# Investigating Lung Function Gene *GSTCD* using Gene Targeting in Human and Mouse Systems

**Kelly Probert**

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

**May 2016**

# Abstract

Respiratory diseases are a significant global health concern and burden, with one of the major respiratory diseases causing a decline in lung function being chronic obstructive pulmonary disease (COPD). Critically, current therapies have limited impact. Recently, single nucleotide polymorphisms at chromosome locus 4q24 have been identified as associated with the lung function measure forced expiratory volume in 1 second and COPD, specifically identifying the gene Glutathione *S*-transferase C-terminal domain containing (*GSTCD*) as a viable candidate gene. Data indicate that GSTCD or associated biological pathways may be a potential target for interventions to alleviate COPD. Further analysis into the pathways and phenotypic effects of this gene may aid our understanding of the mechanisms by which *GSTCD* is associated with and influences lung function and COPD.

Importantly, there are currently no reports describing the function of GSTCD, therefore this project aimed to understand the consequences of reduction and loss of GSTCD within the context of lung biology. This was performed by analysis of human cell knock-down and knock-out mouse models respectively, as well as utilising established protein sequence homology analysis to infer potential function.

Conclusions drawn from these results suggest a possible role for GSTCD in cell growth in a human *in vitro* GSTCD knock-down model, but no gross morphological differences were evident in Gstcd knock-out mice. Additionally, GSTCD sequence homology analysis in several protein prediction servers suggests methyltransferase activity as a potential function for GSTCD, differing from the initial prediction of glutathione S-transferase enzyme activity.

Overall, this research represents an important and novel development for the lung function and COPD associated gene *GSTCD*, providing interesting avenues for future exploration.

# Publications

**K Probert**, S Miller, AK Kheirallah and IP Hall (2015) "**Developmental Genetics of the COPD Lung**" *COPD Research and Practice*. doi:10.1186/s40749-015-0014.

LV Wain, N Shrine, S Miller, V Jackson, I Ntalla, M Soler-Artigas, CK Billington, AK Kheirallah, R Allen, JP Cook, **K Probert** *et al.* (2015) "**Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE**)" *The Lancet Respiratory Medicine*. doi: 10.1016/S2213-2600(15)00283-0.

M Obeidat, S Miller, **K Probert**, CK Billington, AP Henry *et al.* (2013) "***GSTCD* and *INTS12* Regulation and Expression in the Human Lung**" *PLoS ONE* doi: 10.1371.

# Acknowledgements

I would like to thank my supervisors Professor Ian Hall and Doctor Ian Sayers for the opportunity to conduct my Master's degree with them and for all of their support throughout my time in the Division of Respiratory Medicine, particularly during these studies. Especially for always finding time to discuss work progress in their busy schedules and believing in my abilities to present research at international conferences. I would also like to thank Doctor Charlotte Dean for allowing me to become part of her research group and for supporting me as my supervisor during my collaborative placement at Imperial College London.

I am enormously grateful to the Division of Respiratory Medicine as a whole, it has been an extremely welcoming and encouraging group of people who have become friends along with helping me develop an extensive range of research techniques. I would also like to thank Doctor Charlotte Billington and Doctor Emily Hodge in particular for guiding me and for sharing numerous chatting therapy sessions over cups of tea.

Thank you to my family for always encouraging me in my education and putting up with me throughout my studies as well as for having a thoughtful meal waiting for me after very long days in the lab.

Finally, a massive thank you to Joshua Robbins who has always been there for me and been incredibly supportive throughout, I really appreciate everything you have done and do.

# Contents


1 Introduction ..... 1

1.1 Lung Function ..... 2

1.2 Chronic Obstructive Pulmonary Disease ..... 3

1.3 Environmental and Genetic Factors of COPD ..... 6

1.4 Genome Wide Association Studies ..... 7

1.4.1 Genome Wide Association Studies Identify Significant Association between *GSTCD* and $FEV_1$ ..... 9

1.5 Biological Function of GSTCD ..... 14

1.6 *GSTCD* Gene Structure and Conserved Domains ..... 16

1.7 GSTCD Expression Profiles and Localisation in Humans ..... 18

1.8 Human GSTCD Knock-down *in vitro* Model ..... 23

1.9 Mice as Model Organisms ..... 25

1.9.1 Gstcd Knock-out Generation Methodology ..... 27

1.10 Mammalian Lung Development ..... 29

1.11 Cell Growth in the Lungs ..... 33

1.11.1 Cell Growth Assays ..... 33

1.12 Inferring Protein Function ..... 36

2 Hypothesis and Aims ..... 38

3 Materials and Methods ..... 40

Cell Biology ..... 41

3.1 Human Primary Cell Culture ..... 41

3.2 Transient Transfection ..... 42

3.3 Proliferation Assays ..... 43

3.3.1 Cell Counting ..... 43

3.3.2 MTT assay ..... 44

3.3.3 Cell Counting kit 8 assay (CCK-8) ..... 44

3.3.4 Click-iT® EdU Microplate Assay Protocol ..... 45

3.3.5 CyQUANT® NF Assay Protocol ..... 46

Molecular Biology ..... 47

3.4 RNA and Protein Extraction ..... 47

3.4.1 Isolating RNA Lysate ..... 47
3.4.2 RNA Extraction from Lysate ..... 47
3.4.3 Harvesting Protein Lysate ..... 48
3.4.4 BioRad Protein Determination Assay ..... 49
3.5 cDNA Synthesis ..... 51
3.7 PCR and Quantitative-PCR ..... 53
3.7.1 Polymerase Chain Reaction ..... 53
3.7.2 Quantitative-PCR ..... 54
Transgenic Mouse Model Approaches ..... 56
3.8 Gstcd Knock-out Mice ..... 56
Methods to Infer Structure and Function of Proteins ..... 57
3.9 Comparative Protein Modelling ..... 57
Protein Analyses ..... 59
3.10 SDS-PAGE and Western Blotting ..... 59
3.10.1 SDS-PAGE ..... 60
3.10.2 Western Blotting ..... 60
3.11 Immunohistochemistry ..... 62
3.12 Haematoxylin and Eosin Staining ..... 64
Statistical Analyses ..... 65
3.13 Statistics ..... 65
4 Results ..... 66
4.1 GSTCD siRNA Knock-Down in Human Bronchial Epithelial Cells Results in Reduced Cell Growth ..... 67
4.1.1 GSTCD Antibody Optimisation ..... 68
4.1.2 Confirmation of GSTCD siRNA Knock-down ..... 71
4.1.3 Cell Growth Assay Optimisation ..... 76
4.1.4 Investigating the Effect of GSTCD Knock-down on Cell Growth in hBECs ..... 82
4.1.5 Discussion ..... 85
4.2 Knock-out of *GSTCD* Does Not Result in Gross Phenotypic Changes in Mice ..... 97
4.2.1 Gstcd Knock-out Confirmation ..... 98
4.2.2 Gstcd Localises in the Bronchial Epithelium in Mouse Lung Tissue 99

4.2.3 Gross Lung Structure Does Not Appear Morphologically Altered in Adult Gstcd Knock-out Mice ..... 101
4.2.4 Mouse Lung Development Does Not Appear Adversely Affected Morphologically by Gstcd Knock-out ..... 103
4.2.5 Optimising Antibody Markers of Cell Growth in Mouse Lung Tissue ..... 106
4.2.6 Gstcd Protein Expression in Gstcd Knock-out Mouse Lung ..... 108
4.2.7 Discussion ..... 112
4.3 GSTCD Protein Structure ..... 119
4.3.1 Secondary Structure Prediction Suggests a Core Region for GSTCD ..... 120
4.3.2 Conserved Domains Predicted for GSTCD Suggest Methyltransferase Activity ..... 132
4.3.3 Intrinsic Disorder Prediction Suggests Minimal Degree of Disorder in GSTCD Protein Structure ..... 140
4.3.4 Discussion ..... 141
5 Discussion ..... 147
5.1 Conclusions ..... 149
5.2 Limitations ..... 151
5.3 Future Directions ..... 152
6 Appendix ..... 154
6.1 Experimental Reagents ..... 155
6.1.1 PCR ..... 155
6.1.2 SDS-PAGE and Western Blotting ..... 156
6.1.3 Immunohistochemistry ..... 159
6.2 Sequence Alignments ..... 160
6.2.1 GSTCD Antibody Immunogen Alignments ..... 160
6.2.2 GSTCD Sequence Alignment to GST tag ..... 162
6.2.3 GSTCD siRNA Sequence Alignments ..... 164
6.3 Additional GSTCD Localisation in Mouse Lung Tissue ..... 170
6.4 Gstcd Knock-out Mice ..... 171
6.4.1 Gstcd Knock-out Genotyping Confirmation ..... 171
6.4.2 Necropsy Analysis of Gstcd Knock-out Mice ..... 171
6.5 Western Blotting Results ..... 175

6.5.1 Quantification of GSTCD Knock-down hBEC samples in Western Blotting ..... 175
6.6 Inferred GSTCD Homologous Proteins ..... 176
6.6.1 GSTCD Top 10 Homologous Proteins Inferred from Protein Prediction Servers ..... 176
7 References ..... 179

# List of Figures

Figure 1. The respiratory disorder emphysema results in enlarged air spaces and is characterized by destruction of lung tissue, specifically by the break-down of alveolar membranes. ..... 4

Figure 2. The respiratory disorder chronic bronchitis results in reduced airway diameter and is characterised by airway inflammation and excess mucus..... 5

Figure 3. Genotyping in GWA studies can identify SNPs as risk factors for disease in large cohorts (Hirschhorn & Daly 2005). ..... 8

Figure 4. Regional association plots for chromosome 4q24 locus associated with lung function for SpiroMeta (Repapi et al. 2010) (A) and CHARGE (Hancock et al. 2010) (B) consortia. .....11

Figure 5. Graphical representations of *GSTCD* DNA and protein sequence architecture for the two experimentally confirmed variants. .....17

Figure 6. GSTCD protein expression in human adult control and COPD lung tissue (Obeidat et al. 2013). .....21

Figure 7. GSTCD protein expression in human foetal lung tissue (Obeidat et al. 2013). .....22

Figure 8. siRNA mediated knock-down of gene expression (adapted from Santa Cruz Biotechnology RNAi 2015). .....24

Figure 9. International Mouse Phenotyping Resource of Standardised Screens (IMPReSS) phenotyping pipeline for knock-out mice within the IMPC (IMPReSS 2015). .....28

Figure 10. Stages of lung development in humans (Bancalari & Polin 2012). .....31

Figure 11. Diagram displaying the Gstcd$^{tm1(NCOM)}$ reporter-tagged deletion allele with selection cassette (coloured shapes) utilised to create a Gstcd knock-out mouse in the NorCOMM project (IMPC: GSTCD(NCOM) 2015; Grimm 2006). .....56

Figure 12. Two antibodies identify GSTCD protein in hBEC cell lysates by Western blotting. .....70

Figure 13. GSTCD siRNA knock-down in hBECs. .....72

Figure 14. GSTCD protein levels are reduced in hBEC lysates following siRNA knock-down. .....74

Figure 15. Semi-quantification of GSTCD protein levels following siRNA B (A) ($p<0.001$) and C (B) ($p<0.01$) knock-down in hBECs relative to controls. ....75
Figure 16. Comparison of cell growth assays using hBECs cultured in minimal and optimal media for 24h shows modest increase of cell growth in optimal media conditions.....78
Figure 17. Raw data for Click-iT® EdU and CyQUANT® NF assays demonstrate expected increase in cell growth for hBECs cultured in optimal media compared to minimal media across a time-course.....80
Figure 18. Click-iT® EdU and CyQUANT® NF assays demonstrate expected increase in cell growth for hBECs cultured in optimal media compared to minimal media across a time-course. .....81
Figure 19. GSTCD siRNA knock-down in hBECs does not affect proliferation rate (A) but does result in reduced total cell number (B) but over time.. .....83
Figure 20. GSTCD siRNA knock-down in hBECs results in reduced total cell number in a second hBEC donor.....84
Figure 21. LacZ staining of lung tissue provides localisation information in Gstcd +/- and Gstcd -/- transgenic mice where Gstcd expression is replaced by LacZ (Cornell University, Frank Lee). .....98
Figure 22. Gstcd protein expression indicates bronchial epithelial localisation as assessed by IHC in Gstcd +/+ C57BL/6N mouse lung tissue. ..... 100
Figure 23. Loss of Gstcd does not result in gross phenotypic effects in adult mouse lung tissue. ..... 102
Figure 24. Preliminary lung histology investigation by H&E staining indicates unaltered structure in Gstcd -/- mice at E14.5..... 104
Figure 25. Preliminary lung histology investigation by H&E staining indicates unaltered structure in Gstcd -/- mice at E18.5..... 105
Figure 26. Proliferating cells were not detectable in wild-type mouse lung tissue using IHC. ..... 107
Figure 27. GSTCD expression was apparent in lung tissue in GSTCD +/+ (n=2), +/- (n=1) and -/- (n=3) lung tissue using IHC..... 110
Figure 28. Top 3D protein models built for human GSTCD variants 1 and 2 using the protein prediction servers: I-TASSER (A), SWISS-MODEL (B) and Phyre2 (C)..... 122

Figure 29. Representative examples of primary homology region identified by the three protein prediction servers: I-TASSER, SWISS-MODEL and Phyre2. .. 135
Figure 30. Top 3D protein models built for human GSTM5 using the protein prediction servers: I-TASSER (A), SWISS-MODEL (B) and Phyre2 (C). ..... 137
Figure 31. Graphical representation of GSTCD conserved domain homology. ... 144
Figure 32. Immunostaining for Gstcd in mouse lung tissue using an alternative primary antibody (Santa Cruz, sc-138332) performed by Michael Kotlikoff research group, Cornell University. ..... 170
Figure 33. Genotyping results for Gstcd +/+ and Gstcd -/- mice. ..... 171

# List of Tables

Table 1. Significant associations identified at locus 4q24 by meta-analysis and GWA studies. ..... 12

Table 2. Summary of GSTCD expression levels in human tissue. Information collated from Protein Atlas from immunohistochemistry protein staining (The Human Protein Atlas 2015) and Ensembl from RNA sequencing data (Ensembl Human GSTCD 2015). ..... 19

Table 3. Fetal lung gene array data comparing GSTCD and INTS12 expression during pseudoglandular and canalicular stages of lung development (Obeidat et al. 2013). ..... 22

Table 4. Stages and events during lung development in humans and mice (Bancalari & Polin 2012; Koopman 2010; Warburton et al. 2010). ..... 32

Table 5. Click-iT® EdU microplate assay kit manufacturer guidance for producing 2X Click-iT® reaction cocktail components (volumes sufficient for one 96-well plate). ..... 46

Table 6. Click-iT® EdU microplate assay kit manufacturer guidance for producing Amplex® UltraRed reaction mixture (volumes sufficient for one 96-well plate). ..... 46

Table 7. Reagents and volumes used for protein lysis solution for extracting protein from cells grown *in vitro*. ..... 49

Table 8. Reaction components for SuperScript™ II First Strand cDNA Synthesis kit after incubation of RNA/primer mix. ..... 52

Table 9. Temperature and incubation length conditions for cDNA synthesis from RNA using SuperScript™ II reverse transcriptase. ..... 52

Table 10. Sequence information for use of GAPDH primers in PCR. ..... 53

Table 11. Stages of PCR cycles and conditions used for cDNA amplification. ..... 53

Table 12. GSTCD primer and probe sequences (Invitrogen) used for TaqMan® Q-PCR, as used for GSTCD amplification previously (Obeidat et al. 2013). ..... 55

Table 13. Q-PCR template master mix of cDNA samples and associated controls. .....55

Table 14. Q-PCR primer master mix of PCR reagents for amplification of template cDNA samples (Table 13). .....55

Table 15. Primary antibodies and associated isotype controls used in IHC.....63

Table 16. Summary comparison of Click-iT® EdU and CyQUANT® NF assays outlining the purpose and advantages of each assay.....91

Table 17. Proteins suggested to have binary interactions with GSTCD (Weimann et al. 2013) show similar reduction in cell viability in RNA interference expression knock-down studies.....95

Table 18. Phenotype data for male and female Gstcd$^{EUCOMM}$ -/- knock-out mice assessed within the IMPC by the IMPReSS phenotyping pipeline (IMPC: GSTCD 2015). ..... 115

Table 19. Investigating GSTCD protein sequence using web-based protein structure prediction servers indicates methyltransferase homology..... 133

Table 20. Investigating GSTM5 protein sequence using web-based protein structure prediction servers as control results. ..... 136

Table 21. PCR reagents and volumes used for amplification of cDNA..... 155

Table 22. Buffer reagents and volumes used for SDS-PAGE gel electrophoresis and Western blotting of protein samples. ..... 156

Table 23. SDS PAGE gel reagents combined in sequence to produce two gel layers. ..... 158

Table 24. Buffer reagents and volumes used for IHC. ..... 159

Table 25. Report summarising findings of necropsy analysis of major organs, bones and glands from Gstcd +/+ and Gstcd -/- knock-out mice performed by collaborators at Cornell University (Teresa Southard). ..... 172

Table 26. Quantification of Western blotting band intensities reveals a significant reduction of GSTCD protein in siRNA B and siRNA C treated hBECs relative to controls. ..... 175

Table 27. Top 10 homology templates against the protein sequences of human GSTCD variants 1 and 2 as assessed by web-based protein structure prediction servers..... 177

# List of Abbreviations

-/- - homozygote knock-out genotype

+/- - heterozygote knock-out genotype

+/+ - wild-type genotype

BEBM - bronchial epithelial basal growth medium

BEGM - bronchial epithelial cell growth medium

BLAST - basic local alignment search tool

CASP - critical assessment of protein structure prediction

CCK-8 - Cell Counting kit 8

CHARGE - cohorts for heart and aging research in genomic epidemiology

COPD - chronic obstructive pulmonary disease

DAB - 3,3'-diaminobenzidine

DEPC - diethylpyrocarbonate

E[9.5] - embryonic day

EdU - 5-ethynyl-2′-deoxyuridine

EEFE1/AIMP3 - eukaryotic elongation translation factor 1 epsilon 1

EMBL-EBI - European molecular biology laboratory - European bioinformatics suite

ESC - embryonic stem cells

$FEV_1$ - forced expiratory volume in 1 second

FVC - forced vital capacity

GAPDH - glyceraldehyde-3-phosphate dehydrogenase

GST - glutathione S-transferase

GSTCD - glutathione S-transferase C-terminal domain containing

GWA - genome wide association

hBECs - human bronchial epithelial cells

HBSS - Hank's balanced salt solution

IgG - immunoglobulin G

IHC - immunohistochemistry

IKMC - international knock-out mouse consortium

IMPC - international mouse phenotyping consortium

IMPReSS - international mouse phenotyping resource of standardised screens

I-TASSER - iterative threading assembly refinement

ITS - insulin, transferrin and sodium selenite

KDM1A - lysine histone demethylase 1A

lacZ - β-galactosidase gene

MM - minimal media

MTT - 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide

NCBI - national centre for biotechnology information

NKX2 - NK2 homeobox

NorCOMM - North American conditional mouse mutagenesis project

OM - optimal media

PAGE - polyacrylamide gel electrophoresis

PBS - Dulbecco's phosphate buffered saline

PBS-BT - PBS – bovine serum albumin (BSA) – Triton™ X-100

PCP - planar cell polarity

PCR - polymerase chain reaction

PDAR – pre-designed assay reagent

PH3 - phospho histone H3

PRMT6 - protein arginine N-methyltransferase 6

ProSP-C - pro-surfactant protein C

Q-PCR - quantitative polymerase chain reaction

RCSB - research collaboratory for structural bioinformatics

RISC - RNA-induced silencing complex

RT - reverse transcriptase

SAM - S-adenosylmethionine

SDS - sodium dodecyl sulphate

siRNA - short interfering RNA

SNP - single nucleotide polymorphism

SUV39H2 - suppressor of variegation 3-9 homolog 2

# 1 Introduction

## 1.1 Lung Function

Spirometry is standardly used to assess lung function in humans, using the measures forced expiratory volume in 1 second ($FEV_1$) and forced vital capacity, i.e. the volume of air expired by a full expiration (FVC) (Vestbo et al. 2013). $FEV_1$ measures the volume of air that can be forcefully expelled from the lungs in the first second after a maximum inspiration. FVC measures the volume change of the lungs from a maximal inspiration to residual volume. These measures are used to identify different types of lung disease such as chronic obstructive pulmonary disease (COPD) based on the divergence of the ratio $FEV_1$/FVC from a healthy individual's average of 80% (Cooper & Mitchell 2003). When the ratio of $FEV_1$/FVC is under 70%, this is referred to as an obstructive defect where individuals typically have difficulty expelling air from the lungs with full exhalations, this indicates the presence of COPD (GOLD 2015). In contrast in restrictive lung disease where individuals have difficulty fully expanding the lungs during inhalation.

Interestingly, it has been demonstrated that spirometry measurements cluster within families, suggesting that lung function has a hereditary component which may influence the development of respiratory disease (Wilk et al. 2009; Zhai et al. 2007). For instance, hereditary factors are suggested to be attributable for between 20-60% of phenotypic variance in lung function measures (Pauwels et al. 2001; Lewitter et al. 1984; Palmer et al. 2001; DeMeo & Silverman 2003) and this is strongly correlated in twins (28% for $FEV_1$ and 59% for FVC) (Hallberg et al. 2010).

## 1.2 Chronic Obstructive Pulmonary Disease

The three key initiatives for COPD, the Global Initiative for Chronic Obstructive Lung Disease (GOLD), American Thoracic Society (ATS) and European Respiratory Society (ERS), guidelines classify COPD as a heterogeneous and debilitating disease characterised by persistent and irreversible airflow obstruction that is generally progressive (Vestbo et al. 2013; Celli et al. 2004). COPD is also associated with chronic inflammation and airflow obstruction is assessed by spirometry measures, as discussed previously (Vestbo et al. 2013; Celli et al. 2004).

COPD is a significant global health concern and burden. Notably, in 2013 COPD became the third most common cause of death in adults worldwide (Lozano et al. 2012) and critically, current therapies have only a moderate impact.

The development of COPD has a strong environmental basis, with cigarette smoking and exposure to poor air quality (biomass exposure) being key risk factors. However, unlike some common chronic diseases, the incidence of COPD has not declined in recent years, instead there continues to be increasing prevalence, morbidity and mortality rates for COPD globally. According to the World Health Organisation, 64 million people worldwide have COPD and more than 3 million people die each year of the disease (Mathers et al. 2008). Within the UK alone it is estimated that 3 million people have COPD and it accounts for 30,000 deaths each year (Rudolph 2010). Critically, COPD is now the third leading cause of death worldwide (Lozano et al. 2012).

In general, COPD is a progressive condition, leading to airway remodelling, inflammation and narrowing of the small airways and/or alveolar destruction (Vestbo et al. 2013). Symptoms predominantly become evident later in life, where individuals are most commonly diagnosed over the age of 50 (NHS 2014) and with a higher prevalence in men than women (Vestbo et al. 2013).

Although the introduction of smoking bans may have helped to lower the incidence of COPD in some countries, critically, not all individuals with COPD are smokers (Shahab et al. 2006). In addition to environmental exposures, as

described previously, variability in lung function is also proposed to be heritable (Pauwels et al. 2001; Lewitter et al. 1984; Palmer et al. 2001; DeMeo & Silverman 2003). Whilst there is currently a range of therapeutic agents available for treatment of COPD which can improve symptoms in some patients, none have been shown to alter the progression of underlying disease. In particular the primary sub-types of COPD which consist of emphysema (alveolar destruction leading to shortness of breath) (Figure 1) and chronic bronchitis (long term inflammation of the bronchial airways) (Figure 2) (Vestbo et al. 2013).

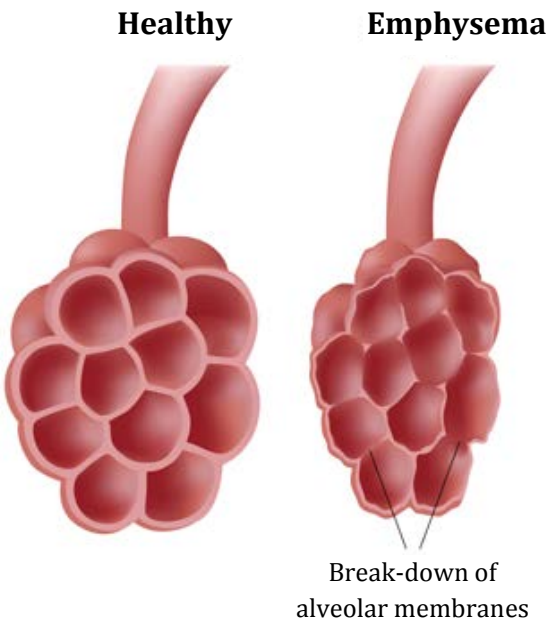



**Figure 1. The respiratory disorder emphysema results in enlarged air spaces and is characterized by destruction of lung tissue, specifically by the break-down of alveolar membranes.** The diagram indicates typical structure of lung alveoli in healthy and emphysema conditions (Breathing Center of Houston 2014). When compared to healthy lungs emphysema leads to a reduction in elasticity of the lungs caused by damage to and break-down of alveolar membranes (Sharafkhaneh et al. 2008). This results in enlarged air spaces which reduces overall lung volume, leading to a reduction in airflow and shortness of breath.

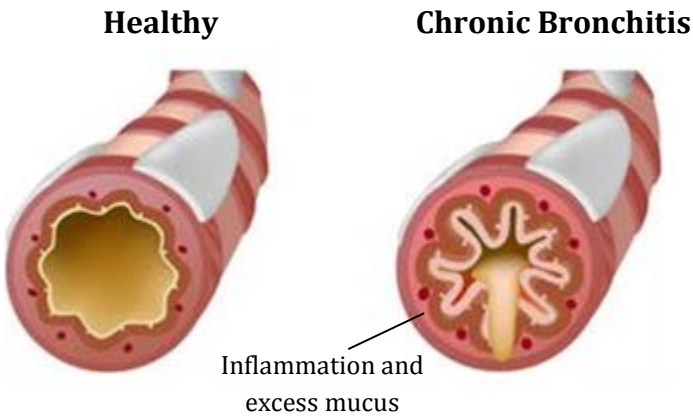



**Figure 2. The respiratory disorder chronic bronchitis results in reduced airway diameter and is characterised by airway inflammation and excess mucus.** The diagram indicates the changes observed in airways of healthy individuals and individuals diagnosed with chronic bronchitis (Ferranti 2014). The condition results in reduction of the diameter of the airways compared to healthy by a combination of overproduction and hypersecretion of mucus and inflammation of the airway walls (Kim & Criner 2013). This leads to accelerated decline in lung function by luminal obstruction of small airways, changes in airway surface tension predisposing to collapse and epithelial remodelling (Kim & Criner 2013).

## 1.3 Environmental and Genetic Factors of COPD

Traditionally, a primary risk factor for COPD is smoking; therefore smokers are characteristically prone to developing COPD. Estimates indicate that after 25 years of smoking 30-40% of smokers will have developed COPD (Løkke et al. 2006). Even non-smokers may be affected as a consequence of general exposure to air pollutants, for example an investigation into long term smoke particulate matter exposure revealed a significant association between an increase in exposure to small particles and a mild decrease in $FEV_1$ across 20 years (Abbey et al. 1998). In addition, biomass emissions are also a notable risk factor globally, in general consisting of smoke inhalation via indoor pollution or occupational exposure.

However, genetic predisposition is also a known risk factor for developing COPD which increases an individual's susceptibility. The most commonly studied genetic example in COPD is α1-antitrypsin deficiency, where individuals are homozygous for a deleterious mutation in SERPINA1 (Primhak & Tanner 2001; Luisetti & Seersholm 2004). Typically, α1-antitrypsin targets and inhibits human neutrophil elastase enzymes. In α1-antitrypsin deficiency neutrophil elastase activity is enhanced relative to normal conditions, ultimately leading to destruction of the alveoli and reducing lung function. This can eventually lead to emphysema in individuals as young as 30 (Needham 2004). 1-2% of COPD cases are attributable to this mutation (DeMeo & Silverman 2004) and approximately 1.1 million cases of severe AAT deficiency are estimated worldwide (Luisetti & Seersholm 2004).

Additionally, the majority of COPD patients are smokers, but not all smokers develop COPD, therefore further indicating a genetic predisposition to COPD (Postma & Boezen 2004) in addition to hereditary factors contributing to lung function. Importantly, only ~15-20% of smokers develop COPD (Wilk et al. 2009; Warburton et al. 2006; Tzortzaki et al. 2006; Davis & Novotny 1989) hence potentially indicating several heritable components or interactions between genetics and environmental factors in the development of COPD.

