ, O'Loughlin J, Billington CK, Thakker D, Hall IP, Sayers I. Functional genomics of GPR126 in airway smooth muscle and bronchial epithelial cells. *Faseb j* 2021; 35: e21300.
127. Liu B, Billington CK, Henry AP, Bhaker SK, Kheirallah AK, Swan C, Hall IP. Chloride intracellular channel 1 (CLIC1) contributes to modulation of cyclic AMP-activated whole-cell chloride currents in human bronchial epithelial cells. *Physiological reports* 2018; 6: e13508.
128. Harvey BG, Strulovici-Barel Y, Vincent TL, Mezey JG, Raviram R, Gordon C, Salit J, Tilley AE, Chung A, Sanders A, Crystal RG. High correlation of the response of upper and lower lobe small airway epithelium to smoking. *PLoS One* 2013; 8: e72669.
129. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology* 2006; 7: 3.
130. Kheirallah AK, de Moor CH, Faiz A, Sayers I, Hall IP. Lung function associated gene Integrator Complex subunit 12 regulates protein synthesis pathways. *BMC genomics* 2017; 18: 248.
131. Garibyan L, Avashia N. Research techniques made simple: polymerase chain reaction (PCR). *The Journal of investigative dermatology* 2013; 133: e6.
132. Towbin H, Staehelin T, Gordon J. Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc Natl Acad Sci U S A* 1979; 76: 4350-4354.
133. Kurien BT, Scofield RH. Western blotting: an introduction. *Methods in molecular biology* 2015; 1312: 17-30.
134. Ramos-Vara JA. Principles and Methods of Immunohistochemistry. *Methods in molecular biology* 2017; 1641: 115-128.
135. Hornbeck PV. Enzyme-Linked Immunosorbent Assays. *Curr Protoc Immunol* 2015; 110: 2.1.1-2.1.23.
136. Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. *Cell* 2009; 136: 642-655.
137. Zhang F, Wen Y, Guo X. CRISPR/Cas9 for genome editing: progress, implications and challenges. *Human molecular genetics* 2014; 23: R40-46.
138. Rapiteanu R, Karagyozova T, Zimmermann N, Wayne G, Martufi M, Belyaev NN, Singh K, Betts J, Beinke S, Maratou K. Highly efficient genome editing in primary bronchial epithelial cells establishes FOXJ1 as essential for ciliation in human airways. *bioRxiv* 2019: 533869.
139. Reed ER, Moses E, Xiao X, Liu G, Campbell J, Perdomo C, Monti S. Assessment of a highly multiplexed RNA sequencing platform and comparison to existing high-throughput gene expression profiling techniques. *Frontiers in genetics* 2019; 10: 150.

140. Smyth GK. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor: Springer; 2005. p. 397-420.
141. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; 43: e47.
142. Kukurba KR, Montgomery SB. RNA Sequencing and Analysis. *Cold Spring Harb Protoc* 2015; 2015: 951-969.
143. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; 30: 2114-2120.
144. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics* 2015; 51: 11.14.11-11.14.19.
145. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013; 29: 15-21.
146. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 2005; 102: 15545-15550.
147. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003; 34: 267-273.
148. Slatkin M. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* 2008; 9: 477-485.
149. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* 2013; 368: 20120362.
150. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010; 26: 2336-2337.
151. Clough E, Barrett T. The Gene Expression Omnibus Database. *Methods in molecular biology* 2016; 1418: 93-110.
152. Kho AT, Bhattacharya S, Tantisira KG, Carey VJ, Gaedigk R, Leeder JS, Kohane IS, Weiss ST, Mariani TJ. Transcriptomic analysis of human lung development. *Am J Respir Crit Care Med* 2010; 181: 54-63.
153. Ezzie ME, Crawford M, Cho JH, Orellana R, Zhang S, Gelinas R, Batte K, Yu L, Nuovo G, Galas D, Diaz P, Wang K, Nana-Sinkam SP. Gene expression networks in COPD: microRNA and mRNA regulation. *Thorax* 2012; 67: 122-131.
154. Carvalho-Silva D, Pierleoni A, Pignatelli M, Ong C, Fumis L, Karamanis N, Carmona M, Faulconbridge A, Hercules A, McAuley E, Miranda A, Peat G, Spitzer M, Barrett J, Hulcoop DG, Papa E, Koscielny G, Dunham I. Open Targets Platform: new developments and updates two years on. *Nucleic acids research* 2019; 47: D1056-D1065.
155. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, Hasan S, Karamanis N, Maguire M, Papa E. Open Targets: a platform for therapeutic target identification and validation. *Nucleic acids research* 2016; 45: D985-D994.

156. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S. Towards a knowledge-based human protein atlas. *Nature biotechnology* 2010; 28: 1248.
157. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. Proteomics. Tissue-based map of the human proteome. *Science* 2015; 347: 1260419.
158. Huang JY, Labrecque JA. From GWAS to PheWAS: the search for causality in big data. *Lancet Digit Health* 2019; 1: e101-e103.
159. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. *Nat Genet* 2018; 50: 1593-1599.
160. O'Loughlin J, Hall RJ, Bhaker S, Portelli MA, Henry A, Pang V, Bates DO, Sharp TV, Sayers I. Extended lifespan of bronchial epithelial cells maintains normal cellular phenotype and transcriptome integrity. *ERJ open research* 2021; 7: 00254- 02020.
161. Miller S, Henry AP, Hodge E, Kheirallah AK, Billington CK, Rimington TL, Bhaker SK, Obeidat M, Melen E, Merid SK, Swan C, Gowland C, Nelson CP, Stewart CE, Bolton CE, Kilty I, Malarstig A, Parker SG, Moffatt MF, Wardlaw AJ, Hall IP, Sayers I. The Ser82 RAGE Variant Affects Lung Function and Serum RAGE in Smokers and sRAGE Production In Vitro. *PLoS One* 2016; 11: e0164041.
162. Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, Asplund A, Sjostedt E, Lundberg E, Szigyarto CA, Skogs M, Takanen JO, Berling H, Tegel H, Mulder J, Nilsson P, Schwenk JM, Lindskog C, Danielsson F, Mardinoglu A, Sivertsson A, von Feilitzen K, Forsberg M, Zwahlen M, Olsson I, Navani S, Huss M, Nielsen J, Ponten F, Uhlen M. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & cellular proteomics : MCP* 2014; 13: 397-406.
163. Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, Fumis L, Miranda A, Carvalho-Silva D, Buniello A, Burdett T, Hayhurst J, Baker J, Ferrer J, Gonzalez-Uriarte A, Jupp S, Karim MA, Koscielny G, Machlitt-Northen S, Malangone C, Pendlington ZM, Roncaglia P, Suveges D, Wright D, Vrousgou O, Papa E, Parkinson H, MacArthur JAL, Todd JA, Barrett JC, Schwartzentruber J, Hulcoop DG, Ochoa D, McDonagh EM, Dunham I. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res* 2021; 49: D1311-d1320.
164. Pendergrass S, Brown‐Gentry K, Dudek S, Torstenson E, Ambite J, Avery C, Buyske S, Cai C, Fesinmeyer M, Haiman C. The use of phenome‐wide association studies (PheWAS) for exploration of novel genotype‐phenotype relationships and pleiotropy discovery. *Genetic epidemiology* 2011; 35: 410-422.
165. Henry AP, Probert K, Stewart CE, Thakker D, Bhaker S, Azimi S, Hall IP, Sayers I. Defining a role for lung function associated gene GSTCD in cell homeostasis. *Respir Res* 2019; 20: 172.
166. Liu B, Henry AP, Azimi S, Miller S, Lee FK, Lee JC, Probert K, Kotlikoff MI, Sayers I, Hall IP. Exposure to lipopolysaccharide (LPS) reduces

contractile response of small airways from GSTCD-/- mice. *PLoS One* 2019; 14: e0221899.
167. Wey E, Lyons GE, Schäfer BW. A human POU domain gene, mPOU, is expressed in developing brain and specific adult tissues. *European journal of biochemistry* 1994; 220: 753-762.
168. Brandsma C-A, Van den Berge M, Hackett T-L, Brusselle G, Timens W. Recent advances in chronic obstructive pulmonary disease pathogenesis: from disease mechanisms to precision medicine. *The Journal of pathology* 2020; 250: 624-635.
169. Kon S, Honda M, Ishikawa K, Maeda M, Segawa T. Antibodies against nephronectin ameliorate anti-type II collagen-induced arthritis in mice. *FEBS Open Bio* 2020; 10: 107-117.
170. Dunsmore SE, Rannels DE. Extracellular matrix biology in the lung. *Am J Physiol* 1996; 270: L3-27.
171. Burgstaller G, Oehrle B, Gerckens M, White ES, Schiller HB, Eickelberg O. The instructive extracellular matrix of the lung: basic composition and alterations in chronic lung disease. *Eur Respir J* 2017; 50.
172. White ES. Lung extracellular matrix and fibroblast function. *Ann Am Thorac Soc* 2015; 12 Suppl 1: S30-33.
173. Bonnans C, Chou J, Werb Z. Remodelling the extracellular matrix in development and disease. *Nat Rev Mol Cell Biol* 2014; 15: 786-801.
174. Kim H, Liu X, Kohyama T, Kobayashi T, Conner H, Abe S, Fang Q, Wen FQ, Rennard SI. Cigarette smoke stimulates MMP-1 production by human lung fibroblasts through the ERK1/2 pathway. *Copd* 2004; 1: 13-23.
175. Wells JM, Parker MM, Oster RA, Bowler RP, Dransfield MT, Bhatt SP, Cho MH, Kim V, Curtis JL, Martinez FJ, Paine R, 3rd, O'Neal W, Labaki WW, Kaner RJ, Barjaktarevic I, Han MK, Silverman EK, Crapo JD, Barr RG, Woodruff P, Castaldi PJ, Gaggar A. Elevated circulating MMP-9 is linked to increased COPD exacerbation risk in SPIROMICS and COPDGene. *JCI insight* 2018; 3.
176. Gao J, Zhan B. The effects of Ang-1, IL-8 and TGF-β1 on the pathogenesis of COPD. *Mol Med Rep* 2012; 6: 1155-1159.
177. Burgess JK, Mauad T, Tjin G, Karlsson JC, Westergren-Thorsson G. The extracellular matrix - the under-recognized element in lung disease? *The Journal of pathology* 2016; 240: 397-409.
178. Obeidat Me, Miller S, Probert K, Billington CK, Henry AP, Hodge E, Nelson CP, Stewart CE, Swan C, Wain LV, Artigas MS. GSTCD and INTS12 regulation and expression in the human lung. *PloS one* 2013; 8: e74630.
179. Kurosawa T, Yamada A, Suzuki D, Morimura N, Sasagane Y, Itabe H, Kamijo R. Nephronectin Expression Is Up-Regulated by BMP-2. *Biological & pharmaceutical bulletin* 2016; 39: 1211-1215.
180. Tsukasaki M, Yamada A, Yoshimura K, Miyazono A, Yamamoto M, Takami M, Miyamoto Y, Morimura N, Kamijo R. Nephronectin expression is regulated by SMAD signaling in osteoblast-like MC3T3-E1 cells. *Biochem Biophys Res Commun* 2012; 425: 390-392.
181. Toraskar J, Magnussen SN, Chawla K, Svineng G, Steigedal TS. Nephronectin mediates p38 MAPK-induced cell viability via its integrin-binding enhancer motif. *FEBS Open Bio* 2018; 8: 1992-2001.
182. Agrawal N, Dasaradhi PV, Mohmmed A, Malhotra P, Bhatnagar RK, Mukherjee SK. RNA interference: biology, mechanism, and applications. *Microbiol Mol Biol Rev* 2003; 67: 657-685.
183. Hung JH, Weng Z. Analysis of Microarray and RNA-seq Expression Profiling Data. *Cold Spring Harb Protoc* 2017; 2017.

184. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, Ritchie ME. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Res* 2016; 5.
185. Wickham H. ggplot2: elegant graphics for data analysis. Springer; 2016.
186. Chen H, Boutros PC. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 2011; 12: 35.
187. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Stein TI, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan-Golan Y, Kohn A, Rappaport N, Safran M, Lancet D. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* 2016; 54: 1.30.31-31.30.33.
188. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2011; 39: D52-57.
189. Gründl M, Engel FB, Gaubatz S. GAS2L3: Coordinator of cardiomyocyte cytokinesis? *Cell Cycle* 2017; 16: 1853-1854.
190. Dong J, Jin G, Wu C, Guo H, Zhou B, Lv J, Lu D, Shi Y, Shu Y, Xu L, Chu M, Wang C, Zhang R, Dai J, Jiang Y, Yu D, Ma H, Zhao X, Yin Z, Yang L, Li Z, Deng Q, Cao S, Qin Z, Gong J, Sun C, Wang J, Wu W, Zhou G, Chen H, Guan P, Chen Y, Liu X, Liu L, Xu P, Han B, Bai C, Zhao Y, Zhang H, Yan Y, Liu J, Amos CI, Chen F, Tan W, Jin L, Wu T, Hu Z, Lin D, Shen H. Genome-wide association study identifies a novel susceptibility locus at 12q23.1 for lung squamous cell carcinoma in han chinese. *PLoS Genet* 2013; 9: e1003190.
191. Wolter P, Schmitt K, Fackler M, Kremling H, Probst L, Hauser S, Gruss OJ, Gaubatz S. GAS2L3, a target gene of the DREAM complex, is required for proper cytokinesis and genomic stability. *Journal of cell science* 2012; 125: 2393-2406.
192. Yang YF, Pan YH, Tian QH, Wu DC, Su SG. CBX1 Indicates Poor Outcomes and Exerts Oncogenic Activity in Hepatocellular Carcinoma. *Translational oncology* 2018; 11: 1110-1118.
193. Sullivan BA, Karpen GH. Centromeric chromatin exhibits a histone modification pattern that is distinct from both euchromatin and heterochromatin. *Nature structural & molecular biology* 2004; 11: 1076-1083.
194. Black BE, Foltz DR, Chakravarthy S, Luger K, Woods VL, Jr., Cleveland DW. Structural determinants for generating centromeric chromatin. *Nature* 2004; 430: 578-582.
195. Zhu J, Hu H, Wang J, Yang Y, Yi P. PRR11 Overexpression Facilitates Ovarian Carcinoma Cell Proliferation, Migration, and Invasion Through Activation of the PI3K/AKT/beta-Catenin Pathway. *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology* 2018; 49: 696-705.
196. Zhao Q. RNAi-mediated silencing of praline-rich gene causes growth reduction in human lung cancer cells. *International journal of clinical and experimental pathology* 2015; 8: 1760-1767.
197. Zhang C, Zhang Y, Li Y, Zhu H, Wang Y, Cai W, Zhu J, Ozaki T, Bu Y. PRR11 regulates late-S to G2/M phase progression and induces premature chromatin condensation (PCC). *Biochem Biophys Res Commun* 2015; 458: 501-508.
198. Liu Z, Lu C, Zhao G, Han X, Dong K, Wang C, Guan JZ, Wang Z. Downregulation of miR-218 by nicotine promotes cell proliferation through targeting CDK6 in non-small cell lung cancer. *J Cell Biochem* 2019: 1-8.

199. Wang H, Chen H, Zhou H, Yu W, Lu Z. Cyclin-Dependent Kinase Inhibitor 3 Promotes Cancer Cell Proliferation and Tumorigenesis in Nasopharyngeal Carcinoma by Targeting p27. *Oncology research* 2017; 25: 1431-1440.
200. Chang SL, Chen TJ, Lee YE, Lee SW, Lin LC, He HL. CDKN3 expression is an independent prognostic factor and associated with advanced tumor stage in nasopharyngeal carcinoma. *Int J Med Sci* 2018; 15: 992-998.
201. Zhou C, Wang P, Tu M, Huang Y, Xiong F, Wu Y. DEPDC1 promotes cell proliferation and suppresses sensitivity to chemotherapy in human hepatocellular carcinoma. *Biosci Rep* 2019; 39: pii: BSR20190946.
202. Zhao H, Yu M, Sui L, Gong B, Zhou B, Chen J, Gong Z, Hao C. High Expression of DEPDC1 Promotes Malignant Phenotypes of Breast Cancer Cells and Predicts Poor Prognosis in Patients With Breast Cancer. *Frontiers in oncology* 2019; 9: 262.
203. Wang Q, Li A, Jin J, Huang G. Targeted interfering DEP domain containing 1 protein induces apoptosis in A549 lung adenocarcinoma cells through the NF-kappaB signaling pathway. *OncoTargets and therapy* 2017; 10: 4443-4454.
204. Honda M, Segawa T, Ishikawa K, Maeda M, Saito Y, Kon S. Nephronectin influences EAE development by regulating Th17/Treg balance via reactive oxygen species. *Am J Physiol Cell Physiol* 2022.
205. Kinoshita M, Yamada A, Sasa K, Ikezaki K, Shirota T, Kamijo R. Phorbol-12-myristate 13-acetate inhibits Nephronectin gene expression via Protein kinase C alpha and c-Jun/c-Fos transcription factors. *Sci Rep* 2021; 11: 20360.
206. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. *Nat Rev Genet* 2019; 20: 631-656.
207. Chen L, Sun F, Yang X, Jin Y, Shi M, Wang L, Shi Y, Zhan C, Wang Q. Correlation between RNA-Seq and microarrays results using TCGA data. *Gene* 2017; 628: 200-204.
208. Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, Stefano GB. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Medical science monitor basic research* 2014; 20: 138.
209. Tang J, Saito T. Nephronectin Stimulates the Differentiation of MDPC-23 Cells into an Odontoblast-like Phenotype. *Journal of endodontics* 2017; 43: 263-271.
210. Mosmann T. Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *J Immunol Methods* 1983; 65: 55-63.
211. Portelli MA, Stewart CE, Hall IP, Brightling CE, Sayers I. Cigarette Smoke and the Induction of Urokinase Plasminogen Activator Receptor In Vivo: Selective Contribution of Isoforms to Bronchial Epithelial Phenotype. *Am J Respir Cell Mol Biol* 2015; 53: 174-183.