## 1.4 Genome Wide Association Studies

Genome wide association (GWA) studies are a valuable method for identifying common genetic risk factors underlying complex diseases, such as for COPD, by using direct association with potentially causative variation (for example lung function measures). Specifically, spirometric measures of lung function have frequently been reported to have a significant heritable basis and may indicate predisposition to a respiratory disease (Hancock et al. 2010; Silverman et al. 1998; Silverman et al. 2002). GWA studies use these phenotypic measures to identify associations with common risk alleles which influence the probability of developing a respiratory disease (Ober and Yao, 2011). Moreover, spirometric measures of lung function are quantitative phenotypes.

In GWA studies, genotyping identifies single nucleotide polymorphisms (SNPs) to search the most common genetic variation in the genome as the risk factors for disease or for a role in disease in large populations (cohorts) (Hall 2009; Hirschhorn & Daly 2005). The SNPs identified are used to tag linkage blocks within the genome or may be causative themselves (Figure 3) and the statistical significance that a SNP associates with a particular phenotype more than by chance can be tested. Hence, it is possible to predict an individual's potential susceptibility to a specific disorder such as COPD as a result of the type of genetic variant present, although this cannot necessarily predict that a person will definitively develop the disorder.

A particular advantage of this method is that it is a hypothesis free approach looking at the whole genome, thereby allowing new associations to be identified (Obeidat & Hall 2011; Soler Artigas, Wain, et al. 2011). In an aim to advance the strength of GWA studies, consortia have recently been formed, combining individual studies through evaluation by meta-analysis which increases the strength and ability to detect significant gene associations with confidence (Tamari et al. 2011). This is important as the use of large consortia helps to reduce the number of false positive associations identified and enable true significant associations to be seen by vastly increasing cohort numbers.

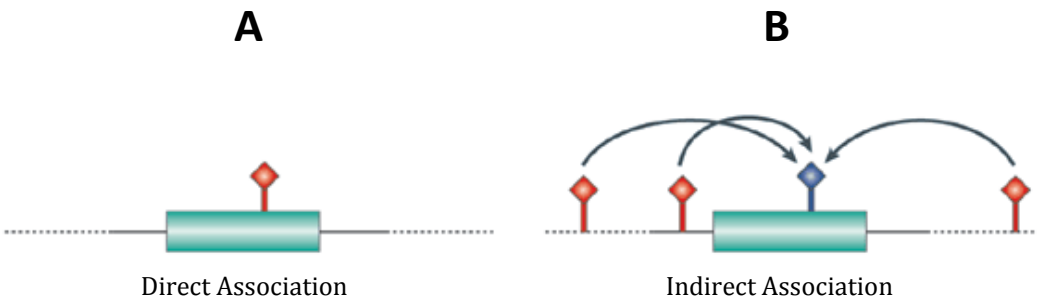



**Figure 3. Genotyping in GWA studies can identify SNPs as risk factors for disease in large cohorts (Hirschhorn & Daly 2005).** In this diagram genes are shown in green and SNPs as red and blue. A direct association may be tested using SNPs in a candidate gene based on prior knowledge of possible function; for example investigating a SNP with a missense mutation suggested to affect the function of a gene (A). However an increasingly common and hypothesis free approach is to investigate indirect association (B). In this instance SNPs are selected for genotyping on the basis of linkage disequilibrium patterns to provide information about many additional SNPs, known as a linkage block. Here, association is tested indirectly at the SNP in blue, as it is in linkage disequilibrium with three other SNPs.

### 1.4.1 Genome Wide Association Studies Identify Significant Association between *GSTCD* and $FEV_1$

To maximise the power to detect risk alleles with modest effect sizes the results of individual GWA studies can be combined in a meta-analysis. By meta-analysing many GWA studies, researchers have revealed a number of SNPs within/near genes which are significantly associated with the lung function measure $FEV_1$. Two large consortia for respiratory research: SpiroMeta consortium and Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium conducted meta-analyses of GWA studies which identified a selection of loci associated with lung function, the most significant of which were novel variants at locus 4q24 (*INTS12-GSTCD-NPNT* region) (Table 1) (Repapi et al. 2010; Hancock et al. 2010). Joint meta-analysis in these cohorts concluded that an intronic SNP at locus 4q24 (rs10516526) in the gene *GSTCD* had the strongest significant association with lung function measures, as assessed by $FEV_1$ ($p=2.18 \times 10-23$) (Figure 4) (Repapi et al. 2010; Hancock et al. 2010). The magnitude of effect on $FEV_1$ at this SNP in *GSTCD* was 52ml per copy of a G allele (Repapi et al. 2010). This change in lung function is equivalent to approximately 3 years of $FEV_1$ decline in non-smokers (Kohansal et al. 2009). This was not repeated in association with $FEV_1$/FVC (Repapi et al. 2010; Hancock et al. 2010). Importantly, the genetic effects seen appear to be independent of both smoking and disease status (Repapi et al. 2010). Therefore, GSTCD may potentially influence the development or severity of COPD and could also be important in other obstructive diseases of the lung.

On a similar note, a significant association has also been suggested with a cardiovascular disease outcome (a major contributor to mortality and morbidity in COPD) at the SNP locus rs10516526 (Levinsson et al. 2014). Results show a significant adverse effect of air pollution exposure on acute myocardial infarction for individuals with AA genotype at this locus. Thus it seems that the G allele at this locus may be beneficial both for lung function and protection against cardiovascular disease (Levinsson et al. 2014). Additionally, several subsequent studies also emphasise *GSTCD* significance with lung function, as many *GSTCD* SNPs have been significantly associated with $FEV_1$, FVC and COPD (Table 1)

(Castaldi et al. 2011; Soler Artigas, Wain, et al. 2011; Soler Artigas, Loth, et al. 2011; Imboden et al. 2012; Hancock et al. 2012; Li et al. 2013; Loth et al. 2014; Yao et al. 2014; Kreiner-Møller et al. 2014; Wain et al. 2015; Artigas et al. 2015; Brehm et al. 2015).

A broad association signal spanning several genes can be detected at locus 4q24 due to a cluster of genes (Figure 4). However, GSTCD was selected for further research on the basis of SpiroMeta-CHARGE joint meta-analysis of n=48,201 individuals showing the strongest association at SNP rs10516526 in GSTCD for $FEV_1$ at the 4q24 locus. Furthermore, subsequent reports at this locus by SpiroMeta investigators and others have also implicated GSTCD most frequently for association with COPD, FVC, acute myocardial infarction and $FEV_1$ in children; strongly indicating an important role for this gene in particular (Table 1).

GSTCD expression levels have also been shown to correlate with lung function measures (Obeidat et al. 2013). Whereby the direction of correlations suggests that higher mRNA levels correlate with better lung function, in this instance indicated by higher percent predicted $FEV_1$ (Obeidat et al. 2013). Interestingly, investigation of GSTCD protein expression levels across lung development revealed a decrease in expression between the pseudoglandular and canalicular stages (Obeidat et al. 2013) (lung development stages outlined in section 1.10 Table 4). This was also correlated at the mRNA level where GSTCD expression significantly decreased with age between the pseudoglandular and canalicular stages (Obeidat et al. 2013). Hence, it could be suggested that GSTCD has a role during lung development, further demonstrating GSTCD as an important gene in the respiratory system.

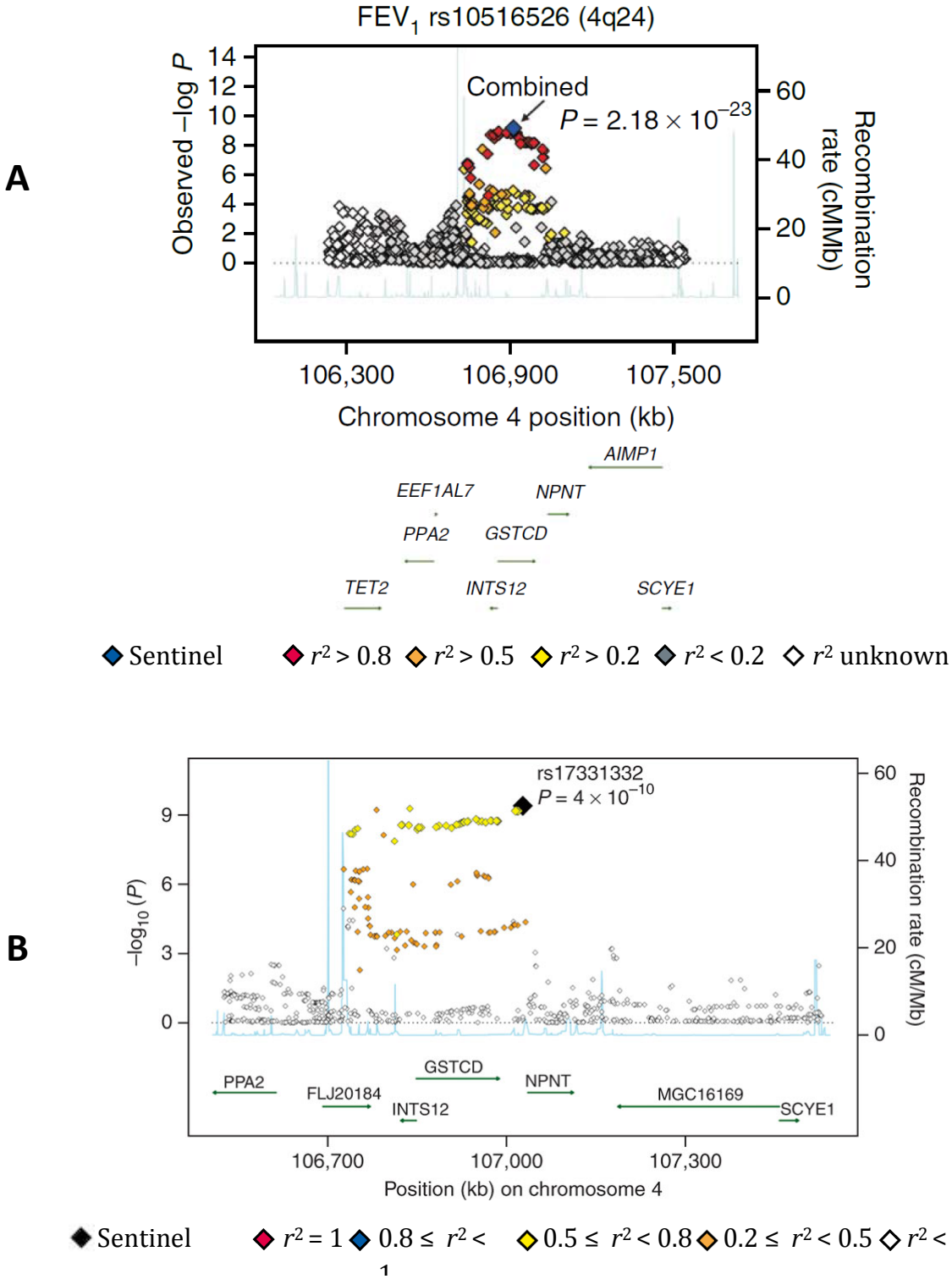



**Figure 4. Regional association plots for chromosome 4q24 locus associated with lung function for SpiroMeta (Repapi et al. 2010) (A) and CHARGE (Hancock et al. 2010) (B) consortia.** Association shown between SNPs at locus 4q24 and $FEV_1$ at genome wide significance, the sentinel/top SNP and correlations ($r^2$) with surrounding SNPs are shown in the indicated colours for each plot. Combined *P* values through all stages of each GWA studies meta-analysis are indicated adjacent to the sentinel SNP (*GSTCD*).

**Table 1. Significant associations identified at locus 4q24 by meta-analysis and GWA studies.**

| Gene | SNP | Association | Reference |
|---|---|---|---|
| *FLJ20184/INTS12/ GSTCD* | rs17035917 | $FEV_1$ | Repapi et al. 2010 |
| *GSTCD* | rs10516526 | $FEV_1$ | |
| *GSTCD* | rs10516526 | FVC | |
| *GSTCD* | rs11731417 | $FEV_1$ | |
| *GSTCD* | rs11727735 | $FEV_1$ | |
| *FLJ20184* | rs17036052 | $FEV_1$ | Hancock et al. 2010 |
| *FLJ20184* | rs17035960 | $FEV_1$ | |
| *FLJ20184/INTS12/ GSTCD* | rs17035917 | $FEV_1$ | |
| *INTS12* | rs11727189 | $FEV_1$ | |
| *INTS12* | rs17036090 | $FEV_1$ | |
| *GSTCD* | rs11097901 | $FEV_1$ | |
| *GSTCD* | rs11728716 | $FEV_1$ | |
| *GSTCD* | rs11731417 | $FEV_1$ | |
| *GSTCD* | rs11727735 | $FEV_1$ | |
| *NPNT* | rs17331332 | $FEV_1$ | |
| *NPNT* | rs17036341 | $FEV_1$ | |
| *FLJ20184/INTS12/ GSTCD* | rs4235415 | COPD | Castaldi et al. 2011 |
| *NPNT* | rs11933466 | COPD | |
| *GSTCD* | rs10516526 | COPD | Soler Artigas et al. 2011 |
| *NPNT* | rs6823809 | $FEV_1$/FVC | Soler Artigas, Loth, et al. 2011 |
| *GSTCD* | rs11731417 | $FEV_1$ | Imboden et al. 2012 |
| *INTS12/GSTCD/NPNT* | rs17331332 | $FEV_1$ | Hancock et al. 2012 |

| | | | |
|---|---|---|---|
| *FLJ20184/INTS12/ GSTCD* | rs17035917 | $FEV_1$ | Li et al. 2013 |
| *FLJ20184/INTS12/ GSTCD* | rs6820671 | $FEV_1$ | |
| *INTS12/GSTCD* | rs1982346 | $FEV_1$ | Ong et al. 2013 |
| *GSTCD* | - | FVC | Loth et al. 2014 |
| *GSTCD* | rs10516526/ rs11726124 | $FEV_1$ | Yao et al. 2014 |
| *FLJ20184/INTS12/ GSTCD/NPNT* | rs17331332/ rs7693333 | $FEV_1$ | |
| *NPNT* | rs11727189/ rs11097903 | FVC | |
| *FLJ20184* | rs17036076 | $FEV_1$ | |
| *INTS12* | rs17035960/ rs17036090 | $FEV_1$ | |
| *GSTCD* | rs11728716/ rs11726124 | $FEV_1$ | |
| *GSTCD* | rs10516526 | Acute Myocardial Infarction | Levinsson et al. 2014 |
| *GSTCD* | rs10516526 | $FEV_1$ at age 7 years, lung growth | Kreiner-Møller et al. 2014 |
| *NPNT* | rs34712979 | $FEV_1$ | Wain et al. 2015 |
| *TET2* | rs2047409 | $FEV_1$ | |
| *NPNT* | rs6856422 | $FEV_1$ | Artigas et al. 2015 |
| *FLJ20184/INTS12/ GSTCD/NPNT* | rs17036129 | Post-bronchodilator $FEV_1$ at age 6-14 years | Brehm et al. 2015 |
| *INTS12/GSTCD* | rs11731417 | Post-bronchodilator $FEV_1$ at age 6-14 years | |
| *NPNT* | rs17036139 | Post-bronchodilator $FEV_1$ at age 6-14 years | |
| *NPNT* | rs11724839 | Post-bronchodilator $FEV_1$ at age 6-14 years | |

## 1.5 Biological Function of GSTCD

Currently, there are no reports describing the function or confirming specific cellular localisation of GSTCD, although it has been indicated that GSTCD may belong to the Glutathione S-Transferase (GST) family of soluble enzymes (of which the gene is named after) due to a shared C-terminal α-helical domain (Repapi et al. 2010).

GST enzymes belong to two distinct multi-gene families; cytosolic and membrane associated, as this function is thought to have evolved on two separate occasions (Sherratt & Hayes 2001). These two GST families indicate virtually no similarity at the primary protein structure level (Sherratt & Hayes 2001). The three key families of proteins that exhibit GST activity are; cytosolic, mitochondrial (cytosolic soluble enzymes) and microsomal (membrane associated enzymes) (Hayes et al. 2005). These enzymes are classically viewed as part of cell defence against a number of harmful chemicals produced endogenously and in the environment, whereby the general reaction involves conjugating glutathione to a variety of endogenous electrophilic targets (Strange et al. 2001; Hayes et al. 2005). Specifically, the largest GST family, soluble GSTs, are primarily involved in the metabolism of foreign chemicals, for instance: environmental pollutants, carcinogens, and cancer chemotherapeutic drugs, as well as detoxifying potentially harmful endogenous reactive compounds (Sherratt & Hayes 2001). Eight classes of soluble GST enzymes are known to exist in mammals, classified by protein sequence identity, these include: alpha, mu, pi, sigma, theta, zeta, omega and kappa, the majority of which localise in the cytoplasm.

The predicted GST function complements reasons for selecting *GSTCD* as the candidate gene from locus 4q24. This anticipated role of a defence mechanism against toxins and carcinogens in the lungs fits with the predicted role of a G allele in GSTCD being protective in relation to $FEV_1$. However, investigation of classical GST activity in a GSTCD over-expression system revealed no change in GST activity (SensoLyte® GST Activity assay) (unpublished, Ceri Stewart, University of Nottingham). The SensoLyte® GST Activity assay is used to assess levels of GST activity in varied samples and conditions as well as for investigating GST substrates (Jirkovsky et al. 2012; Yamaoka & Kitamura 2015). Two vectors expressing human GSTCD were previously transfected into two human airway cell types: airway smooth muscle (hASM) and bronchial epithelial cells (hBECs) alongside transfection of the positive control GST Mu5 (GSTM5) (n=3, unpublished, available on request). This revealed no discernible change in GST activity in either GSTCD over-expression condition (relative to empty vector and transfection reagent only negative controls) in either cell type. Whereas, corresponding over-expression of GSTM5 resulted in 2-3 times increased GST expression relative to the negative controls. Therefore, indicating that GSTCD is less likely to have typical GST activity, although this assay does not assess all possible variations of GST mediated reactions.

Additional functional theories for GSTCD are guided by protein homology searches which speculate that GSTCD shares similar motifs with methyltransferase enzymes, GST enzymes and Eukaryotic Elongation Translation Factor 1 epsilon 1 (EEFE1/AIMP3); a component of the aminoacyl tRNA synthase complex (Obeidat et al. 2013). GSTCD is also suggested to have binary interactions with Lysine histone demethylase 1A (KDM1A), Protein arginine N-methyltransferase 6 (PRMT6) and Histone-lysine N-methyltransferase (suppressor of variegation 3-9 homolog 2, SUV39H2) identified by co-immunoprecipitation and yeast two-hybrid interaction screening approach (Weimann et al. 2013). Overall, little is currently known about the functional activity of GSTCD or its potential interactions.

## 1.6 *GSTCD* Gene Structure and Conserved Domains

According to the two main genome and biotechnology browsers, Ensembl and the National Centre for Biotechnology Information (NCBI), there are two experimentally identified splice variants for *GSTCD,* encoding two separate proteins (variant 1 - 633aa and variant 2 - 546aa) (Figure 5) with variant 1 being the primary transcript. Nevertheless, 9 additional transcripts have been postulated to exist based on manual curation, of which 6 are suggested to also be protein coding (Ensembl Human GSTCD 2015). This project will focus on GSTCD transcripts 1 and 2 as these are the only two experimentally confirmed transcripts. Both transcripts consist of 12 exons with 11 coding exons producing the two protein variants (Figure 5). Although, the functional activity of these proteins remains unknown, hence protein homology searches for functional domains have been the main guide for theoretical functions. Typically, a conserved domain (invariant protein sequence motif) is common to a particular group of proteins sharing a similar function. Consequently, if a specific conserved domain is identified in an unknown protein it can be inferred that the function may be similar to that of other proteins of known function which contain the domain.

Interestingly, GSTCD appears to have only one of the two domains characteristic of the GST family of enzymes (Nebert & Vasiliou 2004) (C-terminal α-helical domain) and lacks the N-terminal thioredoxin-type domain. Furthermore, glutathione is thought to bind GST enzymes between the N- and C-terminal domains, therefore indicating that both domains are necessary for GST activity and that GSTCD may simply have GST-like activity (Nebert & Vasiliou 2004). Alternatively, according to NCBI Basic Local Alignment Search Tool (BLAST) conserved domain search a methyltransferase domain is found to have many hits and the top homology score was with tRNA (uracil-5-) methyltransferase homologue A isoform B.

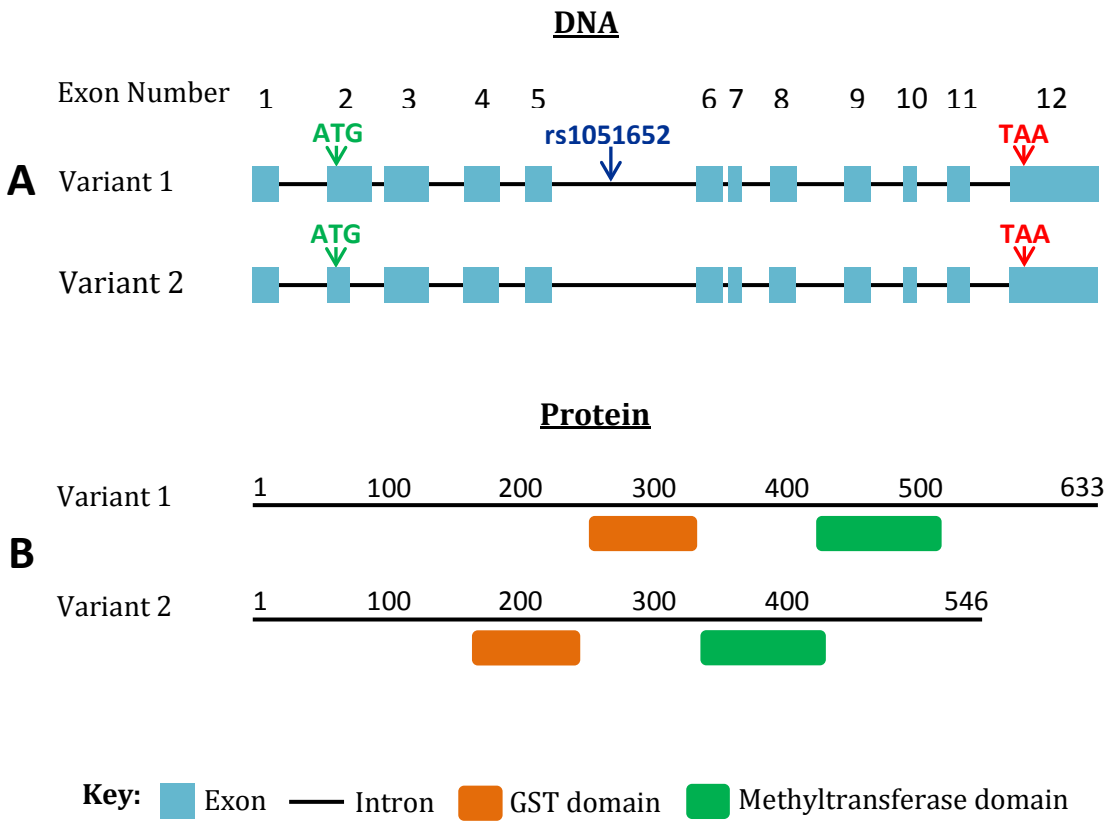



**Figure 5. Graphical representations of *GSTCD* DNA and protein sequence architecture for the two experimentally confirmed variants.** *GSTCD* is considered to consist of a total of 12 exons of which 11 are coding for both published transcripts (A). Approximate locations of start (ATG) and stop (TAA) codons are indicated in green and red text respectively and the intronic sentinel SNP in *GSTCD* identified by two meta-analyses of GWA studies is indicated in blue text (rs10516526, associated with $FEV_1$ p=2.18x10-23) (Repapi et al. 2010; Hancock et al. 2010). Variant 2 shows a truncated exon 2 compared to variant 1 (A) consequently resulting in a truncated variant 2 protein. Conserved domains identified in human GSTCD variant 1 and 2 protein sequences include GST C-terminal α-helical domain (251-332aa and 164-245aa respectively) and S-adenosylmethionine-dependent methyltransferase domain (422-516aa and 335-429aa respectively) (B) (Ensembl: GSTCD Domains and Features. 2014). Translated protein lengths are 633aa (variant 1) and 546aa (variant 2).

## 1.7 GSTCD Expression Profiles and Localisation in Humans

There are currently few studies directly investigating GSTCD, providing little information for overall expression profiles and information is largely obtained as a result of genome wide general searches in disease studies. Therefore inferred locations of expression in human tissue are adapted primarily from Protein Atlas (explores the expression profile of proteins in human tissues using an antibody-based approach) and Ensembl (genome browser) GSTCD expression data, shown collated in Table 2. Protein Atlas determined varied GSTCD expression levels from immunohistochemistry staining and Ensembl utilised RNA sequencing to detect the levels of GSTCD mRNA. Although, the Protein Atlas data is derived from immunohistochemistry using only a single antibody, hence may not be fully conclusive. GSTCD expression levels were identified as highest in Ensembl in the tissues of: cerebellum, lateral ventricle, uterus, epididymis and skin. As of yet the relevance of high GSTCD expression in these tissues remains unknown.

With focus of GSTCD research based on the respiratory system due to the significant associations of GSTCD with $FEV_1$ and COPD, GSTCD expression has been confirmed in human airway smooth muscle cells, bronchial epithelial cells, peripheral blood mononuclear cells and whole lung tissue by quantitative polymerase chain reaction (Q-PCR) (Obeidat et al. 2013). In this comparison of lung cell types human bronchial epithelial cells (hBECs) were found to have the highest level of GSTCD expression (relative to whole lung tissue) (Obeidat et al. 2013).

A previous study in human lung tissue revealed fairly ubiquitous GSTCD expression with localisation predominantly in the bronchial epithelium and alveolar pneumocytes (alveolar epithelial cells) (Figure 6) (Obeidat et al. 2013). Specifically, the sub-cellular localisation of GSTCD is thought to be the cytoplasm (The Human Protein Atlas 2015) concurring with the cytoplasmic localisation of known soluble GSTs.

Furthermore, staining patterns indicate variation in expression levels seen in human lung tissue across development (Figure 7). GSTCD protein expression was

noted to be increased from the early embryonic to the late embryonic stage within developing blood vessels, alveolar epithelial cells and outer region of bronchioles, followed by a decrease in expression at the canalicular stage (Obeidat et al. 2013). Notably, this decrease was also found to be evident at the mRNA level at the same developmental stage (Table 3) (Obeidat et al. 2013) hence possibly indicating an important role for GSTCD in lung development. The cause for this reduction in expression is so far unknown, although could be significant during the pseudoglandular stage (Introduction section 1.10 Table 4) as airway cells become specialised for function during this period.

Potential approaches used to investigate unknown GSTCD function in a relevant human airway cell system include studying the effects of reducing GSTCD protein expression. Specifically, performing protein reduction experiments in a bronchial epithelial cell type, based on Obeidat et al. data on GSTCD localisation in the bronchial epithelium in human lung tissue.

**Table 2. Summary of GSTCD expression levels in human tissue.** Information collated from Protein Atlas from immunohistochemistry protein staining (The Human Protein Atlas 2015) and Ensembl from RNA sequencing data (Ensembl Human GSTCD 2015). Key: High, medium, low and no detectable expression level and – no data available.

| Organ System | Tissue | Protein Atlas | Ensembl |
|---|---|---|---|
| **Blood and Immune System** | Bone Marrow | | |
| | Lymph Node | | |
| | Spleen | | |
| | Tonsil | | - |
| **Cardiovascular System** | Heart Muscle | | |
| **Central Nervous System** | Brain | - | |
| | Cerebellum | | |
| | Cerebral cortex | | |
| | Hippocampus | | |
| | Lateral ventricle | | |
| **Digestive Tract** | Appendix | | |
| | Colon | | |
| | Duodenum | | - |
| | Oesophagus | | - |
| | Oral Mucosa | | - |
| | Rectum | | - |
| | Salivary gland | | - |
| | Small intestine | | - |
| | Stomach | | - |
| **Endocrine Glands** | Adrenal gland | | |
| | Parathyroid gland | | - |
| | Thyroid gland | | - |

| | | | |
|---|---|---|---|
| **Female Reproductive System** | Breast | | |
| | Cervix | | |
| | Fallopian tube | | |
| | Ovary | | - |
| | Placenta | | |
| | Uterus | | |
| | Vagina | | |
| **Liver and Pancreas** | Gallbladder | | |
| | Liver | | - |
| | Pancreas | | |
| **Male Reproductive System** | Epididymis | | |
| | Prostate | | |
| | Seminal vesicle | | |
| | Testes | | |
| **Respiratory System** | Bronchus | | |
| | Lung | | |
| | Nasopharynx | | - |
| **Skin and Soft Tissue** | Adipose tissue | | |
| | Skeletal muscle | | - |
| | Skin | | |
| | Smooth muscle | | |
| **Urinary Tract** | Bladder | | - |
| | Kidney | | |

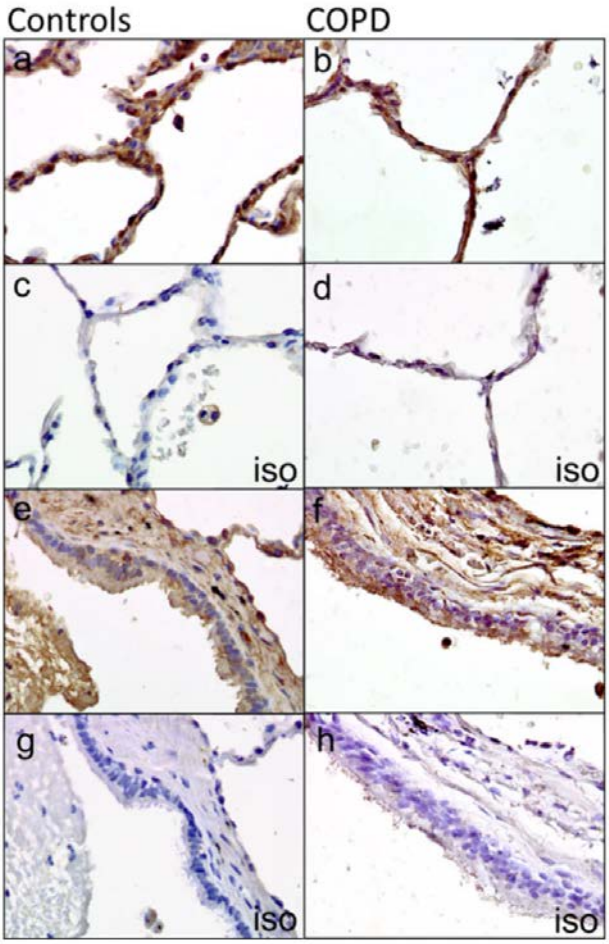



**Figure 6. GSTCD protein expression in human adult control and COPD lung tissue (Obeidat et al. 2013).** Representative images of GSTCD immunohistochemistry staining are shown from three healthy control donors (a and e) with matched isotype controls (c and g) and in three donors with COPD (b and f) with matched isotype controls (d and h). No clear differences were evident in gene expression between control healthy lung tissue and lung tissue from patients with COPD.

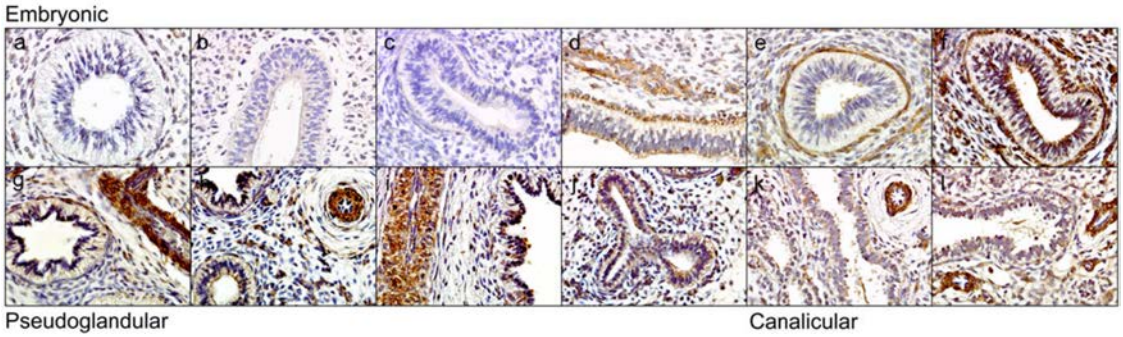



**Figure 7. GSTCD protein expression in human foetal lung tissue (Obeidat et al. 2013).** GSTCD and expression was investigated using immunohistochemistry across developmental stages in human fetal lung tissue: embryonic (19 days, images a and b; 21 days, images c and d; 23 days, images e and f), pseudoglandular (10 weeks, images g and h; 12 weeks, images I and j), canalicular (17 weeks, image k; 19 weeks, image l). Stronger staining is evident around the pseudoglandular stage corresponding to increased levels of GSTCD expression.

**Table 3. Fetal lung gene array data comparing GSTCD and INTS12 expression during pseudoglandular and canalicular stages of lung development (Obeidat et al. 2013).** Key: average expression is between all samples; p-value: unadjusted p-value; adjusted p-value controls for false discovery rate; beta-coefficient: log-odds ratio. n/s: no significance observed. Significant effects are shown in bold and italicised**.**

| Gene Entrez ID | Probe ID | Average Expression | *p*-value | Adjusted p-value | Beta-coefficient | Significant Effect |
|---|---|---|---|---|---|---|
| GSTCD | 220063_at | 5.6144 | 0.0265 | 0.0862 | -0.0037 | n/s |
| ***GSTCD*** | 1554518_at | 5.8000 | 0.0026 | ***0.0150*** | -0.0055 | ***Decreased expression with age*** |
| NA | 241126_at | 3.4619 | 0.3120 | 0.4878 | 0.0012 | n/s |
| ***GSTCD*** | 235387_at | 6.3236 | 0.0001 | ***0.0009*** | -0.0115 | ***Decreased expression with age*** |

## 1.8 Human GSTCD Knock-down *in vitro* Model

Since the function of GSTCD remains unclear, methods investigating the consequences of loss of this protein are beginning to be explored, such as by reducing GSTCD expression in a human *in vitro* cell-based model. The traditional primary approach for applying this method is by short interfering RNA (siRNA) knock-down. Although more recently an alternative method has begun to increase in popularity (CRISPR/Cas system), siRNA knock-down was used in this research as this has been widely utilised as a method for investigating functional biological pathways (Huh et al. 2013; Christofferson et al. 2012; Jin et al. 2011; Huang et al. 2009; Khalil et al. 2008).

Transiently transfecting nucleic acid sequences into cells typically results in expression of introduced genes, although the introduced sequences are not incorporated into the genome and are eventually lost after repeated cell division. In contrast, it was identified that siRNA sequences can be transiently transfected into cells to reduce (knock-down) the expression level of a particular gene (Fire et al. 1998). siRNA knock-down exerts post-transcriptional control on gene expression by directing target mRNA degradation via the RNA-induced silencing complex (RISC) (Fire et al. 1998). siRNA has emerged as a technique to knock-down expression of specific target genes in a variety of organisms hence provides a powerful method for studying gene function (Levenkova et al. 2004).

After transient transfection into a cell, double stranded siRNA sequences remain in the cytoplasm and associate with RISC (Figure 8). At this point the siRNA strands separate with the antisense strand remaining associated with RISC. This acts as a guide to bind to complementary mRNA sequences, resulting in initiation of degradation of the target mRNA by RISC (Fire et al. 1998). Although this method does not completely abolish gene expression, translation of the specific mRNA transcripts into proteins is severely limited.

This approach is useful because it allows targeted reduction of gene expression to easily model minimal protein levels for investigating function in human cells. However, this is an *in vitro* technique, often studying a single cell type in isolation; therefore it is difficult to understand the overall consequences of loss of

function in this model. Instead, a model organism gene knock-out may be generated to study overall phenotypic effects.

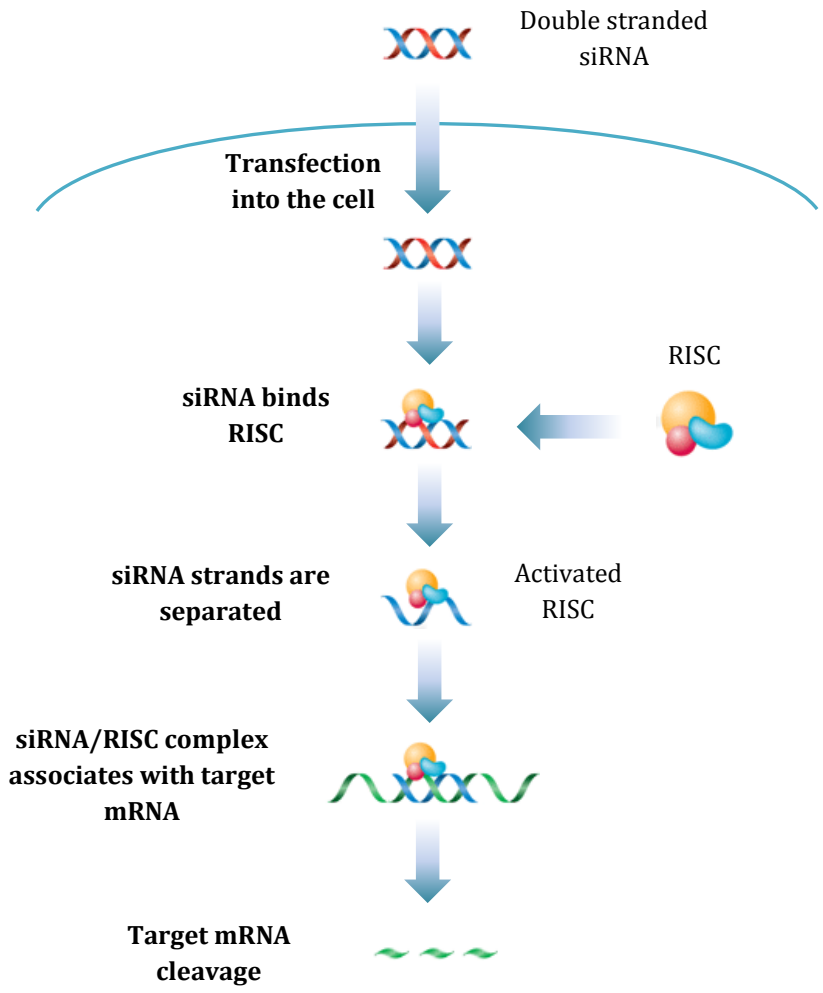



**Figure 8. siRNA mediated knock-down of gene expression (adapted from Santa Cruz Biotechnology RNAi 2015).** Double stranded siRNA is transfected into the cell where it associates with the multi-subunit protein complex, RNA-induced silencing complex (RISC), in the cytoplasm. The antisense strand is separated and used as a guide in RISC to target complementary mRNA sequences (activated RISC). These mRNA transcripts are degraded by the RNase Argonaute 2; the catalytic centre of RISC.

## 1.9 Mice as Model Organisms

Although there is continuing progress in research techniques studying human disease, animal model systems are invaluable tools for investigation as a controlled *in vivo* method (Dean et al. 2013). Many aspects of mammalian development, physiology and metabolism are conserved between humans and mice, making this species an ideal model system for research (Bradley et al. 2012). Specific genes can be targeted and manipulated in a model organism, a widely used method allowing validation and functional characterisation of unknown genes (Grant et al. 1992; Tsien et al. 1996; Caruana & Bernstein 2001; Dahl et al. 2003; Cuthbert et al. 2007; Penzkofer et al. 2014) such as those previously identified in humans from a variety of '-omic' studies (Dean et al. 2013). To determine unknown gene function it is often useful to disrupt gene expression by mutating individual genes and studying effects of this alteration in mice to model effects relevant to human disease. This method allows more in-depth understanding for the implications of disrupting an individual gene than *in vitro* analysis of a single human cell type. The primary reasons for this are that all of the cell types in the organism are present in their normal spatial relationships, with normal intercellular communication and cellular interactions occurring.

Alternatively to studying GSTCD siRNA knock-down in human cells, which does not completely abolish protein production, loss of function effects can be more thoroughly explored through absence of the protein in a knock-out mouse model. When choosing model organisms many things are taken into account and mice are often the organism of choice for research because: ~80% of human genes correspond to a single orthologous gene in the mouse (Jackson-Informatics 2014), there are wide ranging techniques available for mammalian genetic modifications and detailed information is collated in mouse genotype and phenotype databases.

A notable example of a mouse resource is the International Knock-out Mouse Consortium (IKMC), set up with the aim of creating a standardised resource of mice with knock-out mutations for all known protein-coding genes, therefore permitting study of the consequences of complete loss of function for a particular

gene product. The IKMC was then followed on by the International Mouse Phenotyping Consortium (IMPC), establishing an international consortium of mouse centres. These centres have the aim of characterising the phenotypes associated with each of the knock-out mice from the IKMC and identifying any alterations possibly caused by the absence of the knocked-out gene.

Importantly, gene function is suggested to be strongly conserved between humans and mice as these species have a high level of gene and protein similarity. This is evident in direct alignment of human and mouse *GSTCD* DNA and protein sequences which demonstrate a high level of homology with both scores at 84% (NCBI HomoloGene 2014) (further demonstrated in Results section 4.1.5).

Therefore, comparisons between the two species showing high levels of homology, similar arrangement of gene domains and significant amounts of well characterised data available from the IMPC demonstrate that the mouse is a highly applicable model for the analysis of Gstcd knock-out effects.

### 1.9.1 Gstcd Knock-out Generation Methodology

As little is currently known about the lung function associated gene *GSTCD*, knock-out mice were utilised in this research to discover the phenotypic effects caused by loss of Gstcd protein, carried out by investigating knock-out mice generated via the IMPC.

Since 2007 the IKMC has been running a large scale gene targeting pipeline to generate gene knock-out mutations in mouse embryonic stem cells (ESC) (Collins et al. 2007). Gene targeting is the process whereby DNA is inserted into a genome by recombination at a site where expression can be adequately controlled. This technique allows high levels of accuracy for allele design and the potential for conditional knock-out mutagenesis; where gene function can be studied by creating the knock-out in a specific tissue of interest (spatial control) or controlling the knock-out at a particular lifecycle stage (temporal control) (Skarnes et al. 2011).

In this investigation a Gstcd knock-out mouse was established in collaboration with Cornell University which generated the mice using an ESC line created within the IMPC (member: North American Conditional Mouse Mutagenesis project (NorCOMM) (Collins et al. 2007; Nagy et al. 2009)). The colony has been successfully expanded and preliminary analysis performed to confirm loss of Gstcd. Since this time Gstcd has also been selected to go through the IMPC phenotype analysis program, the protocols for which have been standardised within the International Mouse Phenotyping Resource of Standardised Screens (IMPReSS) in association with the IMPC (Figure 9).

Finally, with GSTCD potentially having a role in lung development based on significant changes in mRNA and protein levels at key developmental stages (Obeidat et al. 2013) it is also necessary to understand normal mammalian lung development. Knowledge of normal lung development allows us to draw conclusions for understanding the full effect of knocking out genes and their implications in potentially altered phenotypes.

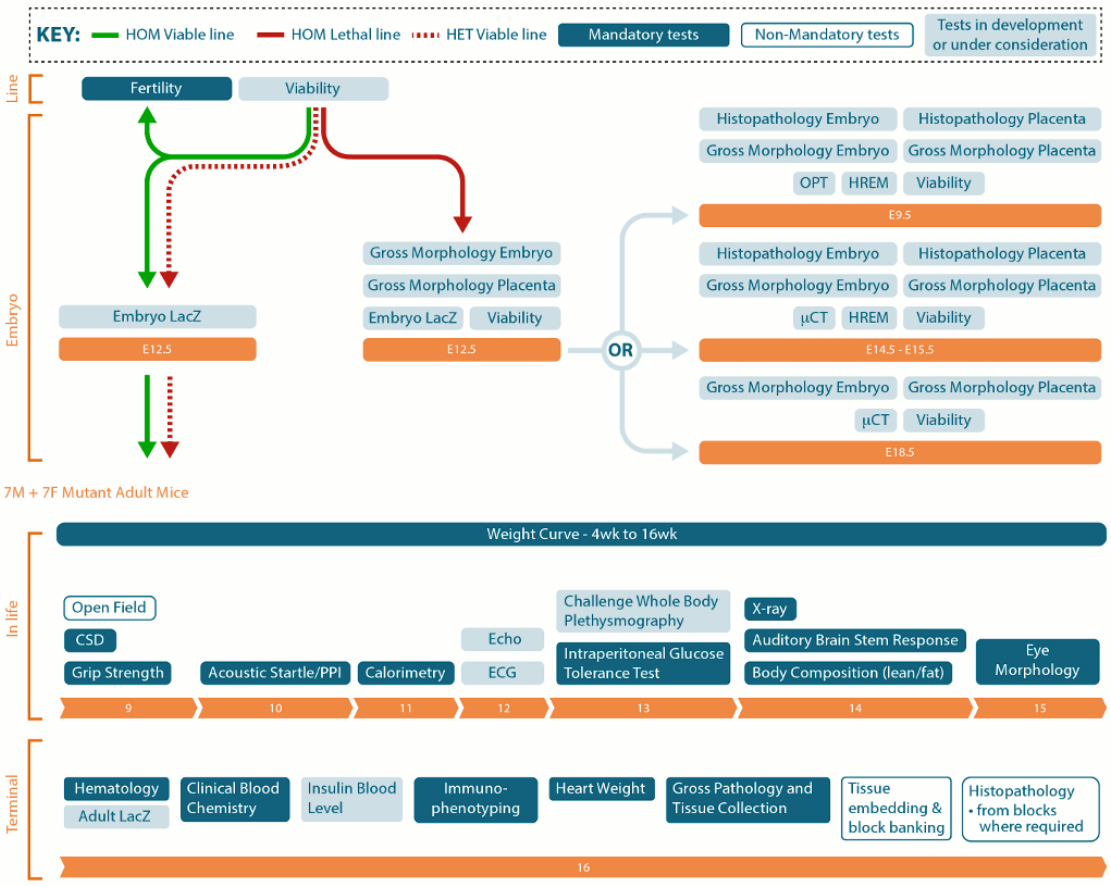



**Figure 9. International Mouse Phenotyping Resource of Standardised Screens (IMPReSS) phenotyping pipeline for knock-out mice within the IMPC (IMPReSS 2015).** The pipeline demonstrates the standard pathways of analysis at several age ranges for knock-out mice assessed within the IMPC using tests for a variety of phenotypes. These depend on the type of knock-out mouse available; homozygous gene knock-out viable (Hom viable line), homozygous gene knock-out lethal (Hom lethal line) or heterozygous gene knock-out viable (Het viable line).

## 1.10 Mammalian Lung Development

Based upon previous evidence demonstrating varied GSTCD expression across lung development and association of GSTCD with $FEV_1$ (Obeidat et al. 2013; Repapi et al. 2010; Hancock et al. 2010), it is relevant to investigate the role of GSTCD in lung development. Firstly, it is important to understand normal lung development prior to investigation of potential differences in a loss of function model.

Initially, the mammalian respiratory system originates from the anterior foregut endoderm in the foetus for the purpose of developing an ideal structure to facilitate gas exchange. Separation of the developing trachea from the oesophagus is the primary step in the formation of the lungs at approximately embryonic day E9 in the mouse embryo (Herriges & Morrisey 2014). By E9.5 (~28 days in humans) the primary lung buds have formed marking the beginning of the embryonic stage of development (E9.5-E12.5) (Table 4, Figure 10) (Herriges & Morrisey 2014; Morrisey & Hogan 2010).

The developmental stages outlined in Table 4 can be further categorised into two overall components of lung development: early (branching morphogenesis) or late (alveolar differentiation) stages and co-ordination between the two processes has only recently been investigated (Chang et al. 2013). Chang *et al.* theorised that branching morphogenesis and alveolar differentiation are co-ordinated processes during lung development; to enable generation of the ideal structure whereby alveoli are located at branch termini (Chang et al. 2013).

During embryogenesis and the following pseudoglandular stage the two lung buds begin a complex process of branching morphogenesis; a highly regulated process generating a tree-like structure of epithelial tubes branching by dichotomy (Herriges & Morrisey 2014). This is a stereotyped and intricate process where virtually identical structures can be seen within inbred mouse strains (Metzger et al. 2008).

The second main component of lung development is alveolar (cellular) differentiation; involving generating the functional units of the lungs by

specialised epithelial cells forming alveoli (Table 4, Saccular and Alveolar stages). There are two characteristic types of alveolar cell formed during lung development: type I and type II alveolar epithelial cells. Alveolar epithelial type I cells make up the majority of alveolar area, approximately 95%, forming a thin diffusion barrier for gas exchange but can be easily damaged (Fehrenbach 2001; McElroy & Kasper 2004). Alternatively type II cells are available as a continued source of both cell types and have the essential role of secreting surfactant (Fehrenbach 2001). This surfactant decreases the surface tension between the alveolar walls, allowing alveolar cells to adapt to the mechanical stress during inhalation and exhalation. In the later stages of gestation, further differentiation of cell types becomes more evident, each with specialised functions, for example: secretory, neuroendocrine, basal and ciliated cells. A notable difference between mice and humans however, is the difference in timing of this stage; only in mice does this stage occur entirely postnatally (Warburton et al. 2010). Overall, throughout lung development and adulthood cell growth and proliferation play a critical role in cellular differentiation and homeostasis in the lungs.

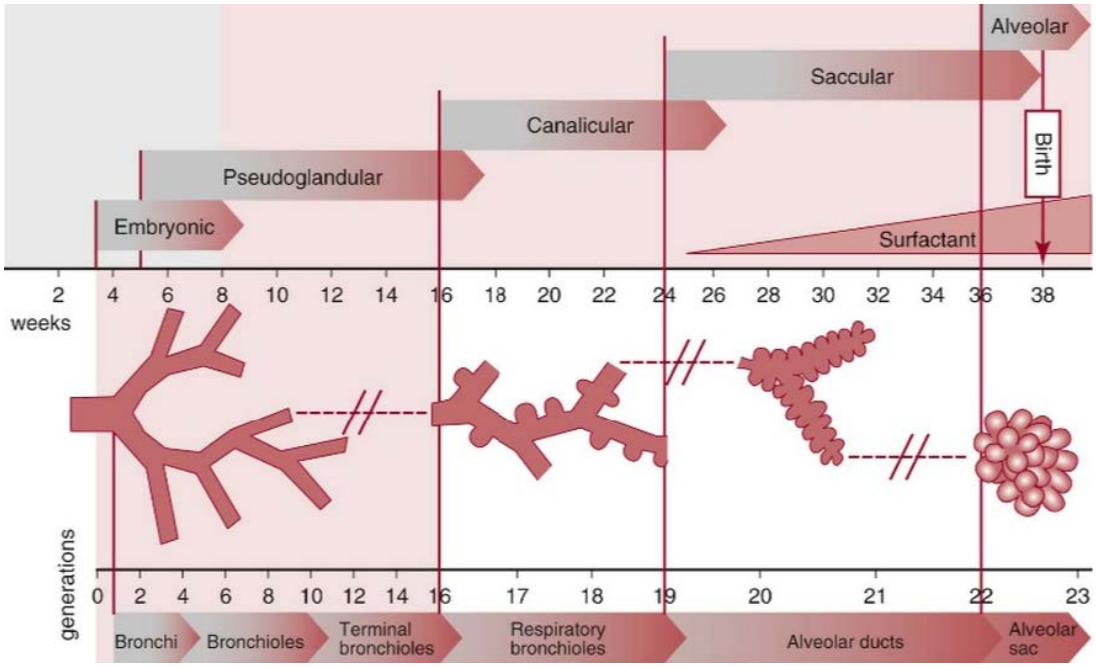



**Figure 10. Stages of lung development in humans (Bancalari & Polin 2012).** Demonstrating the timing (in weeks) of the different lung development stages throughout human gestation (Embryonic, Pseudoglandular, Canalicular, Saccular and Alveolar) and the key structural features formed in each stage (measured as a number of airway branching divisions; generations).

**Table 4. Stages and events during lung development in humans and mice (Bancalari & Polin 2012; Koopman 2010; Warburton et al. 2010).**

| Developmental Stage | Human Gestation Age | Mouse Gestation Age | Events |
|---|---|---|---|
| **Embryonic** | 4-7 weeks | 9-14 days | ▪ Septation of trachea from oesophagus<br>▪ Lung bud forms two main bronchi and individual lobes |
| **Pseudoglandular** | 5-17 weeks | 14-16.5 days | ▪ Branching morphogenesis begins<br>▪ Proximal airway epithelial and mesenchymal differentiation |
| **Canalicular** | 16-26 weeks | 16.5-17.5 days | ▪ Further branching<br>▪ Vascularisation and angiogenesis occurs along the airway<br>▪ Rapid increase in capillary numbers<br>▪ Respiratory bronchioles and alveolar ducts form<br>▪ Alveolar epithelial type II cells differentiate |
| **Saccular** | 24-36 weeks | 17.5 days to postnatal day 5 | ▪ Alveolar epithelial type I cells differentiate<br>▪ Alveolar epithelial type II cell maturation<br>▪ Air sacs are developed<br>▪ Lymphatic network develops<br>▪ Surfactant begins to be produced<br>▪ Initial stage of primary septation for formation of alveoli occurs |
| **Alveolar** | 36 weeks to childhood | Postnatal day 5-30 | ▪ Secondary septation produces alveoli, increasing gas exchange surface area<br>▪ Increasing levels of surfactant produced<br>▪ Majority of gas exchange surface is formed |

## 1.11 Cell Growth in the Lungs

Cell growth and proliferation are essential requirements for cell expansion and airway branching during lung morphogenesis as well as cellular differentiation (Nogawa et al. 1998). Cell growth remains crucial throughout life for lung homeostasis, for instance in the form of constant regeneration and often repair. Critically, alterations to cell growth can contribute to COPD development through airway remodelling (Jeffery 2001; Demedts et al. 2006; Lapperre et al. 2007). As the role of GSTCD in human cells is currently unconfirmed, an initial investigation exploring the effects on cell growth in a loss of function model was selected, due to evidence of a potential protective role of GSTCD in the lungs.

There are a variety of assays dedicated to assessing and quantifying cell growth, several key assays are outlined below.

### 1.11.1 Cell Growth Assays

Changes in cell growth can be assessed and quantified in a variety of ways; an initial measure was cell counting using Trypan blue. Living cells are stained blue enabling quantification of living cells, dead cells and a total cell count, in general with total cell count as the overall measure. The MTT assay became the next standardly used cell growth assay, utilising mitochondrial dehydrogenases. These enzymes are found in viable cells and cleave the tetrazolium ring of MTT, therefore changing the solution colour, which can be measured spectrophotometrically. The basis of this technique is that an increase in cell numbers results in an increase in the amount of MTT formazan formed producing an increase in absorbance.

Whilst the MTT assay has been a traditional measure of cell growth there is doubt about the accuracy and sensitivity of this method, as it only detects living cells and primarily detects cells with high metabolism. Therefore, several more recently developed assays with varied approaches to measuring cell growth are replacing this method.

Similar assays to MTT have been developed which claim to be more sensitive and with more simple protocols. In particular the CCK-8 assay also utilises a highly water-soluble tetrazolium salt (WST-8) which produces an orange-colour formazan dye upon reduction by dehydrogenase activities in cells. This dye is soluble in the tissue culture media and, similarly to MTT, the amount of dye generated is directly proportional to the number of living cells. The detection sensitivity of CCK-8 is claimed to be higher than other tetrazolium salts such as MTT, MTS, WST-1 or XTT.

Importantly, dehydrogenase based assays measure metabolic activity not necessarily total cell growth; hence it is an indirect measure of cell count. Also as an absorbance based assay it is not considered particularly sensitive at detecting minor changes in cell growth. Therefore, fluorescence based assays are becoming more widely used to investigate different measures of cell growth.

For instance, the Click-iT® EdU microplate assay labels actively proliferating cells by incorporating the thymidine analogue EdU into DNA during cell culture, advancing the initial DNA labelling methods involving incorporation of the radioactive nucleoside tritiated-thymidine analogue (Leigh et al. 1995; Reznik et al. 2000; Johnson et al. 2001) and later the BrDU thymidine analogue (Nogawa et al. 1998; Rose et al. 2002; Dasgupta et al. 2009; Herrera et al. 2013). In this instance the nucleoside analogue EdU (5-ethynyl-2´-deoxyuridine) is incorporated into DNA during active DNA synthesis and allows direct measurement of DNA synthesis, followed by antibody based signal amplification. Advantages of this method include: high sensitivity, no dangerous radioactive material is required (tritiated-thymidine) and DNA denaturation is not required (integrity of DNA is not disrupted as in the similar method utilising BrDU) (Zeng et al. 2010; Chehrehasa et al. 2009; Yu et al. 2009).

A second fluorescence based assay that is becoming more commonly used is CyQUANT® NF proliferation assay, which utilises a cell permeable fluorescent DNA binding dye applied to cell cultures at the experimental end-point (Jones et al. 2001). Due to cellular DNA content being highly regulated, it is closely

proportional to cell number, hence providing an accurate representation of potential variations in cell growth between experimental conditions.

Measure of cell growth in a human in vitro model was selected as an initial investigation of reduction of GSTCD effects, although the function of this protein remains speculative. Protein function can be inferred by identifying homologous protein sequences as these homologous sequences often result in similar secondary protein structures; generally indicating related functional activity.

## 1.12 Inferring Protein Function

Alternatively to investigating GSTCD loss of function in human and mouse models, protein function can be inferred based on sequence similarity (homology) to proteins of known function.

Full understanding of protein function at the molecular level commonly relies on knowledge of the 3D protein structure. In particular, domains consisting of protein sequence homology often indicate a shared or similar function. We are able to exploit this based on knowledge that 3D structure of proteins from the same family is more conserved than the amino acid sequences (Fiser 2010). Therefore, if sequence similarity is detected between two proteins, then structural similarity can generally be assumed (Fiser 2010).