212. Bove PF, Hristova M, Wesley UV, Olson N, Lounsbury KM, van der Vliet A. Inflammatory levels of nitric oxide inhibit airway epithelial cell migration by inhibition of the kinase ERK1/2 and activation of hypoxia-inducible factor-1 alpha. *J Biol Chem* 2008; 283: 17919-17928.
213. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; 26: 139-140.
214. Xu X, Wang H, Li H, Cui X, Zhang H. SERPINE1 -844 and -675 polymorphisms and chronic obstructive pulmonary disease in a Chinese Han population. *J Int Med Res* 2016; 44: 1292-1301.

215. Papakrivopoulou E, Vasilopoulou E, Lindenmeyer MT, Pacheco S, Brzóska H, Price KL, Kolatsi-Joannou M, White KE, Henderson DJ, Dean CH, Cohen CD, Salama AD, Woolf AS, Long DA. Vangl2, a planar cell polarity molecule, is implicated in irreversible and reversible kidney glomerular injury. *The Journal of pathology* 2018; 246: 485-496.
216. VanderVorst K, Hatakeyama J, Berg A, Lee H, Carraway III KL. Cellular and molecular mechanisms underlying planar cell polarity pathway contributions to cancer malignancy. Seminars in cell & developmental biology: Elsevier; 2018. p. 78-87.
217. Jessen TN, Jessen JR. VANGL2 protein stability is regulated by integrin αv and the extracellular matrix. *Exp Cell Res* 2019; 374: 128-139.
218. Jessen TN, Jessen JR. VANGL2 interacts with integrin αv to regulate matrix metalloproteinase activity and cell adhesion to the extracellular matrix. *Exp Cell Res* 2017; 361: 265-276.
219. Xie X, Ning Y, Long J, Wang H, Chen X. Diverse CBX family members as potential prognostic biomarkers in non-small-cell lung cancer. *FEBS Open Bio* 2020; 10: 2206-2215.
220. Roskoski R, Jr. Cyclin-dependent protein serine/threonine kinase inhibitors as anticancer drugs. *Pharmacological research : the official journal of the Italian Pharmacological Society* 2019; 139: 471-488.
221. Luo C-W, Hou M-F, Huang C-W, Wu C-C, Ou-Yang F, Li Q-L, Wu C-C, Pan M-R. The CDK6-c-Jun-Sp1-MMP-2 axis as a biomarker and therapeutic target for triple-negative breast cancer. *Am J Cancer Res* 2020; 10: 4325-4341.
222. Tsuji T, Aoshiba K, Nagai A. Alveolar cell senescence in patients with pulmonary emphysema. *Am J Respir Crit Care Med* 2006; 174: 886-893.
223. Radulovich N, Leung L, Ibrahimov E, Navab R, Sakashita S, Zhu CQ, Kaufman E, Lockwood WW, Thu KL, Fedyshyn Y, Moffat J, Lam WL, Tsao MS. Coiled-coil domain containing 68 (CCDC68) demonstrates a tumor-suppressive role in pancreatic ductal adenocarcinoma. *Oncogene* 2015; 34: 4238-4247.
224. Huang N, Xia Y, Zhang D, Wang S, Bao Y, He R, Teng J, Chen J. Hierarchical assembly of centriole subdistal appendages via centrosome binding proteins CCDC120 and CCDC68. *Nat Commun* 2017; 8: 15057.
225. Qu X, Dang X, Wang W, Li Y, Xu D, Shang D, Chang Y. Long Noncoding RNAs and mRNA Regulation in Peripheral Blood Mononuclear Cells of Patients with Chronic Obstructive Pulmonary Disease. *Mediators of inflammation* 2018; 2018: 7501851.
226. Angata T, Ishii T, Motegi T, Oka R, Taylor RE, Soto PC, Chang Y-C, Secundino I, Gao C-X, Ohtsubo K. Loss of Siglec-14 reduces the risk of chronic obstructive pulmonary disease exacerbation. *Cellular and Molecular Life Sciences* 2013; 70: 3199-3210.
227. Dai L, Li J, Xie L, Wang W, Lu Y, Xie M, Huang J, Shen K, Yang H, Pei C, Zhao Y, Zhang W. A Biallelic Frameshift Mutation in Nephronectin Causes Bilateral Renal Agenesis in Humans. *J Am Soc Nephrol* 2021; 32: 1871-1879.
228. Makris S, Paulsen M, Johansson C. Type I Interferons as Regulators of Lung Inflammation. *Frontiers in immunology* 2017; 8: 259.
229. Zhu Y, Song D, Song Y, Wang X. Interferon gamma induces inflammatory responses through the interaction of CEACAM1 and PI3K in airway epithelial cells. *Journal of translational medicine* 2019; 17: 147.

230. Matsumoto K, Inoue H. Viral infections in asthma and COPD. *Respiratory investigation* 2014; 52: 92-100.
231. García-Valero J, Olloquequi J, Montes JF, Rodríguez E, Martín-Satué M, Texidó L, Sancho JF. Deficient pulmonary IFN - β expression in COPD patients. *PloS one* 2019; 14: e0217803.
232. Rao MS, Van Vleet TR, Ciurlionis R, Buck WR, Mittelstadt SW, Blomme EAG, Liguori MJ. Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. *Front Genet* 2018; 9: 636.
233. Bottomly D, Walter NA, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One* 2011; 6: e17820.
234. Hryhorowicz M, Lipinski D, Zeyland J, Slomski R. CRISPR/Cas9 Immune System as a Tool for Genome Engineering. *Archivum immunologiae et therapiae experimentalis* 2017; 65: 233-240.
235. Berns K, Hijmans EM, Mullenders J, Brummelkamp TR, Velds A, Heimerikx M, Kerkhoven RM, Madiredjo M, Nijkamp W, Weigelt B, Agami R, Ge W, Cavet G, Linsley PS, Beijersbergen RL, Bernards R. A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* 2004; 428: 431-437.
236. Jackson AL, Bartz SR, Schelter J, Kobayashi SV, Burchard J, Mao M, Li B, Cavet G, Linsley PS. Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol* 2003; 21: 635-637.
237. Porteus MH, Baltimore D. Chimeric nucleases stimulate gene targeting in human cells. *Science* 2003; 300: 763.
238. Bai Y, Liu Y, Su Z, Ma Y, Ren C, Zhao R, Ji HL. Gene editing as a promising approach for respiratory diseases. *Journal of medical genetics* 2018; 55: 143-149.
239. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 2012; 337: 816-821.
240. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 2011; 471: 602-607.
241. Lino CA, Harper JC, Carney JP, Timlin JA. Delivering CRISPR: a review of the challenges and approaches. *Drug Deliv* 2018; 25: 1234-1257.
242. Martinez-Anton A, Sokolowska M, Kern S, Davis AS, Alsaaty S, Taubenberger JK, Sun J, Cai R, Danner RL, Eberlein M. Changes in microRNA and mRNA expression with differentiation of human bronchial epithelial cells. *Am J Respir Cell Mol Biol* 2013; 49: 384-395.
243. Bellec J, Bacchetta M, Losa D, Anegon I, Chanson M, Nguyen TH. CFTR inactivation by lentiviral vector-mediated RNA interference and CRISPR-Cas9 genome editing in human airway epithelial cells. *Curr Gene Ther* 2015; 15: 447-459.
244. Chu HW, Rios C, Huang C, Wesolowska-Andersen A, Burchard EG, O'Connor BP, Fingerlin TE, Nichols D, Reynolds SD, Seibold MA. CRISPR-Cas9-mediated gene knockout in primary human airway epithelial cells reveals a proinflammatory role for MUC18. *Gene Ther* 2015; 22: 822-829.
245. Wu Q, Jiang D, Matsuda JL, Ternyak K, Zhang B, Chu HW. Cigarette smoke induces human airway epithelial senescence via growth differentiation factor 15 production. *American journal of respiratory cell and molecular biology* 2016; 55: 429-438.

246. Voets O, Tielen F, Elstak E, Benschop J, Grimbergen M, Stallen J, Janssen R, van Marle A, Essrich C. Highly efficient gene inactivation by adenoviral CRISPR/Cas9 in human primary cells. *PloS one* 2017; 12: e0182974.
247. Dakhama A, Al Mubarak R, Pavelka N, Voelker D, Seibold M, Ledford JG, Kraft M, Li L, Chu HW. Tollip Inhibits ST2 Signaling in Airway Epithelial Cells Exposed to Type 2 Cytokines and Rhinovirus. *Journal of innate immunity* 2020; 12: 103-115.
248. Martufi M, Good RB, Rapiteanu R, Schmidt T, Patili E, Tvermosegaard K, New M, Nanthakumar CB, Betts J, Blanchard AD. Single-Step, High-Efficiency CRISPR-Cas9 Genome Editing in Primary Human Disease-Derived Fibroblasts. *The CRISPR journal* 2019; 2: 31-40.
249. Koh KD, Siddiqui S, Cheng D, Bonser LR, Sun DI, Zlock LT, Finkbeiner WE, Woodruff PG, Erle DJ. Efficient RNP-directed Human Gene Targeting Reveals SPDEF Is Required for IL-13-induced Mucostasis. *Am J Respir Cell Mol Biol* 2020; 62: 373-381.
250. Jacobs JJ, Kieboom K, Marino S, DePinho RA, van Lohuizen M. The oncogene and Polycomb-group gene bmi-1 regulates cell proliferation and senescence through the ink4a locus. *Nature* 1999; 397: 164-168.
251. Vonlanthen S, Heighway J, Altermatt HJ, Gugger M, Kappeler A, Borner MM, van Lohuizen M, Betticher DC. The bmi-1 oncoprotein is differentially expressed in non-small cell lung cancer and correlates with INK4A-ARF locus expression. *Br J Cancer* 2001; 84: 1372-1376.
252. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010.
253. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* 2014; 30: 2114-2120.
254. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; 26: 139-140.
255. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003; 4: P3.
256. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; 4: 44-57.
257. Liu R, Holik AZ, Su S, Jansz N, Chen K, Leong HS, Blewitt ME, Asselin-Labat ML, Smyth GK, Ritchie ME. Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res* 2015; 43: e97.
258. Callahan MJ, Nagymanyoki Z, Bonome T, Johnson ME, Litkouhi B, Sullivan EH, Hirsch MS, Matulonis UA, Liu J, Birrer MJ, Berkowitz RS, Mok SC. Increased HLA-DMB expression in the tumor epithelium is associated with increased CTL infiltration and improved prognosis in advanced-stage serous ovarian cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2008; 14: 7667-7673.
259. Cao Y, Zhu Q, Cai X, Wu T, Aierken X, Ahmat A, Liu S, Li N. Glial Cell-Derived Neurotrophic Factor Functions as a Potential Candidate Gene in Obstructive Sleep Apnea Based on a Combination of Bioinformatics and Targeted Capture Sequencing Analyses. *Biomed Res Int* 2021; 2021: 6656943.

260. Bae DH, Kim HJ, Yoon BH, Park JL, Kim M, Kim SK, Kim SY, Lee SI, Song KS, Kim YS. STK31 upregulation is associated with chromatin remodeling in gastric cancer and induction of tumorigenicity in a xenograft mouse model. *Oncology reports* 2021; 45.
261. Xiong J, Xing S, Dong Z, Niu L, Xu Q, Liu P, Yang P. STK31 regulates the proliferation and cell cycle of lung cancer cells via the Wnt/β-catenin pathway and feedback regulation by c-myc. *Oncology reports* 2020; 43: 395-404.
262. Fukuda MN, Sugihara K. Signal transduction in human embryo implantation. *Cell Cycle* 2007; 6: 1153-1156.
263. Chang XZ, Yu J, Zhang XH, Yin J, Wang T, Cao XC. Enhanced expression of trophinin promotes invasive and metastatic potential of human gallbladder cancer cells. *J Cancer Res Clin Oncol* 2009; 135: 581-590.
264. Zhai Y, Liu Y, Qi Y, Long X, Gao J, Yao X, Chen Y, Wang X, Lu S, Zhao Z. The soluble VEGF receptor sFlt-1 contributes to endothelial dysfunction in IgA nephropathy. *PLoS One* 2020; 15: e0234492.
265. Hall IP. FLT1: a potential therapeutic target in sepsis-associated ARDS? *The Lancet Respiratory medicine* 2020; 8: 219-220.
266. Ji X, Hua H, Shen Y, Bu S, Yi S. Let-7d modulates the proliferation, migration, tubulogenesis of endothelial cells. *Molecular and cellular biochemistry* 2019; 462: 75-83.
267. Filipek-Górniok B, Holmborn K, Haitina T, Habicher J, Oliveira MB, Hellgren C, Eriksson I, Kjellén L, Kreuger J, Ledin J. Expression of chondroitin/dermatan sulfate glycosyltransferases during early zebrafish development. *Developmental dynamics : an official publication of the American Association of Anatomists* 2013; 242: 964-975.
268. Chen LH, Kuo WH, Tsai MH, Chen PC, Hsiao CK, Chuang EY, Chang LY, Hsieh FJ, Lai LC, Chang KJ. Identification of prognostic genes for recurrent risk prediction in triple negative breast cancer patients in Taiwan. *PLoS One* 2011; 6: e28222.
269. Baron-Cohen S, Murphy L, Chakrabarti B, Craig I, Mallya U, Lakatošová S, Rehnstrom K, Peltonen L, Wheelwright S, Allison C, Fisher SE, Warrier V. A genome wide association study of mathematical ability reveals an association at chromosome 3q29, a locus associated with autism and learning difficulties: a preliminary study. *PLoS One* 2014; 9: e96374.
270. Dvornyk V, Ponomarenko I, Belyaeva T, Reshetnikov E, Churnosov M. Filaggrin gene polymorphisms are associated with atopic dermatitis in women but not in men in the Caucasian population of Central Russia. *PLoS One* 2021; 16: e0261026.
271. Hasan AN, Ahmad MW, Madar IH, Grace BL, Hasan TN. An in silico analytical study of lung cancer and smokers datasets from gene expression omnibus (GEO) for prediction of differentially expressed genes. *Bioinformation* 2015; 11: 229-235.
272. Li X, Ma C, Luo H, Zhang J, Wang J, Guo H. Identification of the differential expression of genes and upstream microRNAs in small cell lung cancer compared with normal lung based on bioinformatics analysis. *Medicine (Baltimore)* 2020; 99: e19086.
273. Tabib T, Huang M, Morse N, Papazoglou A, Behera R, Jia M, Bulik M, Monier DE, Benos PV, Chen W, Domsic R, Lafyatis R. Myofibroblast transcriptome indicates SFRP2(hi) fibroblast progenitors in systemic sclerosis skin. *Nat Commun* 2021; 12: 4384.

274. Xiao L, Peng H, Yan M, Chen S. Silencing ACTG1 Expression Induces Prostate Cancer Epithelial Mesenchymal Transition Through MAPK/ERK Signaling Pathway. *DNA and cell biology* 2021; 40: 1445-1455.
275. Liu J, Zhong X, Li J, Liu B, Guo S, Chen J, Tan Q, Wang Q, Ma W, Wu Z, Wang H, Hou M, Zhang HT, Zhou Q. Screening and identification of lung cancer metastasis-related genes by suppression subtractive hybridization. *Thorac Cancer* 2012; 3: 207-216.
276. Munye MM, Shoemark A, Hirst RA, Delhove JM, Sharp TV, McKay TR, O'Callaghan C, Baines DL, Howe SJ, Hart SL. BMI-1 extends proliferative potential of human bronchial epithelial cells while retaining their mucociliary differentiation capacity. *Am J Physiol Lung Cell Mol Physiol* 2017; 312: L258-l267.
277. Bovard D, Giralt A, Trivedi K, Neau L, Kanellos P, Iskandar A, Kondylis A, Luettich K, Frentzel S, Hoeng J, Peitsch MC. Comparison of the basic morphology and function of 3D lung epithelial cultures derived from several donors. *Curr Res Toxicol* 2020; 1: 56-69.
278. Yoo SS, Lee JH, Hong MJ, Choi JE, Kang HG, Do SK, Kim JH, Baek SA, Choi SH, Lee WK, Lee YH, Seo H, Lee J, Lee SY, Cha SI, Kim CH, Park JY. Effect of genetic variation in Notch regulator DTX1 on SCLC prognosis compared with the effect on NSCLC prongosis. *Thorac Cancer* 2020; 11: 2698-2703.
279. Fan F, Yin R, Wang L, Zhao S, Lv D, Yang K, Geng S, Yang N, Zhang X, Wang H. ALDH3A1 driving tumor metastasis is mediated by p53/BAG1 in lung adenocarcinoma. *Journal of Cancer* 2021; 12: 4780-4790.
280. Rhodes CJ, Batai K, Bleda M, Haimel M, Southgate L, Germain M, Pauciulo MW, Hadinnapola C, Aman J, Girerd B, Arora A, Knight J, Hanscombe KB, Karnes JH, Kaakinen M, Gall H, Ulrich A, Harbaum L, Cebola I, Ferrer J, Lutz K, Swietlik EM, Ahmad F, Amouyel P, Archer SL, Argula R, Austin ED, Badesch D, Bakshi S, Barnett C, Benza R, Bhatt N, Bogaard HJ, Burger CD, Chakinala M, Church C, Coghlan JG, Condliffe R, Corris PA, Danesino C, Debette S, Elliott CG, Elwing J, Eyries M, Fortin T, Franke A, Frantz RP, Frost A, Garcia JGN, Ghio S, Ghofrani HA, Gibbs JSR, Harley J, He H, Hill NS, Hirsch R, Houweling AC, Howard LS, Ivy D, Kiely DG, Klinger J, Kovacs G, Lahm T, Laudes M, Machado RD, MacKenzie Ross RV, Marsolo K, Martin LJ, Moledina S, Montani D, Nathan SD, Newnham M, Olschewski A, Olschewski H, Oudiz RJ, Ouwehand WH, Peacock AJ, Pepke-Zaba J, Rehman Z, Robbins I, Roden DM, Rosenzweig EB, Saydain G, Scelsi L, Schilz R, Seeger W, Shaffer CM, Simms RW, Simon M, Sitbon O, Suntharalingam J, Tang H, Tchourbanov AY, Thenappan T, Torres F, Toshner MR, Treacy CM, Vonk Noordegraaf A, Waisfisz Q, Walsworth AK, Walter RE, Wharton J, White RJ, Wilt J, Wort SJ, Yung D, Lawrie A, Humbert M, Soubrier F, Trégouët DA, Prokopenko I, Kittles R, Gräf S, Nichols WC, Trembath RC, Desai AA, Morrell NW, Wilkins MR. Genetic determinants of risk in pulmonary arterial hypertension: international genome-wide association studies and meta-analysis. *The Lancet Respiratory medicine* 2019; 7: 227-238.