Methodology for protein structure inference is increasingly based around public protein sequence databases, primarily due to high-throughput structural genomics initiatives producing an ever increasing number of protein sequences. Therefore, this allows prediction of 3D structure by automated protein structure prediction methods; utilising proteins of existing known structure and function to infer information about proteins of unknown function, i.e. template based methods (Yang et al. 2011). Template based modelling is becoming increasingly powerful as the majority of the most common domains are known (Dai & Zhou 2011; Kihara & Skolnick 2003; Zhang et al. 2006; Yang et al. 2011).

Computer algorithms predicting protein structure using template modelling provide valuable insight into relatively unknown proteins of interest such as GSTCD. Protein prediction servers are assessed and ranked by critical assessment of protein structure prediction (CASP) (Kopp et al. 2007; Cozzetto et al. 2009; Mariani et al. 2011; Kryshtafovych et al. 2014). CASP is a worldwide experiment which objectively tests structure prediction methods. Rankings by CASP are considered the key metric of methods and program confidence in protein structure prediction (Quashie et al. 2015).

I-TASSER, SWISS-MODEL and Phyre2 servers were selected to further investigate GSTCD structure and function because they are ranked amongst the best protein

prediction servers as assessed by CASP, are freely available online, automated and user-friendly to non-bioinformaticians. Moreover, multiple servers were utilised as this allows comparison of results and permits conclusions to be drawn with higher confidence if similar results are obtained from several analysis methods, as different bioinformatics approaches are used across the servers. Each server identifies regions of sequence homology with proteins of known structure and function and infers information about the protein of interest such as: secondary structure, theoretical tertiary structure and intrinsic disorder (degree to which a protein lacks a fixed 3D structure).

# 2 Hypothesis and Aims

Genetic polymorphisms spanning the *INTS12-GSTCD-NPNT* (4q24) locus have been reproducibly associated with lung function ($FEV_1$) and COPD diagnosis in many recent studies (Repapi et al. 2010; Hancock et al. 2010; Castaldi et al. 2011; Soler Artigas, Wain, et al. 2011; Soler Artigas, Loth, et al. 2011; Imboden et al. 2012; Hancock et al. 2012; Li et al. 2013; Loth et al. 2014; Yao et al. 2014; Wain et al. 2015; Artigas et al. 2015; Brehm et al. 2015). There has been particular focus on the gene *GSTCD* at this locus due to SNPs in this gene indicating the most significant p-values in several studies. Initial analyses including the expression profile of GSTCD in fetal and adult lung tissue (Obeidat et al. 2013) along with the predicted GST activity for GSTCD make it a plausible candidate gene in this region.

Based on previous research it was hypothesised that GSTCD may contribute to normal mammalian lung development and may affect predisposition to altered lung function in adults, potentially leading to COPD. Therefore I set out to further investigate the potential reasons behind the significant associations of *GSTCD* with lung function and COPD.

The aims of this project were:

1. To investigate whether GSTCD has a role in airway cell growth, using a siRNA knock-down approach in human bronchial epithelial cells
2. To assess lung morphology in Gstcd knock-out and wild type mice to assess the role of GSTCD in lung cell growth
3. To infer the potential function of GSTCD by protein prediction methods based on protein sequence homology

# 3 Materials and Methods

# Cell Biology

## 3.1 Human Primary Cell Culture

Non-differentiated human bronchial epithelial cells (hBECs), isolated from the epithelial lining of the airways, were obtained commercially (Lonza) and grown in optimal bronchial epithelial cell growth medium (BEGM) (Lonza, CC-3170) as described previously (Danahay et al. 2002). Cells were incubated at 37°C in a humidified atmosphere with 5% $CO_2$. Minimal growth media used which was sufficient for cell growth whilst limiting confounding factors (such as the large number of supplements normally added) consisted of bronchial epithelial basal growth medium (BEBM) (basal constituent of BEGM) with insulin, transferrin and sodium selenite (ITS) (Sigma, I3146) as the only supplemented nutrients. ITS supplement was deemed necessary for minimal cell growth because purified forms of the factors normally supplied by serum are required in serum-free media (Hardyman et al. 2013).

hBECs at passage 3 from two donors were used in these experiments. Cells were seeded at densities of 100,000 cells/well in a 6-well plate (Scientific Laboratory Supplies, 3506) or 4,000 cells/well in a 96-well plate (Sigma, CLS3603) initially grown in optimal media and experiments performed when cells reached 30-40% confluency.

## 3.2 Transient Transfection

Short interfering RNA (siRNA) sequences can be transiently transfected into cells to reduce the expression level of a particular gene. After transient transfection siRNA sequences remain in the cytoplasm and bind to complementary mRNA sequences. This results in initiation of degradation via the RNA-induced silencing complex (RISC) due to the formation of double stranded RNA complexes.

hBECs were transiently transfected with siRNA to knock-down GSTCD expression. This was used as a method for investigating changes in functional biological pathways caused by a reduction of target gene expression (Huh et al. 2013; Christofferson et al. 2012; Jin et al. 2011; Huang et al. 2009; Khalil et al. 2008). Three pre-designed siRNA sequences targeted to GSTCD and a scrambled siRNA control were obtained from Cambridge Bioscience (SR312613). The scrambled siRNA acts as a negative control as this contains the same nucleotide composition in a different sequence to the gene targeting siRNAs and does not target a known gene. Sequence alignments to human GSTCD transcripts are outlined in Appendix section 6.2.3.

GSTCD siRNA knock-down sequences:

- Scrambled control siRNA: CGUUAAUCGCGUAUAAUACGCGUAT
- siRNA A: AGAAUGUGGAUACUCCGUUCAAGTG (exon 12)
- siRNA B: GGAUGUUUAAUAUUGGAGUAGCATT (exon 8/9)
- siRNA C: GCCUGUUAGAGUGCAUAAUGAUGAT (exon 3)

1nM of siRNA was transfected into hBECs using INTERFERin® transfection reagent (Polyplus Transfection, 409-10) based on previous optimisation by Amanda Henry for maximal GSTCD knock-down. During optimisation cells were also incubated with siRNA reagents for varied time-courses (12h, 24h, 48h, 72h) (Amanda Henry), identifying 24h and above as the most reliable time-points for GSTCD knock-down.

## 3.3 Proliferation Assays

Several methods of measuring cell proliferation were utilised including: cell counting, MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) assay (Sigma, M5655), Cell Counting kit 8 (CCK-8) (Dojindo, CK04), Click-iT® EdU microplate assay (Life Technologies, C10214) and CyQUANT® NF proliferation assay (Life Technologies C35007).

The ability to quantify differences in cell growth and proliferation reliably and sensitively in hBECs when cells are grown in minimal and optimal media over 24h was used to compare these assays. Cell counting was performed on cells grown in 6-well plates (2ml/well media) and 96-well plates (200µl/well media, black walled) were used for remaining assays (all assays used 4 replicates for each experimental condition) using quantification measures as described above.

Absorbance and fluorescence intensity measures for MTT, CCK-8, Click-iT® EdU and CyQUANT® NF assays were read on a FlexStation® 3 microplate reader (Molecular Devices). Manufacturer's protocols were followed and are outlined below.

### 3.3.1 Cell Counting

Cell counting using Trypan blue was used as a typical measure of total cell number, by counting cells under a light microscope using a haemocytometer and identifying dead cells as those which take up the dye and are stained blue.

Media was aspirated from hBEC cultures followed by 5 minute incubation with 300µl/well 0.5mg/ml trypsin (Sigma, T3924) at room temperature. 300µl/well 0.5mg/ml trypsin inhibitor (Life Technologies, 17075-029 was added and cells carefully scraped from the well using a cell spatula (TPP, 99010). The cell suspension was mixed with 300µl appropriate media and 50µl diluted 1:2 with Trypan Blue solution (Sigma, T8154), this mixture was incubated at room temperature for 5 minutes after which a sample of the cells was added to a haemocytometer and counted under a light microscope. Living and dead cells

were recorded as either unstained or blue stained respectively and total cell counts inferred following the calculations below.

((average count per square x 2) x $10^4$) = Cells/ml

Cells/ml x the original volume of media the cells were taken from (ml) = Total cells

(Total number of cells unstained / Total number of cells counted (stained and unstained)) x 100 = % viability

### 3.3.2 MTT assay

A second commonly used method of measuring proliferation is the MTT assay. The MTT assay utilises mitochondrial dehydrogenases found in viable cells which cleave the tetrazolium ring of MTT, resulting in a colour change in the solution which can be measured by absorbance readings.

A sterile 0.5mg/ml solution of MTT (3-(4,5-dimethylthiazol-2-yl)-2,5 diphenyl tetrazolium bromide) was freshly prepared prior to each experiment in appropriate media (minimal or optimal media) and warmed to 37°C. For experimental wells media was replaced with 200µl appropriate MTT solution for each condition whilst regular minimal or optimal media was added to control wells. Cells were then returned to the incubator (under optimal cell conditions described in 3.1 Human Primary Cell Culture) for 4 hours at which point media was removed from all wells and replaced with 200µl isopropanol to solubilise formazan crystals. Absorbance was measured at wavelength 570nm ($OD_{570}$).

### 3.3.3 Cell Counting kit 8 assay (CCK-8)

Similarly to the MTT assay the CCK-8 assay also exploits cellular dehydrogenase activity to measure cell proliferation using an alternative tetrazolium salt (2-(2-methoxy-4-nitrophenyl)-3-(4-nitrophenyl)-5-(2,4-disulfophenyl)-2H-tetrazolium monosodium salt) (WST-8)).

CCK-8 assay solution was added directly to cells (20µl/well) with no media change required and appropriate media (BEBM-ITS or BEGM) added to matched controls, followed by a 4 hour incubation in optimal cell culture conditions (3.1 Human Primary Cell Culture). Absorbance was measured at wavelength 450nm ($OD_{450}$).

### 3.3.4 Click-iT® EdU Microplate Assay Protocol

Click-iT® EdU microplate assay labels actively proliferating cells by incorporating the thymidine analogue 5-ethynyl-2′-deoxyuridine (EdU) into DNA during cell replication, followed by antibody based signal amplification resulting in a fluorescence reading (Cappella et al. 2008; Yu et al. 2009; Chehrehasa et al. 2009). The assay kit provides components for DNA labelling of cells (EdU) and the antibody based signal amplification utilises antibody specificity for binding to EdU labels. This antibody is also conjugated to horse radish peroxidase (HRP) which then reacts with Amplex® UltraRed substrate; producing a fluorescent product.

The DNA labelling component, EdU, was added directly to experimental wells only in media at a final concentration of 10µM, 24h prior to each experiment end point (with matched volume of appropriate media added to control wells). After 24h, media was removed and cells fixed by incubation with 50µl/well Click-iT® EdU Fixative for 5 minutes. Cells were then labelled by addition of 50µl/well click labelling reaction cocktail (Table 5) applied directly into the fixative solution; plates were protected from light from this point forwards. The reaction cocktail mix was then aspirated from the wells and subsequently cells were washed twice with 200µl/well blocking buffer for 5 minutes. After removal of the blocking buffer, cells were incubated with 50µl/well antibody reaction for 30 minutes, followed by two washes with 200µl/well Amplex® UltraRed buffer. Amplex® UltraRed buffer was removed and cells incubated with 100µl/well Amplex® UltraRed reaction mixture (Table 6). To stop the reaction, 10µl/well Amplex® UltraRed stop reagent was added to the Amplex® UltraRed reaction mixture and fluorescence was measured at 568/585nm (excitation/emission).

**Table 5. Click-iT® EdU microplate assay kit manufacturer guidance for producing 2X Click-iT® reaction cocktail components (volumes sufficient for one 96-well plate).**

| Reaction Components | 1 plate |
|---|---|
| 2X Click-iT® reaction buffer | 4.7ml |
| Copper Sulphate | 120µl |
| Oregon Green® 488 azide | 30µl |
| Click-iT® EdU buffer additive | 1.2ml |
| **Total volume** | **6ml** |

**Table 6. Click-iT® EdU microplate assay kit manufacturer guidance for producing Amplex® UltraRed reaction mixture (volumes sufficient for one 96-well plate).**

| Reaction Components | 1 plate |
|---|---|
| Amplex® UltraRed buffer | 12ml |
| Amplex® UltraRed reagent | 12µl |
| Hydrogen peroxide | 12µl |
| **Total volume** | **~12ml** |

### 3.3.5 CyQUANT® NF Assay Protocol

CyQUANT® NF proliferation assay measures total cell number utilising a cell permeable fluorescent DNA binding dye applied to cell cultures at the experimental end-point (Jones et al. 2001). Since cellular DNA content is highly regulated, it is closely proportional to cell number.

Media was aspirated from cells and replaced by 100µl/well of CyQUANT® dye binding solution (CyQUANT® NF dye reagent diluted 1:500 in supplied Hank's balanced salt solution (HBSS) buffer) to experimental wells or HBSS to control wells. Cells were then protected from light and incubated at 37°C for 30 minutes before fluorescence was measured at 580/520nm (excitation/emission).

# Molecular Biology

## 3.4 RNA and Protein Extraction

RNA and protein were extracted from hBECs to measure levels of GSTCD expression in siRNA knock-down samples relative to controls and to confirm whether the observed reduction of GSTCD RNA levels is also reflected at the protein level.

### 3.4.1 Isolating RNA Lysate

Total RNA can be isolated from mammalian cell cultures by lysing cells and binding RNA to a column to purify. Initially, RNA was obtained from hBECs cultured in 6-well plates using lysis solution from GenElute™ Mammalian Total RNA Miniprep Kit (Sigma, RTN70). This was followed by RNA extraction using the manufacturer's protocol as follows: media was aspirated from cells and wells washed twice using 2ml/well cold Dulbecco's Phosphate Buffered Saline (PBS) (Sigma, D8537). 300µl/well of RNA lysis solution (Genelute™ Mammalian Total RNA Miniprep Kit: 1ml lysis solution plus 10µl β-Mercaptoethanol) was added and cells removed from culture plastic using a cell spatula (TPP, 99010). Duplicate wells were used for each experimental condition to ensure sufficient RNA concentrations were obtained, hence two wells containing 300µl lysed cells were combined and RNA lysates stored at -80°C.

### 3.4.2 RNA Extraction from Lysate

Briefly, for each sample this involved filtering 600µl RNA lysate through a filtration column and centrifuging for 2 minutes at maximum speed. An equal volume of 70% ethanol was added to the flow-through in the collection tube and the solution thoroughly mixed by pipetting. Up to 700µl of lysate/ethanol mix was applied to a binding column and centrifuged for 15 seconds at maximum

speed followed by removal of the flow-through solution (subsequent centrifugation steps repeat these conditions unless otherwise stated). This step was then repeated for the remaining lysate/ethanol mix. The binding column was washed using 500µl Wash Solution 1 and centrifuged.

On-column DNase treatment (Sigma, DNASE70) was performed to eliminate potential DNA contaminants by addition of pre-mixed 10µl DNase I and 70µl DNase digest buffer, incubated at room temperature for 15 minutes. After DNase treatment the binding column was washed with 250µl Wash Solution 1 and centrifuged. The binding column was then transferred to a new 2ml collection tube and washed with 500µl of Wash Solution 2 followed by centrifugation. A second column wash with Wash Solution 2 was carried out with a 2 minute centrifugation and an additional 1 minute centrifugation was used to dry the column. RNA was eluted by transferring the binding column into a new 2ml collection tube with 50µl Elution Solution applied directly onto the binding column and centrifuging for 1 minute. Quality control of RNA and concentration was measured using a Nanodrop™ 2000 spectrophotometer where RNA purity was assessed by the 260/280 ratio (sample values in the expected range of 1.9-2.1 indicating no significant protein contamination). RNA integrity was checked by 1.5% agarose gel electrophoresis (Fisher, 10264544).

### 3.4.3 Harvesting Protein Lysate

Media was aspirated from cells washed twice using 2ml/well cold Dulbecco's Phosphate Buffered Saline (PBS) (Sigma, D8537). To ensure sufficient protein concentrations were obtained each experimental condition was performed in duplicate wells. Protein was extracted from hBECs cultured in 6-well plates by adding 200µl protein lysis solution (Table 7) to the first well and using a cell spatula to detach cells from the culture plastic then transferring the 200µl lysis buffer to the duplicate well and repeating cell removal on ice. Samples were stored at -80°C until protein determination assay.

**Table 7. Reagents and volumes used for protein lysis solution for extracting protein from cells grown *in vitro*.**

| Reagent | Supplier | Catalogue Number | Volume/Quantity |
|---|---|---|---|
| Benzonase nuclease | Merck Biosciences | 70746-4 | 10µl |
| Pepstatin A | Sigma | P4265 | 3µl |
| DTT | Sigma | D9779 | 1µl |
| Lysis Buffer | - | - | 1ml |
| Lysis Buffer: | | | |
| cOmplete, Mini Protease Inhibitor Cocktail Tablets | Roche | 11 836 153 001 | 1 tablet |
| Cytobuster | Merck Biosciences | 71009-3 | 10ml |

### 3.4.4 BioRad Protein Determination Assay

The BioRad colorimetric assay determines a quantitative measure of protein concentration based on the Bradford dye-binding method. This procedure involves binding of Coomassie Brilliant Blue G-250 to proteins resulting in a change in the maximum absorbance level of the solution (from 465 to 595nm) which can be monitored (Bradford 1976). Results of this assay are obtained rapidly and are highly reproducible. Samples of known concentration are measured alongside unknown samples in a series dilution, allowing unknown sample absorbance values to be plotted on a standard curve for predicting protein concentration.

Protein lysates stored at -80°C were thawed on ice followed by a 30 minute incubation on ice to ensure complete lysis. Insoluble material was pelleted from lysates by centrifugation for 10 minutes at 4°C. Each protein sample was diluted 1:2, 1:5 and 1:10 in duplicate and a series dilution of the protein standard, bovine serum albumin (BSA) (Millipore, 82-045-2) established (2000µg/ml - 62.5µg/ml). Bio-Rad protein assay dye concentrate (Bio-Rad, 500-0006) was diluted 1:5 in distilled water and 500µl added to all samples. After a 15 minute

incubation 200µl of each sample was dispensed into a 96-well plate (Scientific Laboratory Supplies, 873429-1) and absorbance read at wavelength 595nm ($OD_{595nm}$) on a Flexstation® 3 microplate reader (Molecular Devices). Concentrations of samples were derived from the standard curve created by the series dilution measurements.

## 3.5 cDNA Synthesis

RNA can be converted into cDNA by reverse transcription to enable application of DNA molecular techniques, such as quantification of mRNA expression levels by quantitative polymerase chain reaction (Q-PCR).

RNA samples were first treated with DNase I (Sigma, 18068-015) to eliminate any traces of DNA by incubating 8µl of each sample (diluted to 1µg of RNA) with 1µl DNase I (1unit/µl) and 1µl 10X DNase I buffer at room temperature for 15 minutes. This reaction was stopped by the addition of 1µl of 25mM EDTA solution and incubation at 65°C for 10 minutes. Samples were then divided into two 5.5µl aliquots (500ng RNA) and labelled as reverse transcriptase (RT) positive or negative samples.

SuperScript™ II First Strand cDNA Synthesis kit (Invitrogen, 11904-018) was utilised to convert the RNA to cDNA. All RT positive and negative samples were initially incubated for 5 minutes at 65°C in a mixture of 1µl random hexamers and 1µl dNTPs with 3µl water. Additional components were added to the samples according to Table 8 which were incubated at 25°C for 2 minutes. Afterwards 1µl Superscript™ II RT was added to RT positives and 1µl water added to control RT negative samples, conditions in Table 9 were then followed. After amplification of cDNA, remaining RNA in all samples was digested by addition of 1µl RNase H and incubation at 37°C for 20 minutes. Samples were diluted 1:5 in DEPC water and stored at -20°C.

**Table 8. Reaction components for SuperScript™ II First Strand cDNA Synthesis kit after incubation of RNA/primer mix.**

| Reagent | X1 RT +ve (µl) | X1 RT –ve (µl) |
|---|---|---|
| 10X RT buffer | 2 | 2 |
| 25mM $MgCl_2$ | 4 | 4 |
| 0.1mM DTT | 2 | 2 |
| RNase OUT | 1 | 1 |
| **Total** | **9** | **9** |

**Table 9. Temperature and incubation length conditions for cDNA synthesis from RNA using SuperScript™ II reverse transcriptase.**

| Temperature (°C) | Time (minutes) |
|---|---|
| 25 | 10 |
| 42 | 50 |
| 70 | 15 |
| Place on ice | 2 |

## 3.7 PCR and Quantitative-PCR

### 3.7.1 Polymerase Chain Reaction

In order to study a gene of interest a large quantity of DNA is often required, therefore DNA is frequently amplified by polymerase chain reaction (PCR). This employs the use of cycles of heat denaturation, annealing of primer DNA sequences and extension from these sequences using *Taq* DNA polymerase to amplify DNA strands exponentially.

As an initial check prior to Q-PCR, PCR was performed to confirm successful cDNA synthesis of RT positive samples and to ensure absence of cDNA in RT negative samples. Reactions were established following reagent volumes outlined in Appendix section 6.1.1 Table 21, using Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH (Sigma, Table 10)) as the housekeeper gene and cDNA RT positive and negative samples amplified following Table 11, as previously optimised (Emily Hodge). Samples were run on a 1.5% agarose electrophoresis gel.

**Table 10. Sequence information for use of GAPDH primers in PCR.**

| Primer Name | Sequence (5'-3') | Binding Location |
|---|---|---|
| GAPDH Forward | AGGTCATCCATGACAACTTTGG | Exon 7 |
| GAPDH Reverse | GCTTCACCACCTTCTTGATGTC | Exon 8 |

**Table 11. Stages of PCR cycles and conditions used for cDNA amplification.**

| Stage | Temperature (°C) | Time | Number of Cycles |
|---|---|---|---|
| Initial denaturation | 94 | 3 min | 1 |
| Denaturation | 94 | 45 sec | 40 |
| Annealing | 55 | 30 sec | |
| Extension | 72 | 1 min 30 sec | |
| Final Extension | 72 | 10 min | 1 |
| Hold | 4 | Hold | - |

### 3.7.2 Quantitative-PCR

Quantitative-PCR (Q-PCR) allows detection and quantitation of gene expression products by detecting fluorescence signals emitted during cycles of DNA amplification. Target specific oligonucleotides used, TaqMan® probes (Applied Biosystems), include a fluorescence dye and a quencher at 5′ and 3′ ends of the sequence respectively. Upon extension of new DNA strands by DNA polymerase the fluorescence dye becomes separated from the quencher and a fluorescent signal is emitted, this is measured at every cycle (40 cycles used). The level of transcript present is related to the amount of fluorescence detected, proportionally increasing with each cycle.

GSTCD mRNA expression levels were quantified by Q-PCR and 18S rRNA used as an endogenous housekeeping gene for normalisation using the $2^{-\Delta\Delta Ct}$ method. 18S rRNA was previously optimised for use as a housekeeping gene in Q-PCR (Ma'en Obeidat) because 18S is known to have invariant expression across a range of tissues and has been identified as the most stable reference gene for normalisation in Q-PCR across a range of human cell culture models (Bas et al. 2004; Aswal et al. 2008; Kuchipudi et al. 2012).

Samples for TaqMan® Q-PCR analysis were set up as follows in Table 13 and Table 14 with duplicate cDNA samples established for analysing GSTCD (Table 12) and the housekeeping gene 18S rRNA pre-designed assay reagent (PDAR) (Applied Biosystems, 4333760F). Samples were run on an Mx3005P Q-PCR machine using MxPro software (Agilent Technologies).

**Table 12. GSTCD primer and probe sequences (Invitrogen) used for TaqMan® Q-PCR, as used for GSTCD amplification previously (Obeidat et al. 2013).** Sequences target both experimentally published GSTCD transcripts: variant 1 and variant 2.

| Sequence Name | Sequence (5'-3') | Binding Location |
|---|---|---|
| GSTCD Probe | ACCTTCCTTTGGGCTGACTGCTGC | Exons 5/6 |
| GSTCD Forward Primer | GGACTCTTGATTGGAATGTTCTCC | Exon 5 |
| GSTCD Reverse Primer | ACCACATAGACAAGGTTGTTCAAC | Exon 6 |

**Table 13. Q-PCR template master mix of cDNA samples and associated controls.**

| Reagents | X1 cDNA Sample (µl) | X1 Control (µl) |
|---|---|---|
| 2X TaqMan® Universal Master Mix | 10 | 10 |
| Water | 3 | 5 |
| cDNA | 2 | - |
| **Total** | **15** | **15** |

**Table 14. Q-PCR primer master mix of PCR reagents for amplification of template cDNA samples (Table 13).**

| | | | | Gene Specific Master Mix solutions | |
|---|---|---|---|---|---|
| | | | | GSTCD | 18S |
| **Reagents** | **Stock Concentration** | **Final Concentration** | **Units** | **X1 Sample (µl)** | **X1 Sample (µl)** |
| Forward Primer | 10000 | 300 | nM | 0.6 | - |
| Reverse Primer | 10000 | 300 | nM | 0.6 | - |
| Probe | 5000 | 100 | nM | 0.4 | - |
| 20X PDAR control | 20 | 1 | X | - | 1 |
| Water | - | - | - | 3.4 | 4 |
| **Total** | | | | **5.00** | **5.00** |

# Transgenic Mouse Model Approaches

## 3.8 Gstcd Knock-out Mice

To investigate the role of Gstcd in the lungs, Gstcd knock-out mice were generated. In collaboration with colleagues at Cornell University (Department of Biomedical Sciences, Michael Kotlikoff research group, including: Jane Lee, Frank Lee, Shaun Reining and Teresa Southard) C57BL/6 GSTCD knock-out embryonic stem cells were obtained from the International Mouse Phenotyping Consortium (IMPC). A member of the IMPC (NorCOMM) utilised a gene targeting approach typically resulting in a null allele (Collins et al. 2007; Nagy et al. 2009), in this instance insertion of the knock-out construct at *Gstcd* exons 4 and 5 would result in transcription of a *lacZ* gene in place of *Gstcd* (Figure 11). The embryonic stem cells were injected into C57BL/6N blastocyst mice and bred for germline transmission of the Gstcd knock-out by collaborators at Cornell University (Michael Kotlikoff group). The colony was expanded and preliminary analysis performed by Cornell University. Reverse transcriptase PCR revealed the absence of expected Gstcd mRNA in homozygous Gstcd mice relative to wild-type (Appendix section 6.4.1) and β-galactosidase (lacZ) staining was conducted to confirm the presence of the *lacZ* insert in Gstcd homozygote and heterozygote mice (Results section 4.2.1). Tissue samples from these mice were sent to University of Nottingham for analysis.

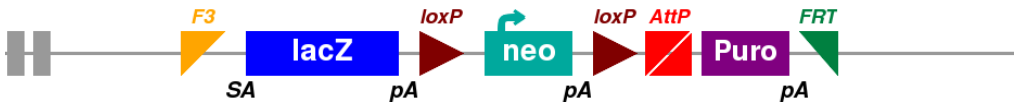



**Figure 11. Diagram displaying the Gstcd$^{tm1(NCOM)}$ reporter-tagged deletion allele with selection cassette (coloured shapes) utilised to create a Gstcd knock-out mouse in the NorCOMM project (IMPC: GSTCD(NCOM) 2015; Grimm 2006).** NorCOMM contributed knock-out mice to the International Knock-out Mouse Consortium (IKMC) (predecessor to the IMPC) using a promoter driven cassette (promoter located in the neomycin resistance gene (neo)). Solid grey boxes are exons 2 and 3. LacZ expression replaces deleted exons 4 and 5 upon integration at *Gstcd* in the mouse genome and selection cassette (neo and Puro) included for isolation of cells containing the inserted gene cassette.

# Methods to Infer Structure and Function of Proteins

## 3.9 Comparative Protein Modelling

Preliminary analysis of potential protein structure and function were investigated based on known structures of homologous proteins. Three principle approaches can be used for protein prediction:

- Comparative Modelling
  - Evolutionarily related homologous templates
- Threading Methods
  - Matches the query protein sequence onto 3D structures of known proteins, including proteins with no evolutionary relationship to the query protein
- *Ab initio* Modelling
  - If no structurally related proteins are available in the Protein Data Bank library structure must be built from scratch, however this is only reliably successful in proteins <120aa and requires significant expertise

Comparative modelling refers to constructing a secondary structure model from the protein of interest amino acid sequence and homology with an evolutionarily related structure (template) (Biasini et al. 2014) whereas threading includes all possible structures for comparison. In this instance, three primary template based protein structure modelling servers were employed to infer potential 3D protein structures and function of GSTCD based on sequence homology (Iterative Threading Assembly Refinement (I-TASSER) (threading method), SWISS-MODEL and Phyre2 (comparative modelling methods)). *Ab initio* modelling was not deemed appropriate for this analysis as this requires detailed knowledge of protein structure bioinformatics and because GSTCD is larger than the maximum recommended protein size for this method, 120aa (GSTCD - 633aa).