281. Heo Y, Heo J, Han SS, Kim WJ, Cheong HS, Hong Y. Difference of copy number variation in blood of patients with lung cancer. *Int J Biol Markers* 2021; 36: 3-9.
282. Zhou H, Brekman A, Zuo WL, Ou X, Shaykhiev R, Agosto-Perez FJ, Wang R, Walters MS, Salit J, Strulovici-Barel Y, Staudt MR, Kaner RJ, Mezey JG, Crystal RG, Wang G. POU2AF1 Functions in the Human

Airway Epithelium To Regulate Expression of Host Defense Genes. *J Immunol* 2016; 196: 3159-3167.
283. Sun X, Hou T, Cheung E, Iu TN, Tam VW, Chu IM, Tsang MS, Chan PK, Lam CW, Wong CK. Anti-inflammatory mechanisms of the novel cytokine interleukin-38 in allergic asthma. *Cell Mol Immunol* 2020; 17: 631-646.
284. Guillen-Guio B, Lorenzo-Salazar JM, Ma SF, Hou PC, Hernandez-Beeftink T, Corrales A, García-Laorden MI, Jou J, Espinosa E, Muriel A, Domínguez D, Lorente L, Martín MM, Rodríguez-Gallego C, Solé-Violán J, Ambrós A, Carriedo D, Blanco J, Añón JM, Reilly JP, Jones TK, Ittner CA, Feng R, Schöneweck F, Kiehntopf M, Noth I, Scholz M, Brunkhorst FM, Scherag A, Meyer NJ, Villar J, Flores C. Sepsis-associated acute respiratory distress syndrome in individuals of European ancestry: a genome-wide association study. *The Lancet Respiratory medicine* 2020; 8: 258-266.
285. Cao Z, Lis R, Ginsberg M, Chavez D, Shido K, Rabbany SY, Fong GH, Sakmar TP, Rafii S, Ding BS. Targeting of the pulmonary capillary vascular niche promotes lung alveolar repair and ameliorates fibrosis. *Nat Med* 2016; 22: 154-162.
286. Chang WH, Ho BC, Hsiao YJ, Chen JS, Yeh CH, Chen HY, Chang GC, Su KY, Yu SL. JAG1 Is Associated with Poor Survival through Inducing Metastasis in Lung Cancer. *PLoS One* 2016; 11: e0150355.
287. Sun C, Zhou T, Xie G, Fu S, Gao L, Liao J, Wu Y, Wang G. Proteomics of exhaled breath condensate in stable COPD and non-COPD controls using tandem mass tags (TMTs) quantitative mass spectrometry: A pilot study. *J Proteomics* 2019; 206: 103392.
288. Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, Bowler R, Reisdorph N. Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. *Sci Rep* 2018; 8: 17132.
289. Ambrocio-Ortiz E, Pérez-Rubio G, Ramírez-Venegas A, Hernández-Zenteno RJ, Paredes-López A, Sansores RH, Ramírez-Díaz ME, Cruz-Vicente F, Martínez-Gómez ML, Falfán-Valencia R. Protective Role of Genetic Variants in HSP90 Genes-Complex in COPD Secondary to Biomass-Burning Smoke Exposure and Non-Severe COPD Forms in Tobacco Smoking Subjects. *Curr Issues Mol Biol* 2021; 43: 887-899.
290. Apostolou A, Kerenidi T, Michopoulos A, Gourgoulianis KI, Noutsias M, Germenis AE, Speletas M. Association between TLR2/TLR4 gene polymorphisms and COPD phenotype in a Greek cohort. *Herz* 2017; 42: 752-757.
291. Allam V, Faiz A, Lam M, Rathnayake SNH, Ditz B, Pouwels SD, Brandsma CA, Timens W, Hiemstra PS, Tew GW, Neighbors M, Grimbaldeston M, van den Berge M, Donnelly S, Phipps S, Bourke JE, Sukkar MB. RAGE and TLR4 differentially regulate airway hyperresponsiveness: Implications for COPD. *Allergy* 2021; 76: 1123-1135.
292. Di Stefano A, Sangiorgi C, Gnemmi I, Casolari P, Brun P, Ricciardolo FLM, Contoli M, Papi A, Maniscalco P, Ruggeri P, Girbino G, Cappello F, Pavlides S, Guo Y, Chung KF, Barnes PJ, Adcock IM, Balbi B, Caramori G. TGF-β Signaling Pathways in Different Compartments of the Lower Airways of Patients With Stable COPD. *Chest* 2018; 153: 851-862.
293. Celedón JC, Lange C, Raby BA, Litonjua AA, Palmer LJ, DeMeo DL, Reilly JJ, Kwiatkowski DJ, Chapman HA, Laird N, Sylvia JS,

Hernandez M, Speizer FE, Weiss ST, Silverman EK. The transforming growth factor-beta1 (TGFB1) gene is associated with chronic obstructive pulmonary disease (COPD). *Human molecular genetics* 2004; 13: 1649-1656.
294. Chen H, Lu R, Zhang YG, Sun J. Vitamin D Receptor Deletion Leads to the Destruction of Tight and Adherens Junctions in Lungs. *Tissue Barriers* 2018; 6: 1-13.
295. Jiang Z, Knudsen NH, Wang G, Qiu W, Naing ZZC, Bai Y, Ai X, Lee CH, Zhou X. Genetic Control of Fatty Acid β-Oxidation in Chronic Obstructive Pulmonary Disease. *Am J Respir Cell Mol Biol* 2017; 56: 738-748.
296. Wijnhoven JH, Hafmans T, Dekhuijzen PN. Distribution of costameric proteins in the diaphragm of patients with chronic obstructive pulmonary disease. *Respiration* 2006; 73: 529-537.
297. Brusselle G, Pavord ID, Landis S, Pascoe S, Lettis S, Morjaria N, Barnes N, Hilton E. Blood eosinophil levels as a biomarker in COPD. *Respir Med* 2018; 138: 21-31.