Each of the selected protein prediction servers utilises data from the primary 3D protein structure archive; Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank and are amongst the top ranked protein prediction servers, as assessed by critical assessment of protein structure prediction (CASP) (Kopp et al. 2007; Cozzetto et al. 2009; Mariani et al. 2011; Kryshtafovych et al. 2014).

Sequences for both published human GSTCD protein variant sequences (NCBI references: GSTCD variant 1 NP_001026890.2 and GSTCD variant 2 NP_079027.2) were input into the servers and analysed by comparing results from the different protein prediction servers alongside results for the positive control Glutathione S-transferase Mu 5 (NCBI reference: NP_000842.2). In all instances where GSTCD variant 1 and 2 are discussed, these refer to the aforementioned NCBI DNA and protein sequences.

# Protein Analyses

## 3.10 SDS-PAGE and Western Blotting

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) is a technique used for separating out individual proteins from a mixture based on molecular weight, independent of other physical features such as variable overall charge. This involves denaturing proteins using heat and Sodium dodecyl sulphate (SDS). In addition, SDS imparts a negative charge on the denatured proteins which enables them to be separated by molecular weight alone. This occurs as a consequence of all of the negatively charged proteins migrating through a gel matrix towards a positive charge; the speed at which the proteins can move through the matrix is inversely proportional to molecular weight.

After transferring the migrated proteins to a polyvinyl difluoride (PVDF) membrane, individual bands can be visualised by antibody labelling in Western blotting. This exploits antibody specificity for binding to particular protein antigens and using this for signal amplification of a protein of interest. A primary antibody targeted to the protein of interest is first applied, followed by further signal amplification by a HRP conjugated secondary antibody targeted against the immunoglobulin G (IgG) of the species of animal in which the primary antibody was raised. Bands are visualised by an enhanced chemiluminescence detection system, relying on HRP enzyme activity which produces light when catalysing the conversion of chemiluminescent substrates.

SDS-PAGE and Western blotting techniques were used as methods to compare the levels of GSTCD protein in siRNA knock-down treated hBECs, relative to untransfected and scrambled siRNA transfected controls. Several reagents and buffers were produced prior to SDS-PAGE and Western blotting outlined in Table 22 (Appendix section 6.1.2). SDS-PAGE gels were produced following Table 23 (Appendix section 6.1.2) adding reagents in the sequences outlined, using Bio-Rad Mini-Protean II gel electrophoresis apparatus.

### 3.10.1 SDS-PAGE

Sufficient protein from each sample, including recombinant human GSTCD protein as positive control (Abnova, H00079807-P01), was diluted to 1µg/µl in a total volume of 40µl and 10µl 5X reducing gel loading buffer added (Appendix section 6.1.2, Table 22). Samples were denatured at 95°C for 5 minutes then centrifuged at 4°C for 4 minutes at top speed. Samples were added to the SDS-PAGE gels and Spectra Multicolour Broad Range Protein Ladder (ThermoScientific, 26634) was used to gauge protein band sizes. Samples were run through the stacking gel at 100V and the resolving gel at 200V (Appendix section 6.1.2, Table 23) in 1X running buffer using Bio-Rad Mini-Protean II gel electrophoresis apparatus.

#### *3.10.1.1 Transfer of Protein to PVDF Membrane*

Components for transfer of protein samples from gel to PVDF membrane were equilibrated in 1X transfer buffer for 15 minutes prior to assembly, with PVDF membranes having been first activated in 100% methanol followed by three washes in deionised water. Transfer components were assembled into cassettes and protein transfer was performed in 1X transfer buffer (Appendix section 6.1.2, Table 22) at 100V for 1 hour.

### 3.10.2 Western Blotting

#### *3.10.2.1 Probing Proteins using Specific Antibodies*

After electrophoresis transfer PVDF membranes were washed for 15 minutes three times in PBS-T. Membranes were cut into sections for use with different antibodies and/or dilutions and incubated in blocking buffer (Appendix section 6.1.2, Table 22) for 1 hour. Two polyclonal primary antibodies for GSTCD were used at dilutions of 1:500 and 1:1000 (Proteintech, 17502-1-AP and Abnova, H00079807-B01) with β-actin antibody (Abcam, ab8227) diluted 1:5000 used as

a control. Primary antibodies were diluted in blocking buffer and incubated on membranes for 1 hour. Subsequently, membranes were stored at 4°C overnight.

Three membrane washes were carried out as described previously. Membranes were incubated with relevant horseradish-peroxidase conjugated secondary antibodies diluted 1:5000, for 1 hour (polyclonal goat anti-rabbit secondary antibody (Sigma, A0545) for GSTCD (Proteintech) and β-actin and polyclonal goat anti-mouse secondary antibody (Sigma, A0168) for GSTCD (Abnova)). Membranes were washed three times for 15 minutes in PBS-T.

#### *3.10.2.2 Visualisation of Signal*

Protein bands were visualised using photographic film (Fisher, 28-9068-39), following enhanced chemiluminescence detection (ECL) according to manufacturer's instructions (GE Healthcare Amersham, RPN2209), in a developing cassette for appropriate time intervals. This was followed by several washes: washed in developer solution, rinsed off developer solution in water, washed photographic film in fixer solution until film turned clear then rinsed off fixer solution in water (Appendix section 6.1.2, Table 22).

Band intensity of Western blots was semi-quantified using Fiji image processing package (Schindelin et al. 2012). Firstly colours were inverted on all images of Western blots and a selection box of minimal sufficient size to enclose all bands individually (GSTCD and β-actin) was drawn. This same box was placed over all bands separately and light intensity values obtained. GSTCD values for each condition were then normalised against the respective housekeeping β-actin values.

## 3.11 Immunohistochemistry

Immunohistochemistry (IHC) allows the visualisation of protein localisation within specific cells or regions of tissue, utilising labelled antibodies which are used for signal amplification similarly to their use in Western blotting (as described in section 3.10). In this instance, the secondary antibody is conjugated to HRP which, in the presence of hydrogen peroxide, converts 3,3'-Diaminobenzidine (DAB) to an insoluble brown product at sites of immunoreactivity.

Paraffin Sections of C57BL/6 wild-type mouse lung, kidney, spleen and intestine were obtained from the department Leukocyte Biology, Imperial College London, UK and C57BL/6N Gstcd wild-type (+/+), heterozygous (+/-) and homozygous (-/-) knock-out lung and kidney tissues were obtained from Department of Biomedical Sciences, Cornell University, USA.

Analysis was performed on sections of paraffin embedded adult lung tissue age matched at 10 weeks, recommended as the ideal age for adult analysis by the IMPC.

IHC was performed using the EnVision™ Detection Systems Peroxidase/DAB kit (DAKO K5007). Firstly, wax was removed from slides using two 5 minute washes in Histoclear II (National Diagnostics, HS-202), followed by hydrating sections in an ethanol series (100%, 95% and 70%) using two 5 minute washes at each stage. Afterwards slides were washed under a running tap and washed with PBS. Antigen retrieval was performed using citrate buffer heated for 10 minutes and cooled for 20 minutes in a steamer (Breville). PBS – bovine serum albumin (BSA) – Triton™ X-100 (PBS-BT) (Appendix section 6.1.2, Table 24) was used to block sections (100µl/section) prior to incubation with primary antibodies following dilutions outlined in Table 15. Negative controls were established by including: no primary antibody treatment (incubation with PBS-BT), isotype controls (IgG solution from matched animal that primary antibody is raised in, to assess potential non-specific binding). Kidney and spleen tissues were used as positive controls (tissue from an organ known to have high expression of protein of

interest). Tissue sections were treated with the primary antibodies or matched rabbit isotype control overnight at 4°C in a humidified environment.

**Table 15. Primary antibodies and associated isotype controls used in IHC.**

| Antibody | Supplier | Catalogue Number | Dilution |
|---|---|---|---|
| GSTCD | Proteintech | 17502-1-AP | 1:50 |
| Rabbit Isotype | Invitrogen | 10500C | Matched to GSTCD antibody stock concentration: 1:21<br>Followed by: 1:50 |
| | | | |
| Ki67 | Proteintech | 19972-1-AP | 1:50 |
| | | | 1:100 |
| Rabbit Isotype | Invitrogen | 10500C | Matched to Ki67 antibody stock concentration: 1:23<br>Followed by: 1:50 |
| | | | |
| Phospho histone H3 (PH3) | Abcam | ab5176 | 1:50 |
| | | | 1:100 |
| Rabbit Isotype | Invitrogen | 10500C | Matched to PH3 antibody stock concentration: 1:65<br>Followed by: 1:50 |

Tissue slides were washed in PBS for 5 minutes three times and the Envision™ kit secondary antibody applied to all sections (1 drop/section) for 30 minutes at room temperature. PBS washes were repeated and sections incubated with DAB solution for 5 minutes then washed in distilled water to stop the reaction. Nuclei were stained using Mayer's Haematoxylin incubated on sections for 30 seconds and washed off under a running tap for 10 minutes. Lastly, sections were dehydrated back through an ethanol series as above and washed twice in Histoclear II prior to mounting with Vectamount permanent mounting medium (Vector Laboratories, H-5000) and cover-slipping. Slides were imaged on a light microscope (Nikon Diaphot 300 inverted microscope).

## 3.12 Haematoxylin and Eosin Staining

Haematoxylin and Eosin (H&E) staining is used to add contrast to a tissue sample, enabling definition of different tissue structures and components by dying them varied colours (based on pink, purple and blue colours).

The staining method firstly involves application of haematoxylin which turns blue in alkaline conditions; such as when binding to basophilic substances like DNA and RNA (nucleic acids are negatively charged due to a phosphate backbone). Therefore nuclei can be identified by blue staining. Nuclear staining is followed by counterstaining with eosin which binds to acidophilic substances, for instance positively charged proteins, whether these are intra- or extra-cellular. The majority of the cytoplasm is eosinophilic. This step colours many non-nuclear components in varied shades of pink, based on the amount of positively charged substances in the region.

Wax was removed from paraffin embedded lung tissue sections as previously described for IHC (Methods section 3.11) and washed in distilled water. Nuclei in sections were stained by 5 minute incubation with Mayer's haematoxylin (Sigma, MHS16) followed by rinsing in running tap water for 10 minutes. Colour in sections was differentiated in 0.3% acid alcohol (hydrochloric acid (Fisher Scientific, H/1200/PB17) and ethanol (Sigma, 322221-2) followed by washing in running tap water for 5 minutes. Sections were then placed in 1% eosin (Millipore, 115935) for 5 minutes and washed in running tap water for 10 minutes. Slides were then dehydrated in an ethanol series and mounted as described in Methods section 3.11.

# Statistical Analyses

## 3.13 Statistics

Data were analysed using GraphPad Prism 6 software (GraphPad Software, Inc.). Combined data from independent experiments were displayed as the mean value ± standard error of the mean unless otherwise stated.

Experiments comparing several treatment conditions (such as varied media and siRNA knock-down) were normalised to the control condition within each replicate. The average of each treatment condition was calculated and divided by the average value of the control condition within each replicate. Replicate experimental results were then assessed for statistical significance as described below.

Statistical significance was assessed using an unpaired t-test to compare the means of two groups, this was applied to raw data values for cell growth assays containing only two experimental conditions. Statistical significance in experiments consisting of only one experimental condition after normalisation was evaluated by a one sample t-test. This assessed deviation in the mean of the experimental condition against a hypothetical mean, in this instance 100% after normalisation against a standard control.

Statistical significance in all other experiments was assessed by one-way ANOVA, comparing the means of three or more unmatched groups, followed by Dunnett's test when comparing multiple conditions to a control. Analysis of triplicate readings from a single experiment was displayed as the mean value ± standard deviation, as mentioned in the appropriate figure. For all statistical analysis as described above, a p-value of $<0.05$ was considered to be significant.

# 4 Results

## 4.1 GSTCD siRNA Knock-Down in Human Bronchial Epithelial Cells Results in Reduced Cell Growth

Polymorphisms spanning the GSTCD gene region at locus 4q24 have been associated with the lung function measure $FEV_1$ in many recent studies. This has also been followed by identification of several associations with COPD diagnosis, hence indicating a likely role for GSTCD in the respiratory system. A previous study in human lung tissue revealed GSTCD localisation in bronchial epithelium and alveolar epithelial cells, as well as confirming that hBECs demonstrate the highest levels of GSTCD mRNA across a range of cell types (Obeidat et al. 2013).

Cell growth and proliferation are essential homeostasis requirements throughout lung development and adulthood. Cell growth is particularly evident and important for the two lung cell types in which GSTCD was identified to primarily localise: alveolar epithelial and bronchial epithelial cells. Both cell types are constantly renewed, required by 'front-line' airway surfaces. For example alveolar epithelial cells are required for the important role of surfactant production and providing a source of alveolar epithelial type I cells. Alveolar epithelial type I cells are essential for gas exchange in the lungs and are easily damaged, hence these cells are in continual need of replenishing from the progenitor type II cells. Additionally, bronchial epithelial cells line the outer surface of the airways and are central to lung defence against inhaled particles and pathogens; therefore these cells have a slow but constant level of renewal (Crystal et al. 2012). Critically, perturbations to cell growth can contribute to lung disease in the form of COPD and chronic bronchitis through airway remodelling (Jeffery 2001; Demedts et al. 2006; Lapperre et al. 2007).

The primary cell based model for alveolar epithelial cells is the alveolar epithelial type II cell line, A549. Alternatively, bronchial epithelial cells are a primary cell model and were selected for initial investigation of GSTCD because morphology and behaviour in primary cells are considered more representative of cells grown *in vivo* (during the early passages of growth *in vitro*) compared to a cell line. Bronchial epithelial cells are also a convenient airway relevant model to

study cell homeostasis outcomes, therefore this work has focussed on understanding GSTCD in hBECs.

The aim of this section was to investigate whether GSTCD has a role in cell growth, an important feature of bronchial epithelial cells. This was performed using a siRNA knock-down approach in hBECs to study cellular effects caused by reduction of GSTCD expression.

### 4.1.1 GSTCD Antibody Optimisation

Prior to investigating GSTCD protein knock-down and Gstcd localisation in a murine model, I first determined the specificity of different GSTCD antibodies by Western blotting with hBEC cell lysate and subsequently demonstrated Gstcd immunostaining in wild-type mouse lung sections.

Currently there is very little published research about GSTCD, previous research has indicated localisation in the bronchial epithelium and alveolar epithelial cells (Obeidat et al. 2013). Hence two antibodies (Proteintech and Abnova) were selected and their specificity in identifying GSTCD protein from hBEC samples was investigated. Western botting was performed using both antibodies at dilutions of 1:500 and 1:1000 and human recombinant GSTCD protein was used as a positive sample control. All hBEC samples were obtained from a single stock of hBEC cells. The positive antibody control utilised was β-actin, standardly used as a housekeeping gene in expression studies.

β-actin bands (42kDa) were identified in the hBEC sample run on both SDS-PAGE gels and were absent in the recombinant GSTCD samples (Figure 12A and B). Individual bands were present in hBEC lysate at the expected molecular weight of the primary GSTCD variant (variant 1, 71kDa) for Proteintech antibody at both dilutions used (Figure 12A). The band seen in the 1:1000 antibody dilution was weaker than that using 1:500 antibody dilution. Additionally, the Proteintech antibody detected a band of higher molecular weight (~105-110kDa) in GSTCD recombinant protein lysate (positive control) (Figure 12A). The band identified in the positive control appeared to be at a higher molecular weight than expected

for the recombinant GSTCD protein (97.4kDa) which includes a protein purification tag (glutathione S-transferase (GST), 26.4kDa). Conversely to Proteintech antibody, the Abnova antibody detected bands of noticeably higher molecular weight than 71kDa (approximately 85-90KDa) in hBEC lysate (Figure 12B). The Abnova antibody also successfully identified recombinant GSTCD protein (Figure 12B), at a similar higher than expected molecular weight as seen with the Proteintech antibody.

In conclusion, the Proteintech GSTCD antibody at 1:500 dilution was selected for continued use as this antibody demonstrated the most accurate result for expected GSTCD molecular weight. Additional confirmation that this antibody correctly identified GSTCD protein was shown by demonstrating reduced levels of GSTCD protein following knock-down with GSTCD siRNA in hBEC cells (Results section 4.1.2.2 Figure 14).

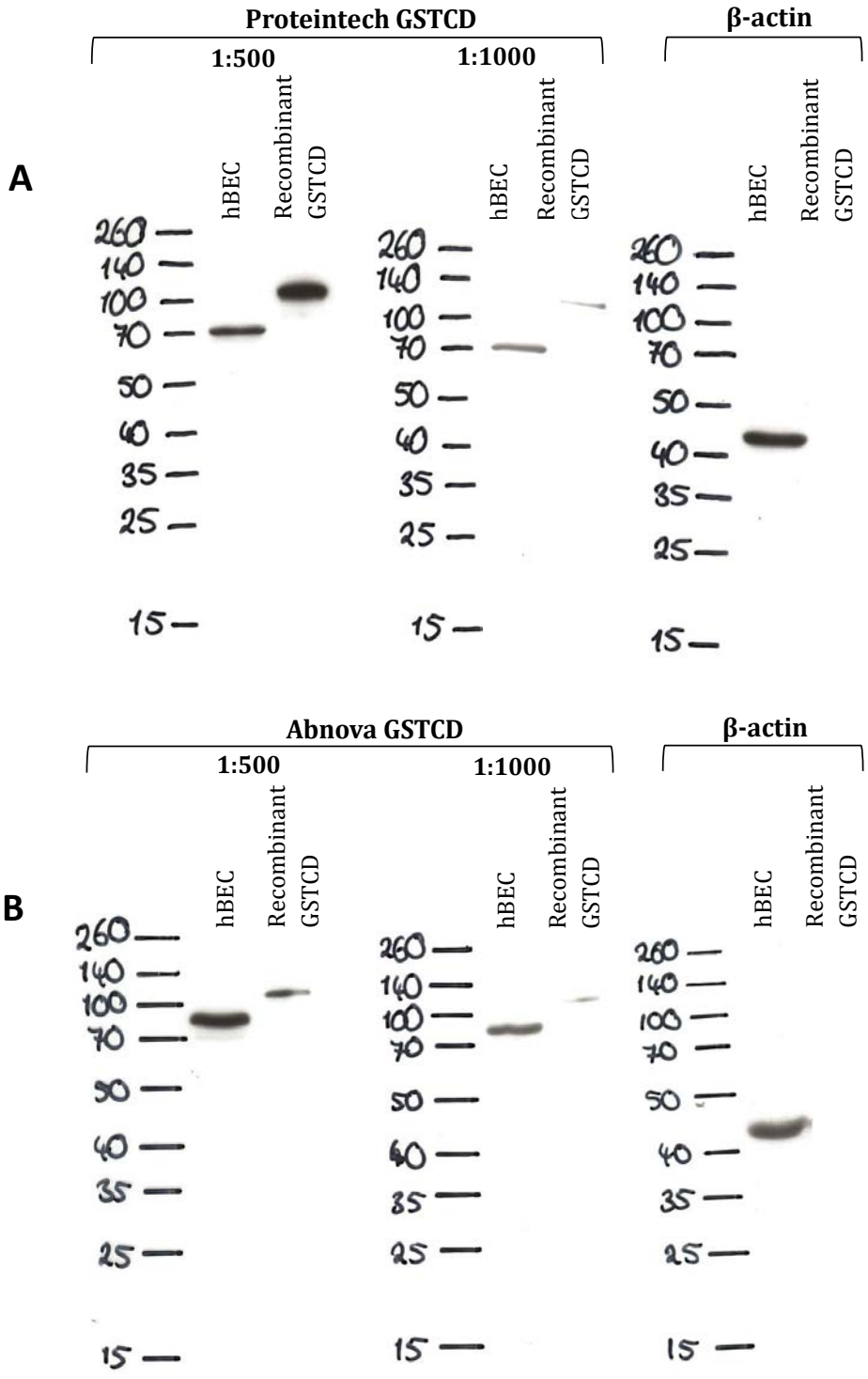



**Figure 12. Two antibodies identify GSTCD protein in hBEC cell lysates by Western blotting.** Optimisation of two GSTCD antibodies by Western blotting reveals predicted GSTCD molecular weight bands (71kDa) present at two dilutions for Proteintech antibody in hBEC samples and appropriate bands present for recombinant GSTCD protein positive controls (approximately 97.4kDa as this includes a GST Tag used for purification) (A). Appropriate bands are detected in recombinant GSTCD sample with Abnova antibody but higher than expected molecular weight bands are present in hBEC samples (approximately 85-90kDa) (B). Control β-actin bands are present for each hBEC sample used (42kDa) (A and B) and absent in recombinant GSTCD protein samples. Data is representative of n=3 experiments.

### 4.1.2 Confirmation of GSTCD siRNA Knock-down

In a human model of loss of function for GSTCD, expression was reduced using a siRNA knock-down approach. This method involves disrupting the translation of GSTCD mRNA by stimulating degradation of GSTCD mRNA transcripts. Reduction of GSTCD expression levels were assessed in mRNA and protein to identify degree of knock-down established and to determine whether a reduction at the mRNA level is similarly translated at the protein level.

#### *4.1.2.1 GSTCD mRNA levels can be reproducibly reduced in primary bronchial epithelial cells*

Previous work has shown that 1nM siRNA incubated for 48h is an optimal incubation length for GSTCD mRNA knock-down and produces 86.7%, 88% and 90% mRNA expression knock-down using three siRNA sequences (siRNA A, B and C respectively, n=3) (unpublished, Amanda Henry, University of Nottingham). The two siRNA sequences which produced the highest levels of knock-down, siRNA B and C, were also used in a time-course experiment (12h, 24h, 48h and 72h) which demonstrated comparable knock-down across time (n=3). I then replicated these results in hBECs for both siRNA sequences in a new batch of siRNA reagents, to confirm reproducibility and robustness of the approach (GSTCD knocked-down by 90% (siRNA B) and 92% (siRNA C)) (Figure 13). The scrambled control appeared to reduce GSTCD mRNA expression relative to untransfected (reduced to 64%). Transfection with the scrambled siRNA is the true control in these experiments as this condition also involves addition of RNA and transfection reagent to the cells, replicating exact conditions used for GSTCD siRNA knock-down (whereas untransfected is simply regular cell growth conditions). This includes using the same siRNA nucleotide composition in the scrambled siRNA control compared to the GSTCD targeting siRNAs (siRNA B and C), in a different sequence which does not target a known gene. Therefore scrambled siRNA is used as the control comparison for subsequent experiments with data normalised against the untransfected condition, as the standard cell growth condition.

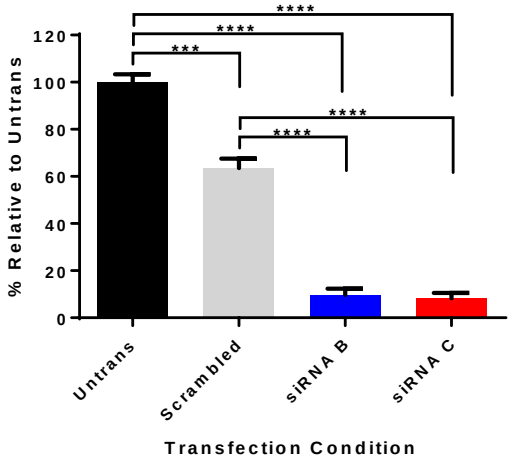



**Figure 13. GSTCD siRNA knock-down in hBECs.** Proportional changes in GSTCD expression are shown normalised to untransfected cells. siRNA B and siRNA C conditions significantly reduced GSTCD expression relative to Untrans (shown above, reduced to 10% ($p<0.0001$ indicated by ****) and 8% ($p<0.0001$) respectively) and scrambled controls (data not shown, reduced to 15% ($p<0.0001$) and 13% ($p<0.0001$) respectively). The scrambled siRNA appears to exert an effect on GSTCD expression, reducing mRNA levels to 64% ($p<0.001$) relative to Untrans, although this condition remains the true control in this experiment by replicating all variables used in GSTCD siRNA knock-down conditions. Average values from triplicate readings in a single experiment are shown therefore standard deviation bars are presented. Key: Untrans - untransfected control, Scrambled - transfection with scrambled siRNA control, siRNA B/C - transfection with GSTCD knock-down siRNA.

#### *4.1.2.2 GSTCD Protein Levels are Significantly Reduced in Primary Bronchial Epithelial Cells Targeted by GSTCD siRNA*

To explore whether GSTCD protein levels are similarly reduced as mRNA levels in optimal GSTCD siRNA knock-down conditions, Western blotting was conducted on protein lysates of knock-down treated hBEC samples (siRNA B and C, n=3). This confirmed notably lower levels of GSTCD protein in the siRNA B and siRNA C samples when compared against both untransfected and scrambled siRNA treated controls (Figure 14). Band intensities obtained by Western blotting are indicative of relative amount of protein per sample. These were semi-quantified for each siRNA experiment (values shown in Appendix section 6.5.1 Table 26) and normalised against the housekeeping gene β-actin to account for potential variations between samples. β-actin is used as a standard endogenous control as its expression levels are understood to remain constant under different experimental conditions and across cells of different tissues. Semi-quantified Western blotting results are summarised in Figure 15, where GSTCD expression was reduced by 63% (siRNA B) and 74% (siRNA C) relative to untransfected control lysates (n=3) with significance values determined by one-way ANOVA (siRNA B $p<0.001$ and C $p<0.01$). Therefore, these results demonstrate a significant reduction of GSTCD protein levels in hBECs as well as in mRNA utilising the GSTCD siRNA B and siRNA C knock-down sequences.

In summary, both siRNA B and siRNA C sequences effectively reduced GSTCD mRNA and protein expression in hBECs compared to both standard controls; untransfected and scrambled siRNA conditions.

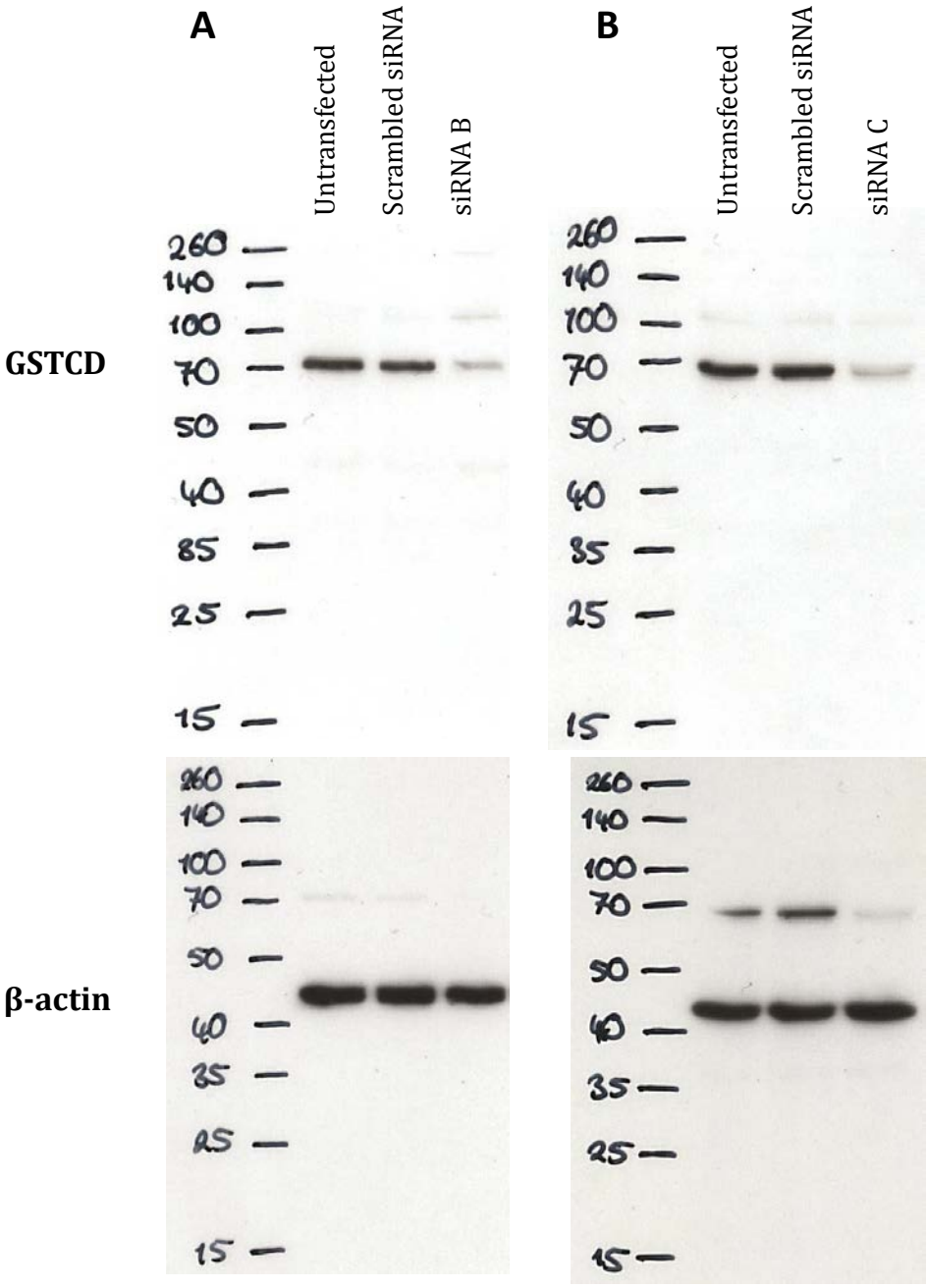



**Figure 14. GSTCD protein levels are reduced in hBEC lysates following siRNA knock-down.** 1nM siRNA B (A) and C (B) was used and incubated with hBECs for 48h (Amanda Henry), representative images from Western blotting shown of n=3 for each siRNA. GSTCD protein levels were noticeably reduced relative to untransfected and scrambled controls for each siRNA in all samples. Top row: Western blotting with protein samples from GSTCD siRNA mediated knock-down performed in hBECs. GSTCD was detected around 71kDa as expected. Bottom row: Blots were stripped and re-probed with an antibody for the housekeeping gene β-actin (42kDa) to confirm even loading of samples and aid quantification, expected bands were demonstrated across samples. GSTCD bands appear in siRNA C β-actin blot, re-amplified after antibody stripping due to a shared secondary antibody utilised for both GSTCD and β-actin primary antibodies. Results are representative of n=3 independent experiments.

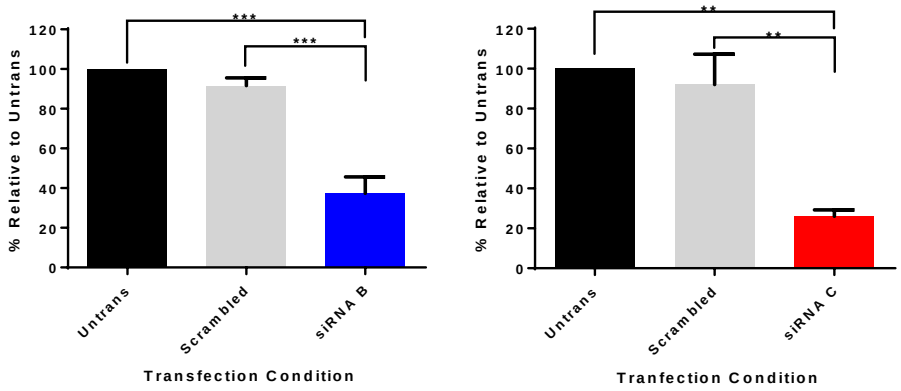



**Figure 15. Semi-quantification of GSTCD protein levels following siRNA B (A) ($p<0.001$) and C (B) ($p<0.01$) knock-down in hBECs relative to controls.** Proportional changes are shown with results normalised against untransfected control. GSTCD protein expression is significantly reduced for siRNA B (reduced to 37% $p<0.001$, indicated by ***) (A) and siRNA C (reduced to 26% $p<0.01$, indicated by **) (B) relative to both untransfected and scrambled controls. Key: Untrans - untransfected control, Scrambled - transfection with scrambled siRNA control, siRNA B/C - transfection with GSTCD knock-down siRNA. Data are representative of n=3 independent experiments.

### 4.1.3 Cell Growth Assay Optimisation

Cell growth was selected as an important measure of cell homeostasis in bronchial epithelial cells to investigate reduction of GSTCD. Firstly, several cell growth assays were trialled for measuring hBEC growth rates in minimal media (BEBM + ITS) and optimal media (BEGM) at a single time-point; aiming to identify the most reliable and sensitive assays capable of detecting a difference in cell growth in minimal and optimal media conditions. Secondly, longer incubation times were then used to assess potential changes over time. This was performed using two assays identified in initial experiments.

#### *4.1.3.1 Several Cell Growth Assays Identify a Change in Cell Growth between Minimal and Optimal Media*

Cell growth assays were compared by ability to sensitively detect and quantify differences in hBEC growth when minimal and optimal media are used to limit and promote optimal cell growth respectively. Preliminary comparisons were performed using a 24h incubation time to assess initial cell growth differences in the two media types. Additionally, assays were compared to identify a potential replacement assay for the previously standardly used MTT assay, as this is no longer considered the most accurate measure of cell growth.

The assays compared included cell counting, a traditional measure of total cell number measuring cells under a microscope. MTT and CCK-8 assays were compared which measure cell metabolism, providing information on approximate total numbers of living cells. The Click-iT® EdU assay was used to measure actively proliferating cells, therefore giving an indication of number of cells which have replicated. Total cell number was also measured by the CyQUANT® NF assay, which measures both living and dead cells (unlike MTT and CCK-8 assays).

Data was normalised to minimal media for comparison between different assays (i.e. 100%), cell growth in optimal media was measured at 106% (cell counting, Figure 16A), 135% (MTT, Figure 16B), 99% (CCK-8, Figure 16C), 141% (Click-iT® EdU, Figure 16D) and 117% (CyQUANT® NF, Figure 16E). The most reliable

assays adept at distinguishing expected differences in cell growth between minimal and optimal media were MTT, Click-iT® EdU and CyQUANT® NF (Figure 16B, D and E). The majority of assays demonstrated very low levels of background in vehicle controls (cells grown in minimal or optimal media, with absorbance/fluorescence measured with only the assay specific buffer solution) whereas the Click-iT® EdU assay exhibited high background signals in vehicle control conditions.

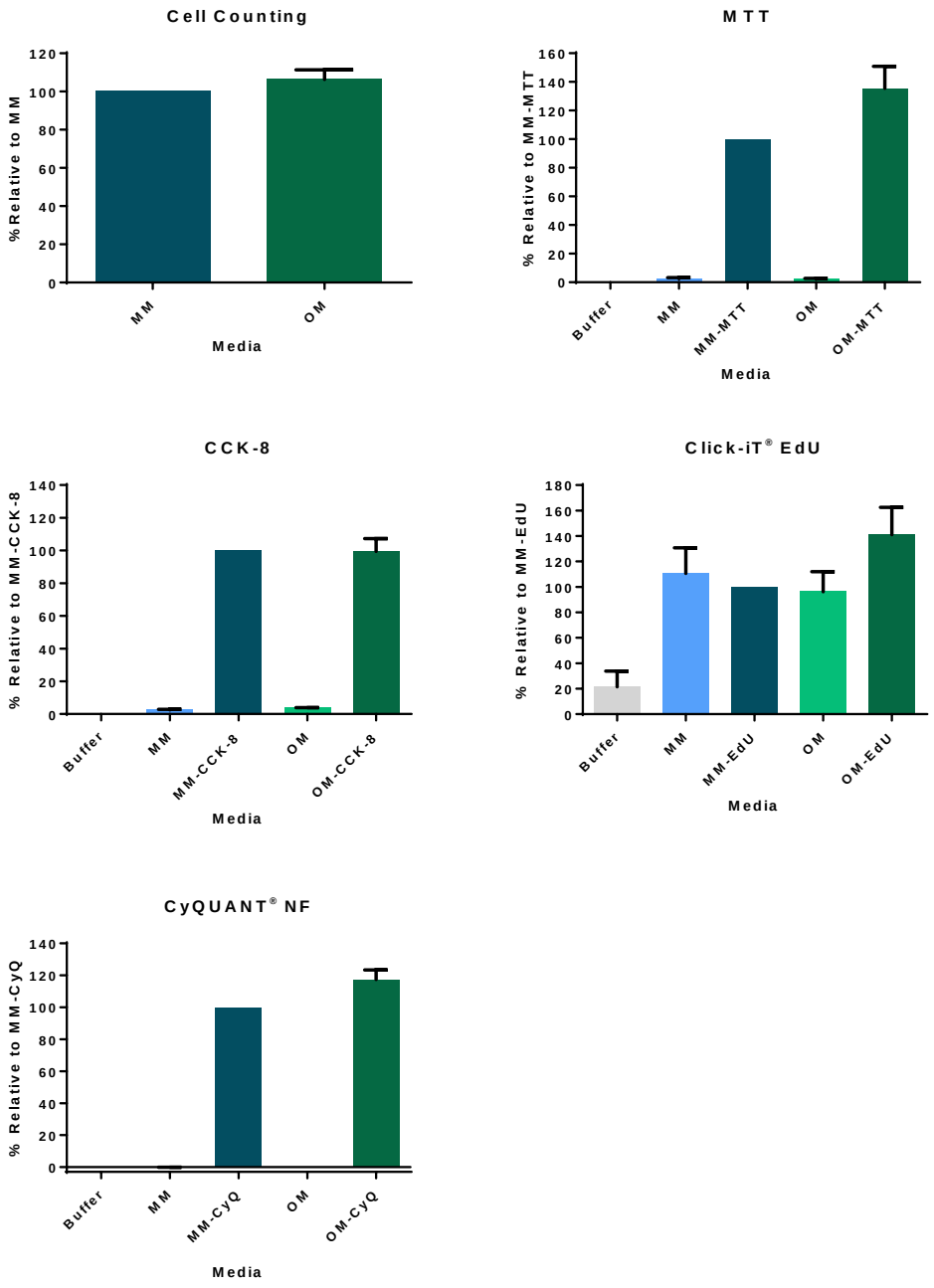



**Figure 16. Comparison of cell growth assays using hBECs cultured in minimal and optimal media for 24h shows modest increase of cell growth in optimal media conditions.** Data is normalised to minimal media for each assay to allow direct comparison. Cell growth was measured by: cell counting (A), MTT assay (B), CCK-8 assay (C), Click-iT® EdU (D), CyQUANT® NF (E). MTT, Click-iT® EdU and CyQUANT® NF assays demonstrate the most sensitive detection of differences in cell growth between minimal and optimal media. No clear differences in cell growth were observed between media conditions for the cell counting and CCK-8 assays. Controls used in Click-iT® EdU assay show unexpected high signals. Key: Buffer - no cells grown, measured with assay specific vehicle control; MM - cells grown in minimal media, measured with assay specific vehicle control; MM-[MTT] - cells grown in minimal media, measured with assay specific reagent; OM - cells grown in optimal media, measured with assay specific vehicle control; OM-[MTT] - cells grown in optimal media, measured with assay specific reagent. Data shown are mean values from n=3 independent experiments, with individual conditions completed in quadruplicate. Since data were normalised to minimal media producing a set value of 100% in one of the only two experimental conditions analysed (minimal or optimal media with assay specific reagent), the one sample t-test was performed to assess significance. No significant variation was observed from the normalised value (100%) in these data.

#### *4.1.3.2 Differences in Cell Growth between Minimal and Optimal Media Conditions are Reliable across a Time-course*

Following initial comparison of several cell growth assays, Click-iT® EdU and CyQUANT® NF assays were taken forward for analysis in a time-course experiment; since these assays identified a clear increase in hBEC growth in optimal media compared to minimal media at 24h and measure all cells as opposed to only living or actively metabolising cells (MTT assay).

Raw data results for both assays demonstrate higher levels of cell growth in optimal media compared to relative media at all time-points (Figure 17). Click-iT® EdU assay indicates a gradual reduction in actively proliferating cells over time for both media conditions equivalently (Figure 17A). This reduction is likely attributable to decreasing availability of supplementary nutrients and increasing cell confluence; therefore by 72h cell growth is no longer exponential. In contrast, CyQUANT® NF optimal media results demonstrate continued increases in total cell number over-time (Figure 17B), reflecting the high levels of actively proliferating cells evident from the Click-iT® EdU assay. Total cell number is notably increased in the optimal media compared to minimal media for the CyQUANT® NF assay, with the difference between these conditions also increasing over time. Interestingly, the CyQUANT® NF assay indicates that total cell numbers in minimal media are fairly consistent across the time course, suggesting that total cell number is simply maintained despite cells proliferating in a similar pattern, albeit at a decreased level, as cells grown in optimal media (as demonstrated by Click-iT® EdU).

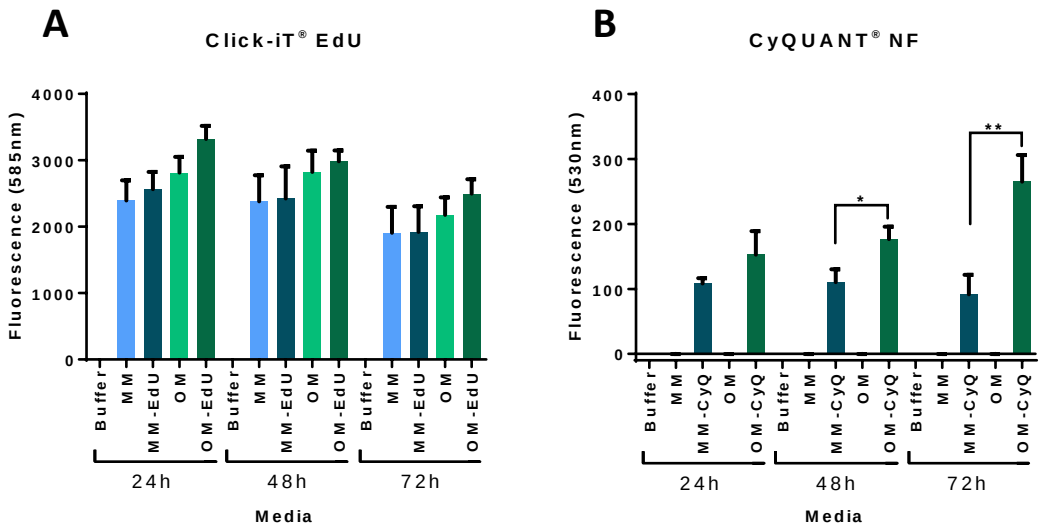



**Figure 17. Raw data for Click-iT® EdU and CyQUANT® NF assays demonstrate expected increase in cell growth for hBECs cultured in optimal media compared to minimal media across a time-course.** Click-iT® EdU assay indicates a higher level of actively proliferating cells (A) and CyQUANT® NF assay confirms increases in total cell number at all time-points (B) (significance at 48h $p<0.05$ (represented by *) and 72h $p<0.01$ (represented by **)) for optimal media experimental conditions compared to minimal media experimental conditions. High background signals are also seen in Click-iT® EdU assay control conditions across time. Key: Buffer - no cells grown, measured with assay specific vehicle control; MM - cells grown in minimal media, measured with assay specific vehicle control; MM-[MTT] - cells grown in minimal media, measured with assay specific reagent; OM - cells grown in optimal media, measured with assay specific vehicle control; OM-[MTT] - cells grown in optimal media, measured with assay specific reagent. Data shown are mean values from n=3 independent experiments, with individual conditions completed in quadruplicate. An unpaired t-test was used to assess significance in raw data between the two experimental conditions; minimal or optimal media with assay specific reagent.

Results were also normalised to the control condition minimal media to allow direct comparison of both assays. Results for Click-iT® EdU showed increased levels of actively proliferating cells in optimal media relative to minimal media as expected with little variation across time, proliferation in optimal media increased to 131% (24h), 133% (48h) and 138% (72h) (Figure 18A). Similarly, in the CyQUANT® NF media comparisons there was a distinct difference in cell growth between media types increasing over time, with cell growth increased in optimal media compared to minimal media at all time-points (140% at 24h, 162% at 48h and 325% at 72h) (Figure 18B). Background signals remained high in the vehicle control conditions for Click-iT® EdU at all time-points.

The time-course of Click-iT® EdU and CyQUANT® NF demonstrated that these assays were also capable of detecting expected variation in cell growth between

these media types at several time-points, prior to use in analysing cell growth in GSTCD knock-down experiments. Continued use of the Click-iT® EdU assay despite high background signals was based on extensive use in published research (Yokota et al. 2012; Petrilli et al. 2014; Salic & Mitchison 2008; Xia et al. 2010; Morrison et al. 2014; Lentz et al. 2009; Feng et al. 2011) and that the assay has been widely approved as being highly sensitive. In particular the Click-iT® EdU assay has been demonstrated to be a significant improvement on the earlier thymidine labelling method utilising BrdU (Cappella et al. 2008; Buck et al. 2008).

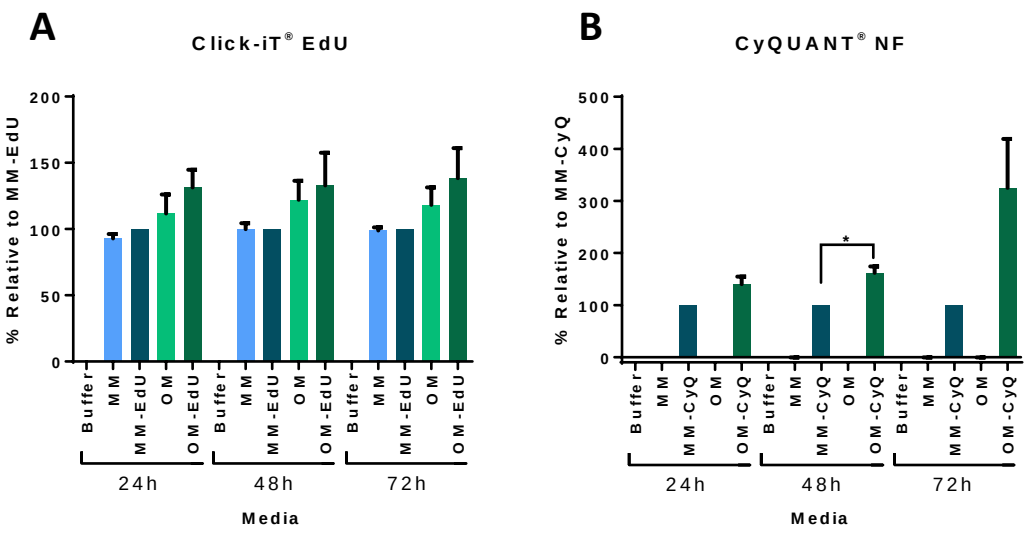



**Figure 18. Click-iT® EdU and CyQUANT® NF assays demonstrate expected increase in cell growth for hBECs cultured in optimal media compared to minimal media across a time-course.** Data is normalised to minimal media for each assay to allow direct comparison. Click-iT® EdU assay shows a higher level of actively proliferating cells (A) and CyQUANT® NF assay confirms increases in total cell number at all time-points (B) (significance at 48h $p<0.05$, represented by *) for optimal media experimental conditions relative to minimal media experimental conditions. High background signals are also seen in Click-iT® EdU assay control conditions across time. Key: Buffer - no cells grown, measured with assay specific vehicle control; MM - cells grown in minimal media, measured with assay specific vehicle control; MM-[MTT] - cells grown in minimal media, measured with assay specific reagent; OM - cells grown in optimal media, measured with assay specific vehicle control; OM-[MTT] - cells grown in optimal media, measured with assay specific reagent. Data shown are mean values from n=3 independent experiments, with individual conditions completed in quadruplicate. Since data were normalised to minimal media producing a set value of 100% in one of the only two experimental conditions analysed (minimal or optimal media with assay specific reagent), the one sample t-test was performed to assess significance.

### 4.1.4 Investigating the Effect of GSTCD Knock-down on Cell Growth in hBECs

To assess potential effects of GSTCD knock-down on cell growth two assays were utilised to investigate actively proliferating cells (Click-iT® EdU assay) and total cell number (CyQUANT® NF assay). GSTCD expression was knocked-down in hBECs using 1nM siRNA B and C incubated across a time course following the methodology optimised previously. Two siRNA sequences were utilised to account for potential off-target effects specific to one siRNA sequence.

No discernible changes in levels of actively proliferating cells were observed in either GSTCD knock-down sample relative to untransfected or scrambled controls using the Click-iT® EdU assay (Figure 19A). However measuring total cell number with CyQUANT® NF assay appeared to show reduced cell numbers in the knock-down samples increasingly across time relative to both untransfected and scrambled siRNA treated hBECs, with siRNA C displaying the strongest effect (Figure 19B). At 72h total cell number was significantly reduced compared to scrambled siRNA for both siRNA B (reduced by 33% $p<0.001$) and siRNA C (reduced by 48% $p<0.0001$) when results are normalised to untransfected. A minor reduction in total cell number was seen in the scrambled control over time, as could potentially be expected from previous results in this hBEC donor, indicating a reduction in GSTCD mRNA in this condition relative to untransfected.

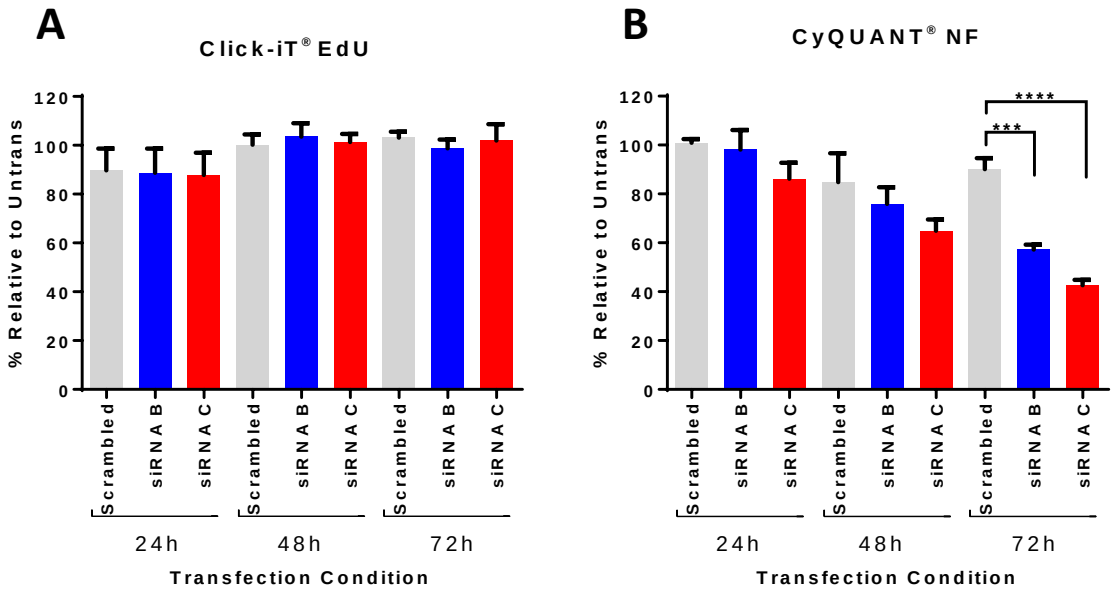



**Figure 19. GSTCD siRNA knock-down in hBECs does not affect proliferation rate (A) but does result in reduced total cell number (B) but over time.** Data is normalised to Untrans controls (100%) bars not shown. No change in proliferation rate was seen in GSTCD knock-down samples (siRNA B and C) relative to untrans or scrambled controls at any time-point with Click-iT® EdU assay (A). Total cell number was noticeably reduced in both GSTCD knock-down samples, with an increasing effect across the time-course and significant reduction evident at 72h compared to scrambled for siRNA B (reduced by 33% $p<0.001$, represented by ***) and siRNA C (reduced by 48% $p<0.0001$, represented by ****) (B). Key: Untrans - untransfected control, Scrambled - transfection with scrambled siRNA control, siRNA B/C - transfection with GSTCD knock-down siRNA. Data show mean values from n=3 independent experiments, with individual conditions completed in quadruplicate. One-way ANOVA was used to assess significance between the three experimental conditions investigated at each time-point (scrambled siRNA, GSTCD siRNA B and GSTCD siRNA C).

To account for possible hBEC donor specific effects in these results, the CyQUANT® NF assay was repeated in a second hBEC donor. This replicated the results from donor 1, showing reduction of total cell number in GSTCD knock-down samples compared to scrambled controls over time, with significant reduction primarily evident at 72h (reduced by 51% $p<0.05$ (siRNA B) and 128% $p<0.001$ (siRNA C)) (Figure 20).

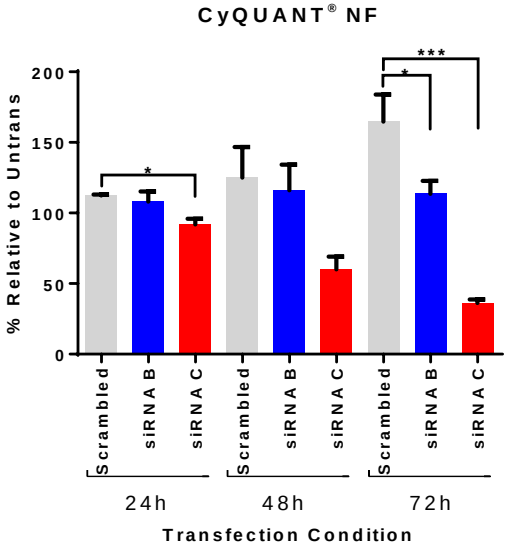



**Figure 20. GSTCD siRNA knock-down in hBECs results in reduced total cell number in a second hBEC donor.** Data is normalised to Untrans controls (100%) bars not shown. Total cell number measured with CyQUANT® assay was similarly reduced in both GSTCD knock-down conditions as previously demonstrated in a different hBEC donor, with an increasing effect across the time-course. Total cell number was significantly reduced compared to scrambled control for siRNA C at 24h (reduced by 20% $p<0.05$, represented by *) and at 72h for siRNA B (reduced by 51% $p<0.05$, represented by *) and siRNA C (reduced by 128% $p<0.001$represented by ***). Key: Untrans - untransfected control cells, Scrambled - transfection with scrambled siRNA control, siRNA B/C - transfection with GSTCD knock-down siRNA. Data show mean values from n=3 independent experiments, with individual conditions completed in quadruplicate. One-way ANOVA was used to assess significance between the three experimental conditions investigated at each time-point (scrambled siRNA, GSTCD siRNA B and GSTCD siRNA C).

Concluding this section, based on evidence of reliable reduction of GSTCD at the mRNA and protein levels by siRNA mediated knock-down in primary human cells, these results suggest a possible role for GSTCD in cell growth when assessing total cell number using the CyQUANT® assay. However when comparing these two measures of cell growth (Click-iT® EdU and CyQUANT® NF assays), the data do not necessarily indicate a role in cell proliferation for GSTCD, instead potentially suggesting a role in cell viability.

### 4.1.5 Discussion

There is currently little published research directly characterising the role of GSTCD in cell functions including cell homeostasis. Previous knowledge has indicated that GSTCD belongs to the GST family of enzymes involved in cellular detoxification, based on the presence of a shared C-terminal alpha helical domain which is characteristic of GST enzymes. Yet GSTCD has not indicated GST functional activity therefore this work has focussed on investigating the potential role of GSTCD in the respiratory system, specifically in hBECs due to previous evidence of GSTCD localisation in this cell type (Obeidat et al. 2013).

Since the function of GSTCD remains unknown and cell growth is an essential homeostasis requirement in the lungs, particularly important in the bronchial epithelium, this was selected for investigation to determine whether GSTCD has a role in maintenance of the cell. This was explored by studying the effects of GSTCD siRNA knock-down on cell growth in hBECs. The aim of this work was to investigate the effects on cell growth when GSTCD levels are reduced in an *in vitro* lung model in human cells.

#### *4.1.5.1 GSTCD Antibody Optimisation and Confirmation of siRNA Knock-down*

siRNA mediated reduction of GSTCD expression has previously been confirmed at the mRNA level by Q-PCR, however changes in mRNA expression levels do not necessarily correspond when translated into protein. Therefore, two GSTCD antibodies (Proteintech and Abnova) were trialled in Western blotting for optimal identification of GSTCD protein, for use in confirming reduced GSTCD protein levels in siRNA knock-down studies and later in mouse lung tissue localisation experiments.

Both antibodies identified the positive control recombinant GSTCD, although at a slightly increased molecular weight than expected. These results suggest that this protein may not be migrating through the SDS-PAGE gel optimally, due to the

presence of a protein purification tag used for production of this protein (glutathione S-transferase (GST)). Alternatively, the molecular weight of the recombinant protein could be marginally larger than the exact sum of GSTCD (97.4kDa) and the GST tag (26.4kDa). Since the recombinant GSTCD includes a protein purification tag, it is feasible to predict that an additional linker sequence between the gene and tag is also present. Therefore, this could be adding to the overall molecular weight. Also, the increased molecular weight seen here appears unlikely to be a result of altered processing during production of the recombinant GSTCD protein, as can occur with proteins produced in bacteria. In this instance the recombinant protein was produced using eukaryotic translation apparatus in a cell free wheat germ expression system (Abnova Technologies 2015). In addition, two different GSTCD antibodies were capable of detecting the recombinant GSTCD; hence it is unlikely that presence of the GST tag resulted in mis-folding of the GSTCD protein.