298. Regan EA, Hersh CP, Castaldi PJ, DeMeo DL, Silverman EK, Crapo JD, Bowler RP. Omics and the Search for Blood Biomarkers in Chronic Obstructive Pulmonary Disease. Insights from COPDGene. *Am J Respir Cell Mol Biol* 2019; 61: 143-149.
299. Sun W, Kechris K, Jacobson S, Drummond MB, Hawkins GA, Yang J, Chen T-h, Quibrera PM, Anderson W, Barr RG. Common genetic polymorphisms influence blood biomarker measurements in COPD. *PLoS genetics* 2016; 12: e1006011.
300. Cordero AIH, Milne S, Yang CX, Li X, Shi H, Sin DD, Obeidat Me. Integrative Omics Reveal Novel Protein Targets For Chronic Obstructive Pulmonary Disease Biomarker Discovery. *medRxiv* 2021: 2021.2001.2011.21249617.
301. Casaburi R, Celli B, Crapo J, Criner G, Croxton T, Gaw A, Jones P, Kline-Leidy N, Lomas DA, Merrill D, Polkey M, Rennard S, Sciurba F, Tal-Singer R, Stockley R, Turino G, Vestbo J, Walsh J. The COPD Biomarker Qualification Consortium (CBQC). *Copd* 2013; 10: 367-377.
302. Miller BE, Tal-Singer R, Rennard SI, Furtwaengler A, Leidy N, Lowings M, Martin UJ, Martin TR, Merrill DD, Snyder J, Walsh J, Mannino DM. Plasma Fibrinogen Qualification as a Drug Development Tool in Chronic Obstructive Pulmonary Disease. Perspective of the Chronic Obstructive Pulmonary Disease Biomarker Qualification Consortium. *Am J Respir Crit Care Med* 2016; 193: 607-613.
303. Barnes PJ. Targeting cytokines to treat asthma and chronic obstructive pulmonary disease. *Nature reviews Immunology* 2018; 18: 454-466.
304. Carolan BJ, Kim YI, Williams AA, Kechris K, Lutz S, Reisdorph N, Bowler RP. The association of adiponectin with computed tomography phenotypes in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2013; 188: 561-566.
305. Celli BR, Wedzicha JA. Update on Clinical Aspects of Chronic Obstructive Pulmonary Disease. *N Engl J Med* 2019; 381: 1257-1266.
306. Pavord ID, Jones PW, Burgel PR, Rabe KF. Exacerbations of COPD. *Int J Chron Obstruct Pulmon Dis* 2016; 11 Spec Iss: 21-30.
307. Jones PW, Chen WH, Wilcox TK, Sethi S, Leidy NK. Characterizing and quantifying the symptomatic features of COPD exacerbations. *Chest* 2011; 139: 1388-1394.
308. Hurst JR, Vestbo J, Anzueto A, Locantore N, Müllerova H, Tal-Singer R, Miller B, Lomas DA, Agusti A, MacNee W. Susceptibility to

exacerbation in chronic obstructive pulmonary disease. *New England Journal of Medicine* 2010; 363: 1128-1138.
309. Duvoix A, Dickens J, Haq I, Mannino D, Miller B, Tal-Singer R, Lomas DA. Blood fibrinogen as a biomarker of chronic obstructive pulmonary disease. *Thorax* 2013; 68: 670-676.
310. Minas M, Mystridou P, Georgoulias P, Pournaras S, Kostikas K, Gourgoulianis KI. Fetuin-A is associated with disease severity and exacerbation frequency in patients with COPD. *Copd* 2013; 10: 28-34.
311. Keene JD, Jacobson S, Kechris K, Kinney GL, Foreman MG, Doerschuk CM, Make BJ, Curtis JL, Rennard SI, Barr RG. Biomarkers predictive of exacerbations in the SPIROMICS and COPDGene cohorts. *American journal of respiratory and critical care medicine* 2017; 195: 473-481.
312. Tong X, Wang D, Liu S, Ma Y, Li Z, Tian P, Fan H. The YKL-40 protein is a potential biomarker for COPD: a meta-analysis and systematic review. *Int J Chron Obstruct Pulmon Dis* 2018; 13: 409-418.
313. Singh D, Agusti A, Anzueto A, Barnes PJ, Bourbeau J, Celli BR, Criner GJ, Frith P, Halpin DMG, Han M, López Varela MV, Martinez F, Montes de Oca M, Papi A, Pavord ID, Roche N, Sin DD, Stockley R, Vestbo J, Wedzicha JA, Vogelmeier C. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease: the GOLD science committee report 2019. *Eur Respir J* 2019; 53.
314. Brandenberger R, Schmidt A, Linton J, Wang D, Backus C, Denda S, Muller U, Reichardt LF. Identification and characterization of a novel extracellular matrix protein nephronectin that is associated with integrin alpha8beta1 in the embryonic kidney. *J Cell Biol* 2001; 154: 447-458.
315. Mohan A, Agarwal S, Clauss M, Britt NS, Dhillon NK. Extracellular vesicles: novel communicators in lung diseases. *Respir Res* 2020; 21: 175.
316. Papakonstantinou E, Karakiulakis G, Batzios S, Savic S, Roth M, Tamm M, Stolz D. Acute exacerbations of COPD are associated with significant activation of matrix metalloproteinase 9 irrespectively of airway obstruction, emphysema and infection. *Respir Res* 2015; 16: 78.
317. Kuphal S, Wallner S, Bosserhoff AK. Loss of nephronectin promotes tumor progression in malignant melanoma. *Cancer Science* 2008; 99: 229-233.
318. Sopel N, Ohs A, Schiffer M, Müller-Deile J. A Tight Control of Non-Canonical TGF-β Pathways and MicroRNAs Downregulates Nephronectin in Podocytes. *Cells* 2022; 11.
319. Kim JH, Cheong HS, Sul JH, Seo JM, Kim DY, Oh JT, Park KW, Kim HY, Jung SM, Jung K, Cho MJ, Bae JS, Shin HD. A genome-wide association study identifies potential susceptibility loci for Hirschsprung disease. *PLoS One* 2014; 9: e110292.