Furthermore, despite the presence of a GST tag in this recombinant protein (which could possibly be deemed inappropriate based on the given name of GSTCD) this was selected as a positive control as it was the only available option including the full length GSTCD protein sequence. Including the full length sequence was considered to be of higher importance because the recombinant protein could then be used as a positive control for multiple GSTCD antibodies which may have varied immunogen sequences against GSTCD (for example Proteintech and Abnova). Also, as previously indicated, GSTCD is unlikely to have typical GST enzyme function and there is no significant sequence alignment between GSTCD and the GST tag protein sequences (Appendix section 6.2.2). Therefore it can be considered appropriate to utilise this recombinant protein, as the antibodies are unlikely to be able to bind to the GST tag to produce a false positive signal.

Moreover, there is an option available for removing the GST tag from this recombinant GSTCD, via the PreScission Protease method (GE Healthcare). The PreScission Protease specifically cleaves between the Gln and Gly residues of the recognition sequence Leu-Glu-Val-Leu-Phe-Gln/Gly-Pro. However, it was not deemed necessary to remove the GST tag for this work as it cannot be guaranteed

that all tags will be removed from the recombinant samples and this method involves utilising a genetically engineered fusion protein which is also labelled with GST.

Both of the GSTCD antibodies trialled here appeared to identify single bands in a hBEC sample. Although, the approximate molecular weight of the band produced by the Abnova antibody appeared to be higher than expected for GSTCD, hence the Proteintech antibody was selected for further use in GSTCD investigations. This was surprising as the Proteintech antibody immunogen is shorter (350aa) than the Abnova antibody immunogen (full length GSTCD sequence, 633aa), hence could be thought of as less specific. The Proteintech antibody definitively revealed correct identification of GSTCD by also detecting significant reductions in protein levels in GSTCD siRNA treated hBECs relative to controls. Nevertheless, the possibility remains that the Abnova antibody detects GSTCD protein which potentially has post-translational modifications, such as phosphorylation, which increase the overall molecular weight of the protein. For instance, post-translational modifications influencing protein activity have been identified in human tissue for several cytosolic GST enzymes such as phosphorylation (GSTM5 and GST pi 1) and acetylation (GST alpha 1, GST alpha 2, GSTM3 and GSTM4) (Ijiri et al. 2011; Ranganathan et al. 2005; Okamura et al. 2009; Rowe et al. 1997).

Alternatively, as GSTCD also contains a predicted methyltransferase domain modifications to these enzymes were explored. Methyltransferases have roles in a diverse range of biological processes including interacting with DNA, RNA and proteins, transferring a methyl group to one or more of these targets (Struck et al. 2012; Petrossian & Clarke 2011). There is evidence that methyltransferase activity is tightly regulated within the cell cycle in humans by a variety of post-translational modifications, including modifications to: DNA methyltransferase 1 by acetylation and ubiquitination (Scott et al. 2014), SET domain bifurcated 1 histone methyltransferase by phosphorylation (Takada et al. 2007) and glycine N-methyltransferase by acetylation (Luka et al. 2008). Therefore, further investigation into GSTCD modifications and potentially interactions (such as dimerization) may be necessary to fully elucidate its molecular weight. In addition it may be useful in future work to confirm whether the Abnova antibody

is similarly capable of identifying the specific siRNA mediated reduction of GSTCD protein.

In summary, this work demonstrates that siRNA mediated knock-down of GSTCD in hBECs effectively reduces GSTCD expression at both the mRNA and protein levels, hence this is a valuable method to study effects of minimal GSTCD levels in a human *in vitro* model.

#### *4.1.5.2 Cell Growth Assays*

The ability to quantify differences in cell growth sensitively in hBECs when cells are grown in minimal and optimal media was used to compare several cell growth assays. This identified Click-iT® EdU and CyQUANT® NF assays as the two assays most adept at distinguishing a difference in cell growth between the two media conditions at 24h. Whilst initial assay comparisons were performed using a 24h incubation time, it is possible that this time-point could be interpreted as too early to observe a significant difference in cell growth between the media conditions. Although, assays capable of detecting a notable difference at this time-point would appear to be more sensitive to small changes in cell growth, which is of benefit when comparing different treatment conditions (such as with siRNA knock-down). When these assays were compared across a time-course both confirmed defined increases in proliferating cells (Click-iT® EdU) and total cell number (CyQUANT® NF) for cells grown in optimal media when compared against minimal media, demonstrating suitable sensitivity for measuring limited/optimal cell growth. In particular, CyQUANT® NF results reached significance at all time-points, thus potentially indicating better assay reproducibility between wells and replicates, as perhaps expected with a more simple protocol.

Of the assays not selected for continued experiments, MTT remained the primary option for detecting minor changes in cell growth. Cell counting and CCK-8 assays were not able to detect notable differences in minimal and optimal media with a limitation of cell counting potentially being the protocol components. It can be

considered difficult to accurately quantify cell numbers using cell counting due to trypsin being used to remove cells from the culture plastic. Inevitably, not all cells will be removed for counting using this method and some may be lysed in the process. This assay was included for comparison as this is a traditional measure of cultured cells. Moreover, MTT and CCK-8 assays are considered measures of cell viability based upon their mechanism of action; measuring metabolic activity via dehydrogenase enzymes, produced only by living cells. Measuring cell viability does not necessarily correspond to total cell numbers, hence changes in cell growth can be difficult to accurately determine with these methods.

Focussing on the Click-iT® EdU and CyQUANT® NF assays provides many benefits (Table 16). For instance the Click-iT® EdU assay has several advantages over previous assays similarly based upon labelling of DNA, such as tritiated thymidine or BrdU method, whereby it eliminates the requirement for radioactivity (safer and more simple protocol) and the harsh denaturation step (able to retain the integrity of DNA) necessary in the respective previous assays. This labelling method has been demonstrated to be highly reproducible and the reduced size of fluorescent azides, compared to monoclonal antibodies used previously, provides improved access to DNA correlating with increased sensitivity (Poujol et al. 2014; Choi & Berdis 2015). Also, the Click-iT® technology is compatible with additional experiments to utilise the same samples, for example immunohistochemistry and immunocytochemistry.

Nevertheless, a point of note here is the high levels of background fluorescence detected in vehicle controls using the Click-iT® EdU assay. With the basis of this assay utilising a thymidine analogue, the high measures in the control wells may be attributable to non-specific signal amplification, particularly when regarding the multi-step protocol. Possible future preventative measures not suggested in the protocol may include: permeabilising cells using 0.1% Triton X-100 (allowing better access to intracellular labelled DNA), blocking endogenous peroxidase activity with 1% hydrogen peroxide prior to cell fixation (limit interference when utilising HRP labelling) and/or incubating the HRP conjugated anti-Oregon green antibody overnight at 4°C for increasing specificity of binding (Haines et al. 2010). However, the Click-iT® EdU assay remained a top choice for measuring

cell growth due to its high level of sensitivity for measuring actively proliferating cells, as also demonstrated previously (Yokota et al. 2012; Petrilli et al. 2014; Salic & Mitchison 2008; Xia et al. 2010; Morrison et al. 2014; Lentz et al. 2009; Feng et al. 2011). It is currently unknown whether other groups have also investigated vehicle controls when using this assay.

On the other hand, CyQUANT® NF assay appears to be a very basic protocol in comparison to the Click-iT® EdU method, with single step application of a fluorescent DNA binding protein and incubation at 37°C. Moreover, this assay advances on a previous CyQUANT® assay by eliminating a compulsory freezing step (Jones et al. 2001) thus avoiding compromising cell structure. CyQUANT® NF boasts many other benefits including: a linear correlation between fluorescence readings and cell number (100-20,000 cells), no washes required and can be used with antibody- or dye-based assays in flow or high content screening (Ignatius 2009).

Also, both assays are compatible with a variety of cell culture plates (well sizes) and, in theory, allow collection of several experimental samples at varying time-points (either stored as fixed cells at 4°C (Click-iT® EdU) or by freezing at -70°C (CyQUANT® NF)) for assay completion to be performed simultaneously, supporting reproducibility between replicates.

Overall, these results indicate that Click-iT® EdU and CyQUANT® NF assays are a good choice for future studies investigating cell growth as both assays sensitively detect expected differences in cell growth between minimal and optimal media. Importantly, these assays also avoid effects on cell number determination influenced by variations in metabolic activity, as per MTT and CCK-8 assays.

**Table 16. Summary comparison of Click-iT® EdU and CyQUANT® NF assays outlining the purpose and advantages of each assay.**

| Application | Click-iT® EdU Assay | CyQUANT® NF Assay |
|---|---|---|
| **Answers the Question:** | Which cells are actively proliferating? | Have the cells multiplied? How many cells are present? |
| **Primary Use** | For direct and accurate detection of new DNA synthesis; measure of cell proliferation | Cell counting via DNA content to assess cytotoxic or proliferative effects independent of cellular metabolism |
| **Advantages** | ▪ Easy EdU protocol replaces difficult BrdU assays<br>▪ Can be used with: imaging, antibody based assays, high content screening, flow cytometry<br>▪ Antigenicity and cell structure preserved<br>▪ Use on fixed cells, no dependence on metabolic activity<br>▪ Does not require cell lysis, long incubations, radioactivity, or removal of stain from cells | ▪ No wash assay<br>▪ Highly accurate, sensitive, and reproducible<br>▪ Short protocol (1 hour)<br>▪ Can be used with antibody- or dye-based assays in flow or high content screening<br>▪ Does not require cell lysis, long incubations, radioactivity, or removal of stain from cells<br>▪ Cells can be frozen and stored prior to assaying<br>▪ No dependence on metabolic activity |
| **Method of Detection** | Measures rate of new DNA synthesis by EdU nucleoside analogue incorporation into DNA with detection by copper catalysed “click” reaction | Quantitates relative number of cells in a population based on total DNA content by measuring DNA binding dye |

#### *4.1.5.3 GSTCD siRNA Knock-Down Results in Reduced Cell Growth*

The two assays most adept at distinguishing a difference in hBEC growth between minimal and optimal media (Click-iT® EdU and CyQUANT® NF) were selected to examine minimal GSTCD expression effects in a siRNA knock-down model. Specifically, an advantage of comparing these two assays is that they assess two different measures of cell growth; actively proliferating cells and total cell number, respectively.

When cell growth was studied utilising these assays in GSTCD siRNA knock-down treated hBECs, no discernible changes were observed in actively proliferating cells suggesting that GSTCD does not have a role in initiating or maintaining cell division. Conversely, total cell number was decreased in both GSTCD knock-down conditions relative to scrambled siRNA controls across two hBEC donors. Although, a possible donor specific effect apparent from these analyses could be that scrambled siRNA transfection appeared to result in varied effects in different donors compared to untransfected conditions. siRNA intersects with a number of other pathways, hence non-specific effects can be triggered despite accounting for balanced nucleotide composition in scrambled siRNA with the gene targeting siRNAs and randomizing the sequence to avoid targeting other genes. It is currently unknown whether the variability between donors seen here is due to the scrambled siRNA or the transfection reagent INTERFERin®. Overall, transfection with the scrambled siRNA is the true control in this experiment as this condition also involves addition of RNA to the cells, replicating exact conditions used for GSTCD siRNA knock-down. Notably, both donors show significant results relative to this condition, therefore it remains highly probable that there is a real effect on total cell number in GSTCD knock-down treated hBECs.

Furthermore, since two different siRNA sequences were used to knock-down GSTCD, which produced similar reductions in total cell numbers, it can be concluded that the reduction is a true result and unlikely to be a consequence of off-target effects from a specific siRNA. Interestingly, siRNA C appears to have a stronger effect on reduction of total cell number, corresponding to data where

siRNA C transfection results in a stronger level of GSTCD knock-down at both the mRNA and protein levels compared to siRNA B. It is possible that GSTCD knock-down is more efficient in this instance due to siRNA C binding to GSTCD transcripts further upstream than siRNA B (exon 3 compared to exon boundary 8/9) (outlined in Appendix section 6.2.3). As a result siRNA C may be more effective at limiting successful translation of transcripts and/or may be more likely to abolish translation of a greater number of potential transcript variants, due to its more N-terminal location.

Since the function and signalling pathways for GSTCD are currently theoretical and based on identification of proteins with shared homology, it is pertinent to investigate whether predicted homologous proteins have similar effects on cell growth and viability when expression is reduced. Firstly, GSTCD was initially associated with the cytosolic GST family of enzymes (alpha, mu, pi, sigma, theta, zeta, omega and kappa) due to a shared conserved domain in the protein sequence. GST enzymes are typically involved in cell defence by cellular detoxification; conjugating glutathione to a variety of targets. Theses enzymes have been shown to influence susceptibility to human diseases linked to oxidative stress, such as cardiovascular diseases, asthma and carcinogenesis (Korashy & El-Kadi 2006). Few studies have so far directly investigated the effects of reduction of GST expression individually in the context of cell viability, although, GST pi 1 siRNA knock-down was found to have no effect on the viability of A549 cells at 72h (Wang et al. 2011). Furthermore, siRNA mediated down-regulation of GST alpha 1 was observed to significantly increase cells in S-phase and for associated cell proliferation (Adnan 2012).

On the other hand, a second conserved domain has also been identified in the GSTCD protein sequence, conferring potential methyltransferase activity. Investigating siRNA knock-down of expression in methyltransferases frequently results in reduction of cell viability in human *in vitro* studies. For example DNA methyltransferase 1 has an important role in maintaining methylation patterns during DNA replication and siRNA knock-down results in inhibition of proliferation and induction of apoptosis (Brown & Robertson 2007; Xu et al. 2010; Li et al. 2014). Additionally, reduction of cap methyltransferase expression,

involved in capping mRNA transcripts, results in loss of cell viability (Shafer et al. 2005). However, it could be suggested that examples such as these would evidently affect cell growth and viability, due to involvement of these methyltransferases in processes critical to functioning of the cell (such as by directly influencing gene expression by modifying DNA and mRNA). Therefore, further evidence that reduction of expression of a variety of methyltransferases results in decreased cell viability is demonstrated by Majumder et al. The group demonstrate that the consequences of siRNA knock-down of coactivator-associated arginine methyltransferase (catalyses methylation of histone H3) are inhibition of proliferation and induction of apoptosis (Majumder et al. 2006). Moreover, similar effects on cell viability are also shown for the three proteins so far predicted to potentially interact with GSTCD (Table 17).

Therefore, summarising effects on cell growth and viability in GSTCD homologous proteins, similarities appear to be primarily evident with methyltransferases. Although, these effects are not necessarily apparent in GSTCD knock-out mice, as homozygote knock-out mice are viable and grow normally as discussed in the proceeding section.

**Table 17. Proteins suggested to have binary interactions with GSTCD (Weimann et al. 2013) show similar reduction in cell viability in RNA interference expression knock-down studies.**

| Methyltransferase | Activity | Expression Knock-down Effects | References |
|---|---|---|---|
| **KDM1A** | Modifies histone H3 by demethlyation | - Significant reduction in cell viability (72h)<br>- Strong reduction in number of proliferating cells<br>- Phenotypic changes predominantly due to induction of apoptosis | Pajtler et al. 2013<br>Althoff et al. 2013<br>Yang et al. 2014 |
| **PRMT6** | Modifies proteins by methylation | - Considerable reduction of cell proliferation<br>- Significant suppression of cell growth | Stein et al. 2012<br>Harrison et al. 2010<br>Yoshimatsu et al. 2011 |
| **SUV39H2** | Modifies histone H3 by methylation | - Severely impairs cell viability (combined gene expression disruption with SUV39H1)<br>- Induces chromosomal instabilities | Peters et al. 2001 |

In summary, reduction of GSTCD function did not appear to affect actively proliferating cells (as has been shown in knock-down experiments for many typical methyltransferases) but did result in reduced total cell numbers (reflecting typical methyltransferase knock-down results). By comparing all assay results for human GSTCD siRNA knock-down this data could be suggesting a potential role for GSTCD in apoptosis, since total cell number appears to be reduced despite actively proliferating cells seeming to be unaffected. Whilst care was taken to avoid disrupting cells, another possibility could be that the GSTCD knock-down treated cells may have resulted in impaired adhesion; hence some cells may have been lost upon media removal prior to labelling protocols. Therefore whilst these results provide a novel insight into the consequences of GSTCD knock-down, the specific reason behind the reduction in cell number seen here is currently unclear.

Overall, this remains an important and novel development for the lung function and COPD associated gene GSTCD, providing interesting avenues for future exploration.

## 4.2 Knock-out of *GSTCD* Does Not Result in Gross Phenotypic Changes in Mice

Detailed knowledge of GSTCD function is yet to be elucidated. A loss of function model was initially investigated in human cells *in vitro* by siRNA mediated reduction of GSTCD expression in hBECs to investigate a potential role in cell growth. However, this model does not represent a complete loss of function. Therefore Gstcd knock-out mice were generated with the aim of further investigating the phenotypes which may be affected by loss of this protein *in vivo*. Specific focus was on cell growth in the form of lung structure, due to the association identified in GSTCD expression knock-down in hBECs shown in Results section 4.1.

Since phenotypic associations with SNPs in *GSTCD* have focussed around the respiratory system, lung structure and morphology was investigated in Gstcd -/- and Gstcd +/+ littermate mice to examine gross changes in the lungs. Sections were obtained from Gstcd knock-out mice, generated at Cornell University (Michael Kotlikoff group). Lung function measure associations with SNPs located at locus 4q24 and in particular with GSTCD have been identified in adult lungs and varied expression levels have also been demonstrated across human lung development. Specifically, GSTCD protein expression was found to increase during the pseudoglandular stage (branching morphogenesis and epithelial and mesenchymal differentiation occur) and begin decreasing in the proceeding canalicular stage (for example when alveolar epithelial type II cells differentiate) (Obeidat et al. 2013). Changes in the levels of GSTCD mRNA also concur with these results at both developmental stages (Obeidat et al. 2013) suggesting a role for GSTCD in lung development. Therefore, morphology of Gstcd +/+, +/- and -/- lungs were compared at embryonic day (E) E14.5, E18.5 and at adult (10 weeks) stages.

### 4.2.1 Gstcd Knock-out Confirmation

Investigation into a Gstcd knock-out mouse was established in collaboration with Cornell University which generated the mice using an ESC line created within the IMPC. The colony was successfully expanded and preliminary analysis performed, by reverse transcriptase PCR revealing absence of Gstcd mRNA (Appendix section 6.4.1 Figure 33). Notably, lacZ staining was present in lung tissue in homozygotes and heterozygotes, indicating that Gstcd expression has been replaced by the lacZ reporter tag from the Gstcd knock-out construct (Figure 21).

As a component of the Gstcd knock-out construct, the lacZ reporter tag was inserted into Gstcd exons 4-5, resulting in deletion of these exons and initiation of lacZ expression independently of the Gstcd promoter. LacZ staining for the Gstcd knock-out construct revealed similar localisation to Gstcd IHC staining in mouse lung tissue (Results section 4.2.2), in cells in the bronchial epithelium and in specific cells in the alveolar tissue.

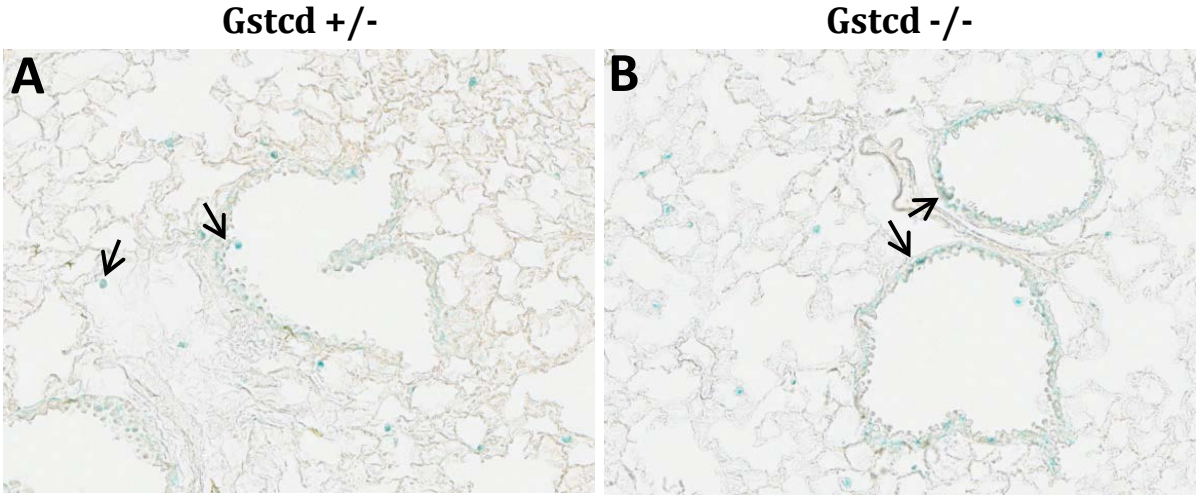



**Figure 21. LacZ staining of lung tissue provides localisation information in Gstcd +/- and Gstcd -/- transgenic mice where Gstcd expression is replaced by LacZ (Cornell University, Frank Lee).** In mouse lung tissue expression of the lacZ reporter tag from the Gstcd knock-out construct indicates localisation in the bronchial epithelium and in specific cells in the alveolar tissue (indicated by arrows), for Gstcd +/- (A) and Gstcd -/- (B) knock-out mice. Images are shown at x20 magnification.

### 4.2.2 Gstcd Localises in the Bronchial Epithelium in Mouse Lung Tissue

Previous work from our group has shown that GSTCD localises to a number of different regions within human adult lung, including bronchial epithelial cells and selected cells in the alveolar tissue (Obeidat et al. 2013). The cellular localisation of Gstcd was investigated in mouse lung tissue to determine whether the same localisation pattern is present in this model organism. This was performed using the previously optimised Proteintech GSTCD antibody (Results section 4.1.1) in IHC.

Figure 22 appears to show positive Gstcd staining in bronchial epithelial cells and in a sub-set of cells within the alveolar tissue (Figure 22E and F, indicated by arrows).

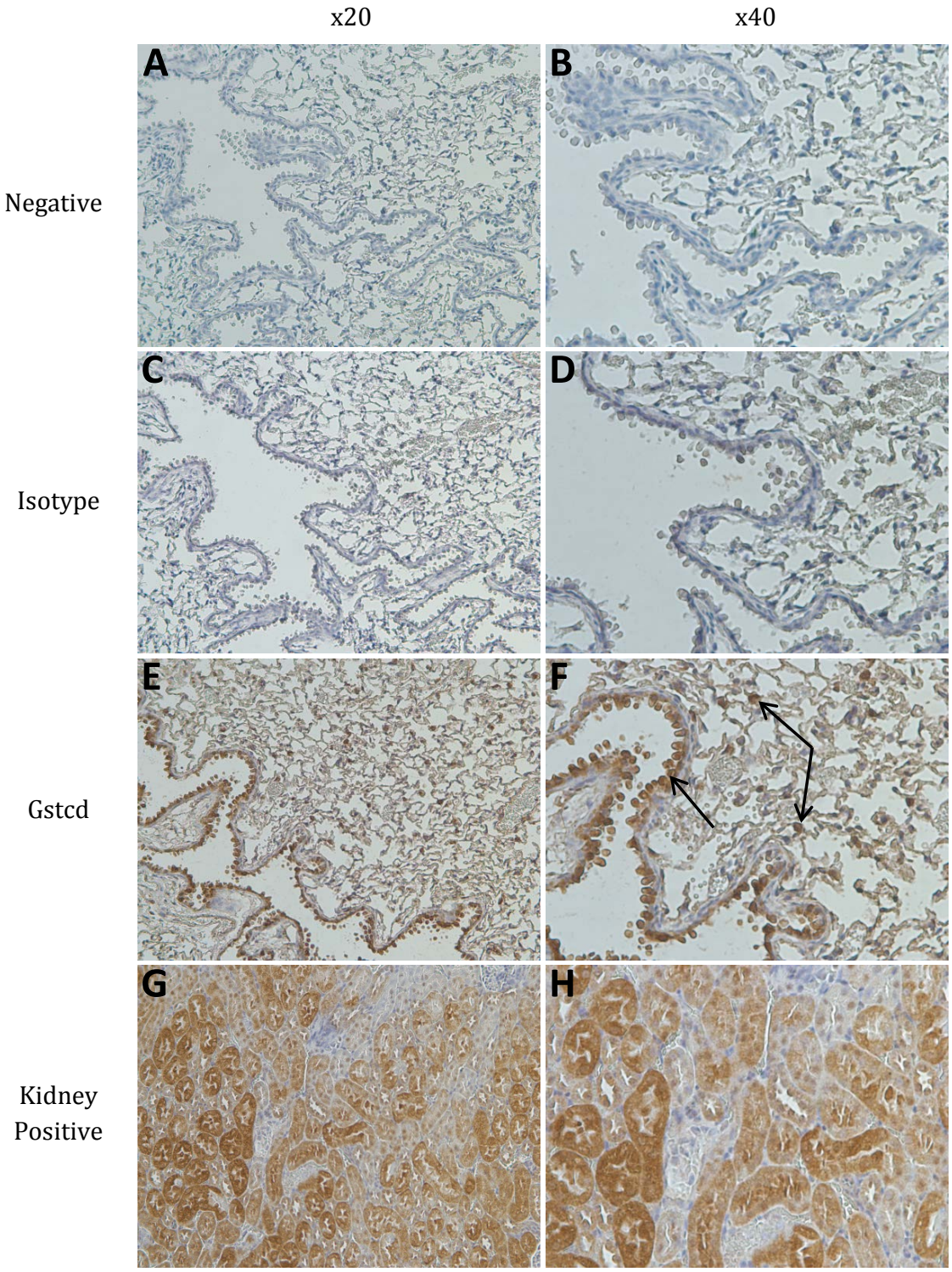



**Figure 22. Gstcd protein expression indicates bronchial epithelial localisation as assessed by IHC in Gstcd +/+ C57BL/6N mouse lung tissue.** Representative images shown of Gstcd expression with negative and isotype controls in lung tissue and positive control performed in kidney tissue. No staining is evident in either negative control (A-D). Predominant Gstcd staining in the lung indicates bronchial epithelial specificity in lung tissue and specific cells across the alveolar tissue (E, F, arrows indicate localised staining) and predicted positive control staining is seen in kidney tissue (G, H). Sections were stained with Proteintech antibody at 1:50 and images shown at x20 (A, C, E, G) and x40 (B, D, F, H) magnification. Analysis was performed on sections of paraffin embedded adult lung tissue age matched at 10 weeks. Images are representative of n=3 independent experiments.

### 4.2.3 Gross Lung Structure Does Not Appear Morphologically Altered in Adult Gstcd Knock-out Mice

As little is currently known about the role of GSTCD in the airways including the contribution to normal lung development and homeostasis, knock-out mice were utilised in this research to examine phenotypic effects caused by loss of Gstcd protein, with specific focus on the lungs.

During expansion of the Gstcd knock-out mouse colony, Gstcd -/- mice were observed to reach adulthood and breed normally. A necropsy analysis investigating major organs, bones and glands in two male and female mice also found no overt differences between Gstcd +/+ and Gstcd -/- mice (Appendix section 6.4.2 Table 25)

To study lung morphology in Gstcd knock-out mice, histological analysis was performed following haematoxylin and eosin (H&E) staining. Comparison of H&E stained Gstcd +/+ and Gstcd -/- adult lungs revealed no notable morphological differences between these genotypes (Figure 23). However, regions of damaged tissue were visible across Gstcd +/- sections (Figure 23D and E) and heterogeneous tissue structure was present across individual lung sections (Figure 23A-B, D-E and G-H). The heterogeneity and appearance of the tissue indicated possible variations in degree of lung inflation influencing how stretched the airways and alveoli become. These histological variations preclude the ability to quantify possible differences between genotypes hence this has not yet been explored.

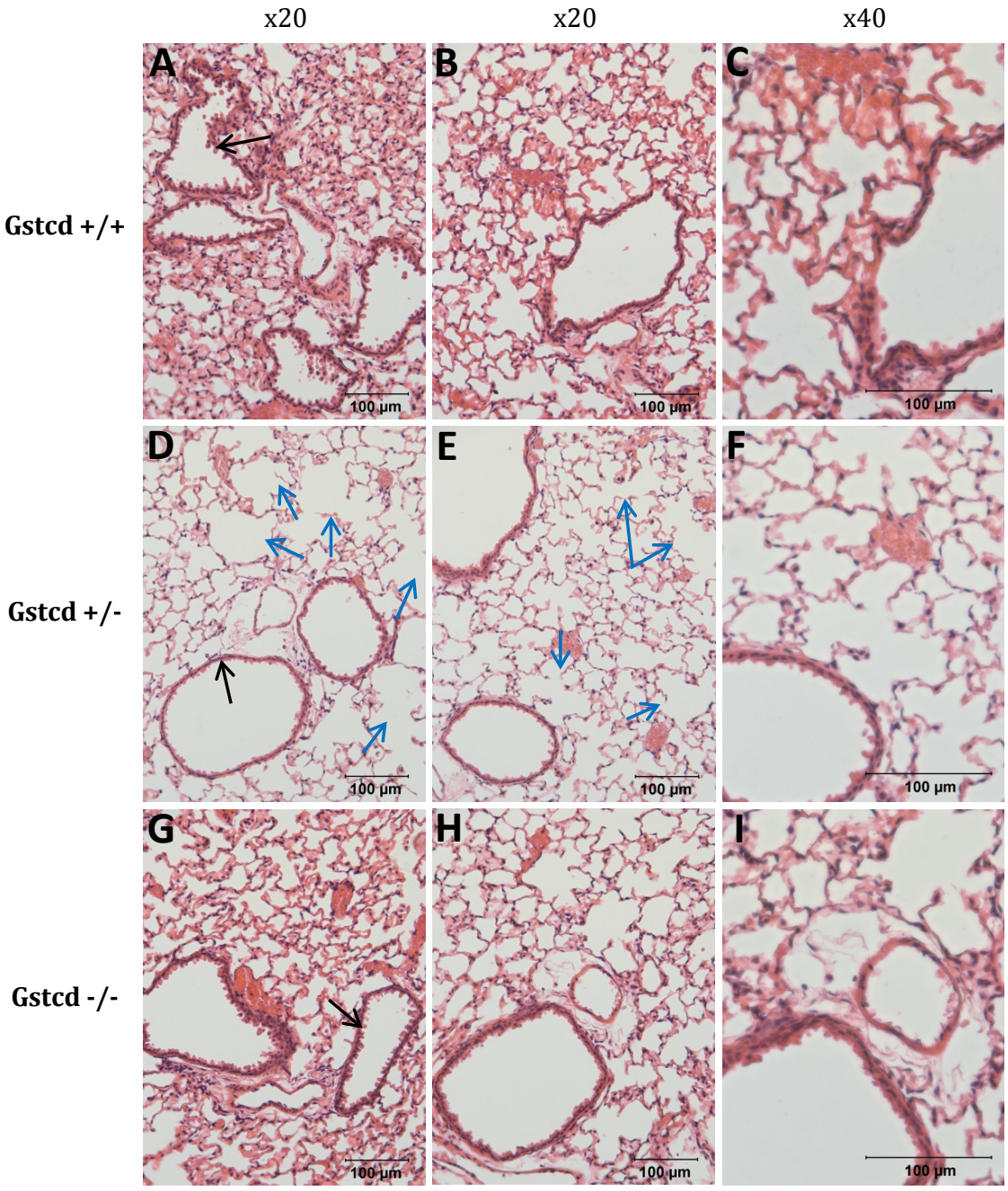



**Figure 23. Loss of Gstcd does not result in gross phenotypic effects in adult mouse lung tissue.** H&E histological analysis of adult Gstcd +/- (D-F, representative of n=1) and Gstcd -/- (G-I, representative of n=3) knock-out mouse lung tissue reveals no notable differences in lung structure relative to Gstcd +/+ (A-C, representative of n=2). Slight variations in bronchial epithelial lining protrusions (A, D, G, indicated by black arrows) and alveolar density are potentially due to variations in inflation (A)) as these occur across lung tissue within individual lung lobe sections. Regions of alveolar tissue damage, potentially a consequence of over-inflation of lungs prior to tissue processing, are also evident (D and E, indicated by blue arrows). Two example images for each genotype are shown at x20 magnification to demonstrate variations in alveolar density in a single lung section and the most representative images for comparison across genotypes are also shown at x40. Analysis was performed on sections of paraffin embedded adult lung tissue littermate and age matched (10 weeks). Images are representative of lung sections cut across each whole lung.

### 4.2.4 Mouse Lung Development Does Not Appear Adversely Affected Morphologically by Gstcd Knock-out

Previous research has demonstrated differential GSTCD gene and protein expression across lung development in humans, peaking in the pseudoglandular stage (Obeidat et al. 2013). Therefore two standard developmental stages were selected to investigate whether loss of Gstcd affects lung development in mice. Development in mice is quantified by measure of days post coitum, developing lungs were assessed at embryonic days 14.5 (pseudoglandular stage) and 18.5 (saccular stage, studying potential alterations that have occurred in the canalicular stage) (E14.5 and E18.5).

Sections from Gstcd +/+, Gstcd +/- and Gstcd -/- mouse lungs at both E14.5 and E18.5 were examined by H&E staining. No structural differences were perceptible between genotypes in lungs from either developmental stage (Figure 24 and Figure 25). Due to low replicate numbers statistical analyses could not be performed.

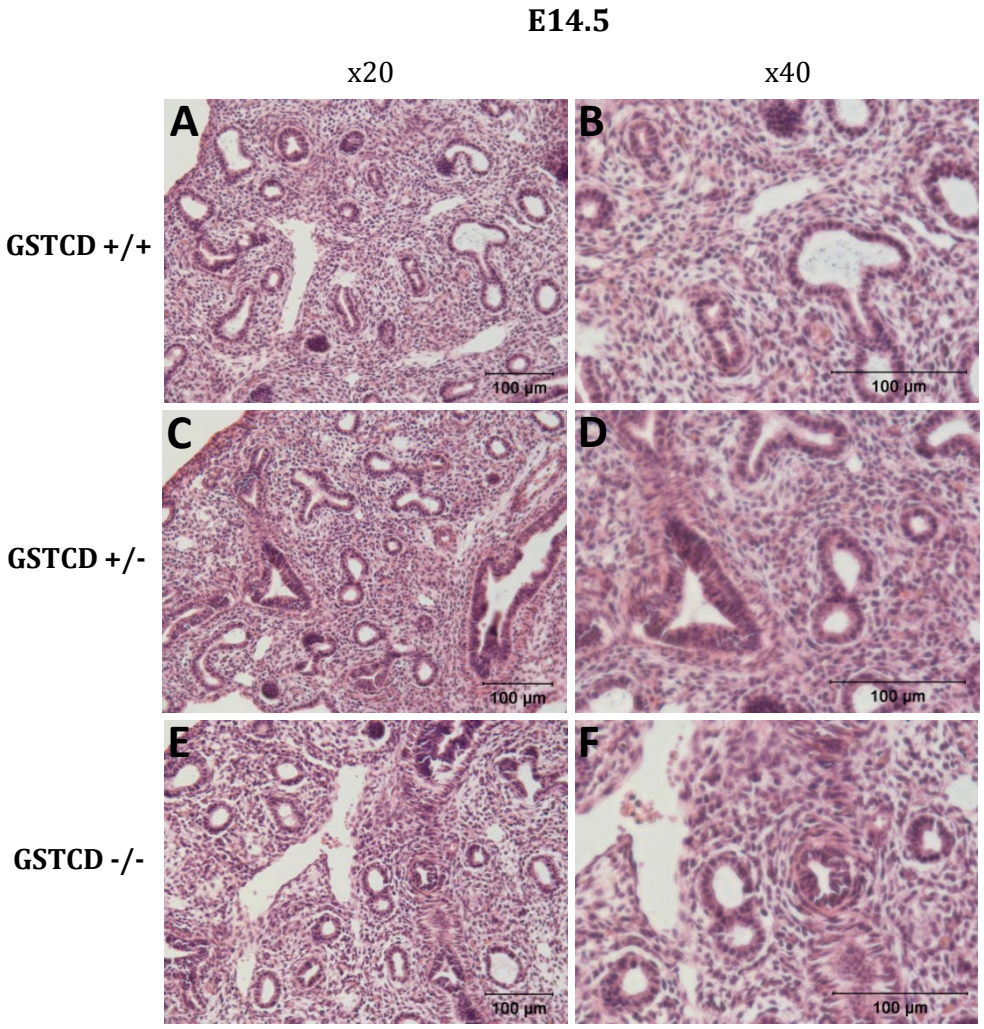



**Figure 24. Preliminary lung histology investigation by H&E staining indicates unaltered structure in Gstcd -/- mice at E14.5.** Representative images for genotypes at the developmental stage E14.5 are outlined as Gstcd +/+ (A and B, representative of n=3), Gstcd +/- (C and D, representative of n=1) and Gstcd -/- (E and F, representative of n=2). Sections show comparable airway structures across genotypes, with no observable differences. Quantitative approaches have not yet been performed due to low sample numbers. Images are shown at x20 and x40 magnification. Analysis was performed on sections of paraffin embedded littermate matched developing mouse lungs at embryonic day 14.5. Images are representative of lung sections cut across each whole lung.

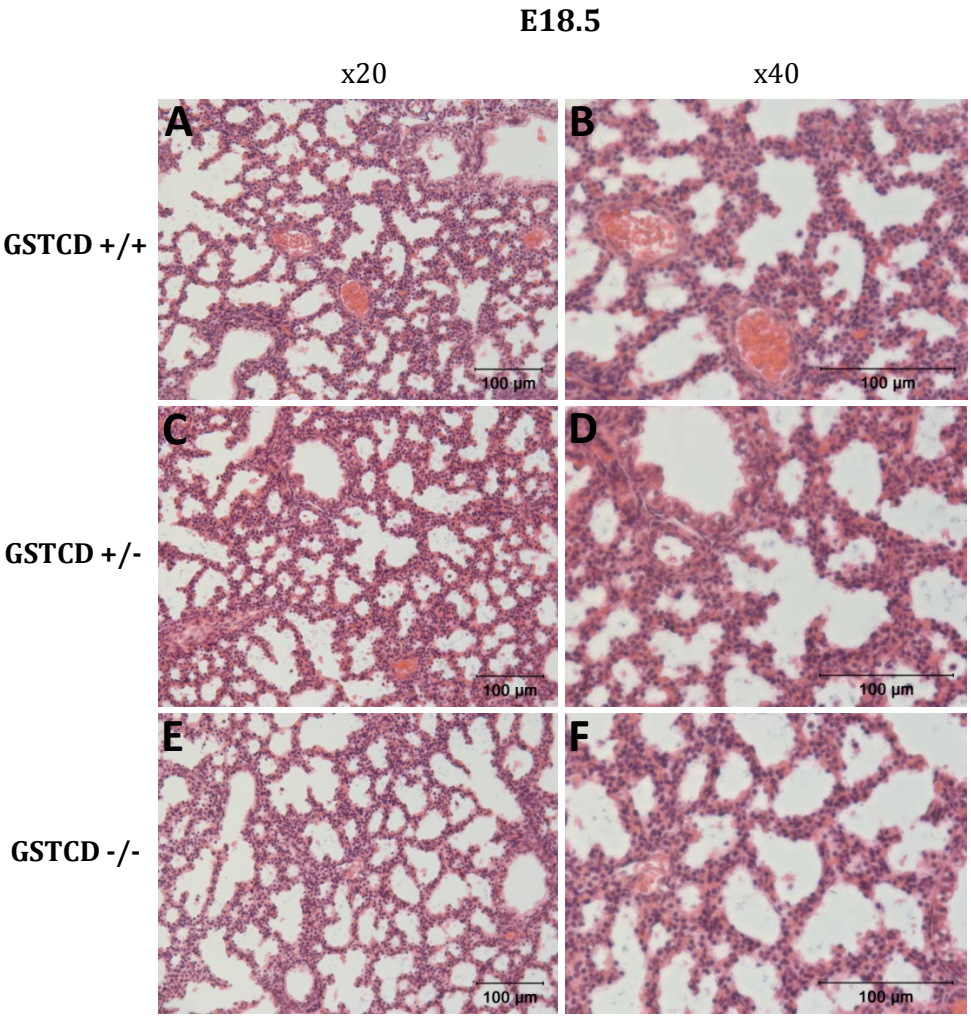



**Figure 25. Preliminary lung histology investigation by H&E staining indicates unaltered structure in Gstcd -/- mice at E18.5.** Representative images for genotypes at the developmental stage E18.5 are outlined as Gstcd +/+ (A and B, representative of n=4), Gstcd +/- (C and D, representative of n=2) and Gstcd -/- (E and F, representative of n=3). Sections show comparable airway structures across genotypes, with no observable differences in alveolar structure or airway distribution are evident. Quantitative approaches have not yet been performed due to low sample numbers. Images are shown at x20 and x40 magnification. Analysis was performed on sections of paraffin embedded littermate matched developing mouse lungs at embryonic day 18.5. Images are representative of lung sections cut across each whole lung.

Whilst gene targeting in primary human cells can be informative as shown in Results section 4.1, gene deletion in the entire mammalian organism has scope to provide novel insight into gene function, particularly with respect to cell growth and homeostasis. With this in mind in addition to morphological differences I also set out to investigate the effect of GSTCD knock-out on homeostasis markers for cell growth.

### 4.2.5 Optimising Antibody Markers of Cell Growth in Mouse Lung Tissue

Since GSTCD knock-down in hBECs has indicated a potential role in cell growth (Results section 4.1), future work may focus on this effect in GSTCD knock-out mouse lung tissue. Initially, two cell growth marker antibodies were selected for optimisation in wild-type mouse tissue in IHC; these were the standard markers of cell division Ki67 and phospho-histone H3 (PH3). Ki67 is an antigen expressed in cells during active phases of the cell cycle (late G1, S, G2 and M phases) including cells which may result in apoptosis prior to cell division. Whereas, PH3 selectively stains for cells only in M phase of the cell cycle; cells which are committed to division.

In this instance, staining was only evident in the positive control tissue for both antibodies (Figure 26), indicating either a very low level of proliferation in the lung tissue at time of fixation or anomalous tissue samples.

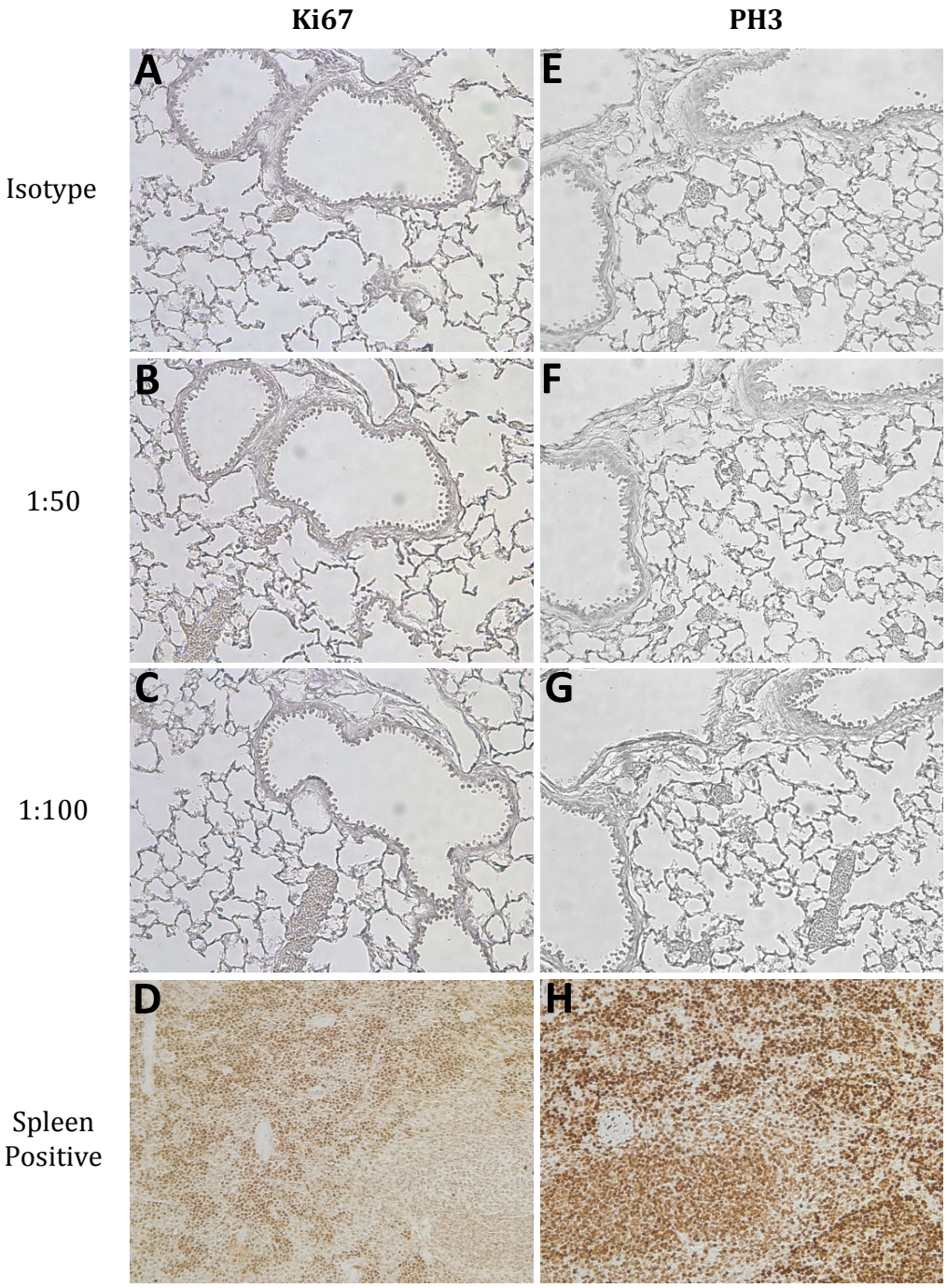



**Figure 26. Proliferating cells were not detectable in wild-type mouse lung tissue using IHC.** Representative images shown from IHC performed with proliferation markers Ki67 and PH3 including isotype controls in lung and positive controls performed in spleen. No staining is evident in either isotype control (A and E) and no cells appeared to be stained for either proliferation antibody at 1:50 or 1:100 dilutions (B, C, F and G). Positive staining as expected is seen in spleen tissue for both antibodies (D and H). Images are shown at x20 magnification. Analysis was performed on sections of paraffin embedded adult lung tissue age matched at 10 weeks. Images are representative of n=3 independent experiments.

### 4.2.6 Gstcd Protein Expression in Gstcd Knock-out Mouse Lung

IHC was performed with the aim of confirming absence and reduction of Gstcd protein in Gstcd -/- and +/- mice respectively, however varied staining was seen indicating Gstcd expression in mouse lung tissue across genotypes (Figure 27). This is unlikely to be a consequence of technique during IHC due to the unstained isotype controls and the presence of an internal negative control, whereby airway smooth muscle does not appear to be stained (indicated by arrows, Figure 27). Staining was also evident in Gstcd -/- kidney tissue, reflecting similar results to Gstcd +/+ kidney.

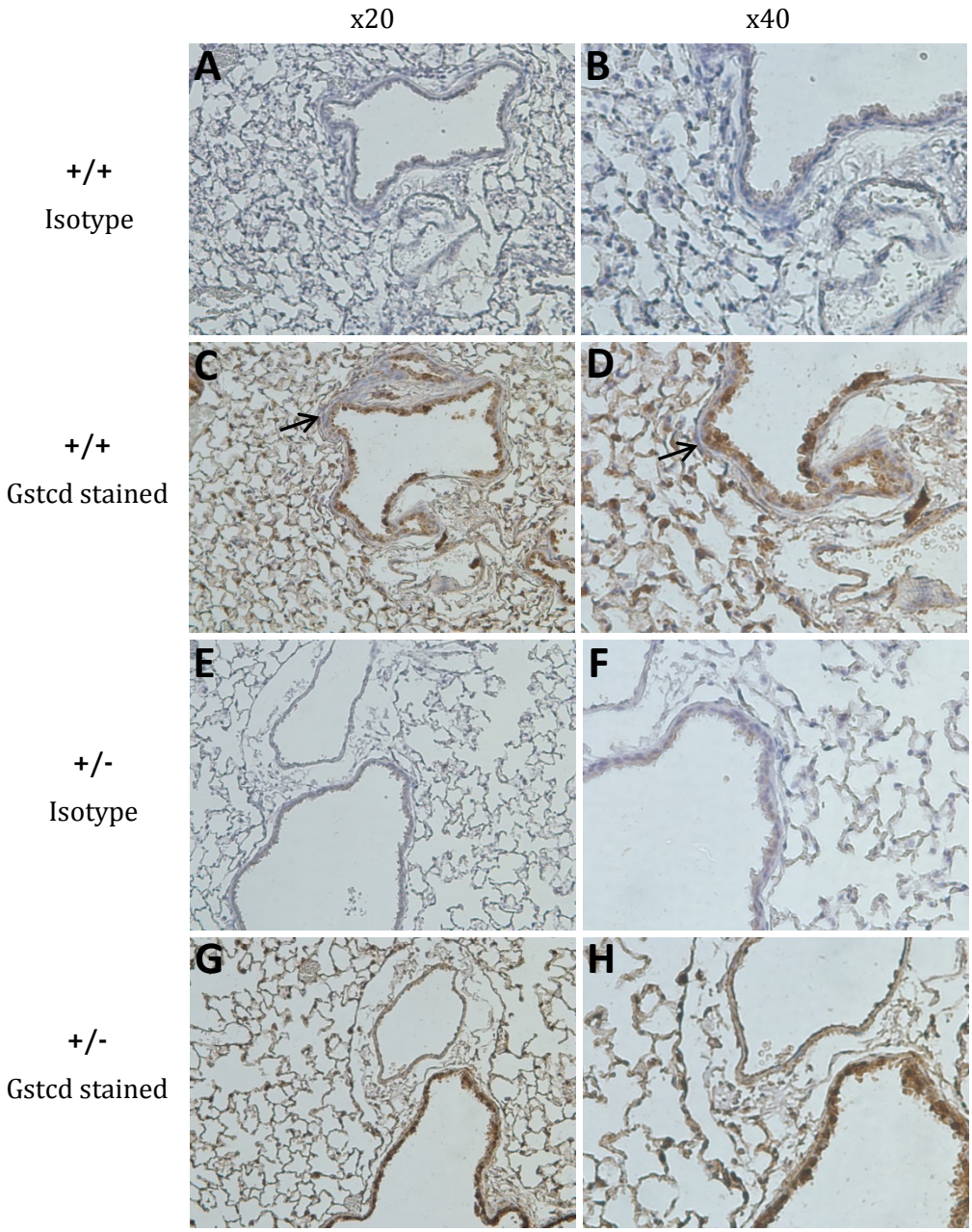



Figure continues on the following page

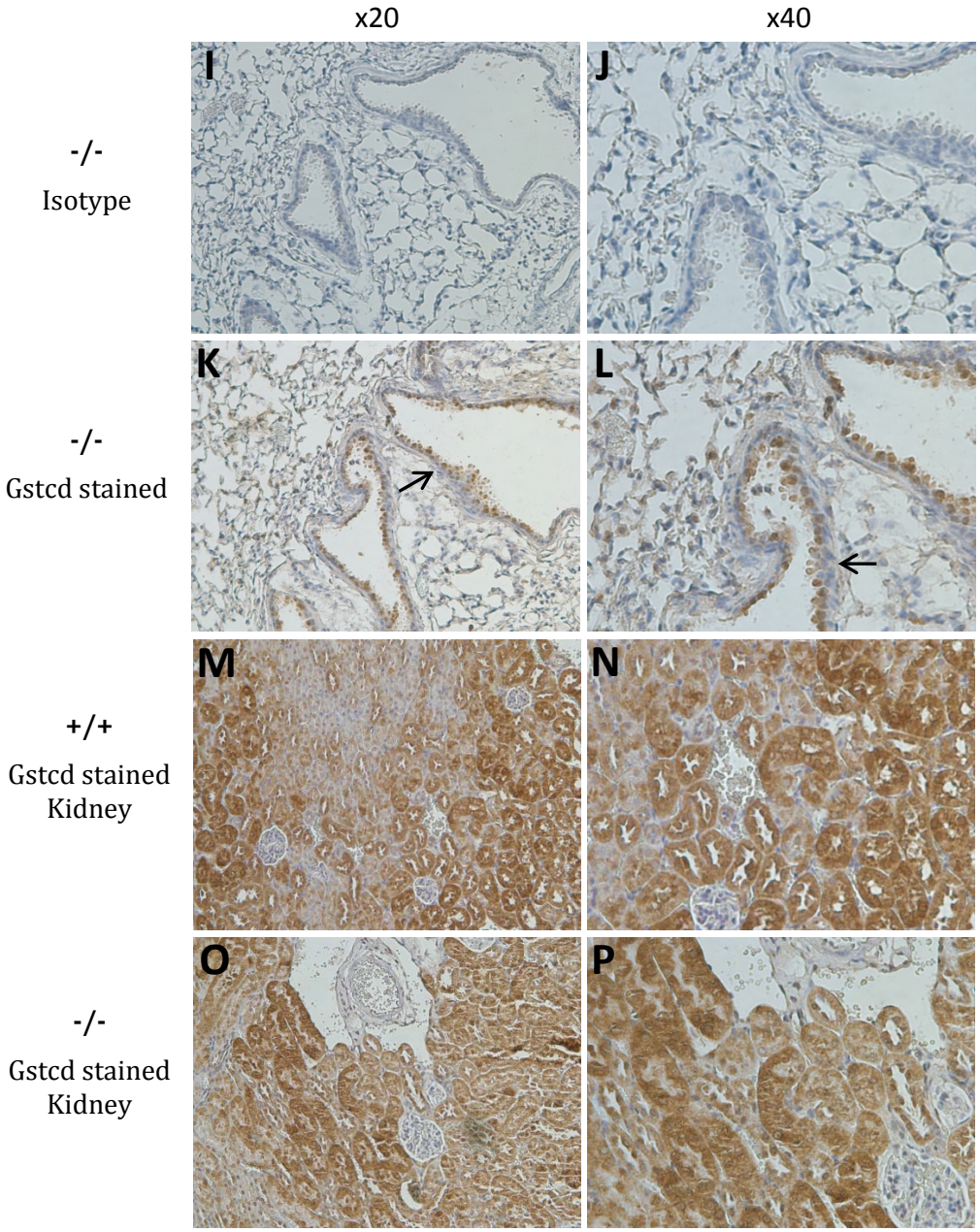



**Figure 27. GSTCD expression was apparent in lung tissue in GSTCD +/+ (n=2), +/- (n=1) and -/- (n=3) lung tissue using IHC.** Representative images shown from IHC performed with GSTCD antibody in mouse lung tissue including isotype controls. Staining is not evident in isotype controls (A, B, E, F, I and J) or in airway smooth muscle (indicated by arrows, C, D, K and L). Gstcd staining appeared to mainly localise in the bronchial epithelium of airways across genotypes (C, D, G, H, K and L) with non-specific staining seen across the +/- lung sections (G and H). Staining is also observed in Gstcd +/+ and -/- kidney sections (M-P). Images are shown at x20 and x40 magnification. Analysis was performed on sections of paraffin embedded adult lung tissue littermate and age matched (10 weeks).

Overall, the results in this section exhibit preliminary exploration of loss of function effects for Gstcd by knock-out in a mammalian model organism. No notable differences have so far been observed for the Gstcd -/- mice compared to Gstcd +/+ either globally or specifically within the lungs. To support this work, effects of loss of function for GSTCD were then compared in a human cell based model, examining GSTCD siRNA knock-down in primary human airway cells (hBECs).

### 4.2.7 Discussion

Due to SNPs within the region of *GSTCD* having been significantly associated with lung function and COPD but no defined function known for this protein, Gstcd knock-out mice were generated to investigate loss of function effects. It is particularly advantageous to explore GSTCD loss of function in a mammalian model organism because studying this in a human *in vitro* model via siRNA knock-down does not result in complete absence of GSTCD expression.

Loss of Gstcd expression in knock-out mice was confirmed by RT-PCR and lacZ staining by collaborators at Cornell University (Michael Kotlikoff group). I aimed to characterise loss of function effects on the lungs in this knock-out mouse model by assessing adult and developmental lung structure between genotypes.

#### *4.2.7.1 Gstcd Localises in the Bronchial Epithelium in Mouse Lung Tissue*

Investigating Gstcd cellular localisation in mouse lung tissue by IHC revealed localisation in the bronchial epithelium and in specific cells throughout the alveolar tissue; possibly indicating a specific role for this protein in the lungs. These results support previous research of GSTCD localisation in human lung tissue (Obeidat et al. 2013), re-emphasising the relevance of using knock-out mice as a model organism. Notably, similar localisation patterns in mouse lung tissue have also been demonstrated by an alternative GSTCD antibody (Appendix section 6.3 Figure 32). In particular, lacZ staining of Gstcd -/- mice indicates a similar localisation to both antibodies, therefore enhancing the likelihood that this localisation pattern is accurate, despite both antibodies being polyclonal. The lacZ staining only occurs at the sites where Gstcd is normally expressed, where Gstcd expression has been interrupted and replaced by transcription of the inserted *lacZ* reporter gene.

The location and appearance of these cells suggests that they could be alveolar epithelial type II cells or macrophages however further co-staining would be required to confirm this. Due to the specific localisation patterns in these results, this further confirms that GSTCD is likely to have an important role in the lungs,

agreeing with previous research demonstrating significant associations with lung function measures and COPD.

#### *4.2.7.2 Lung Structure Does Not Appear Morphologically Altered in Developing or Adult Gstcd Knock-out Mice*

Preliminary investigation of lung morphology revealed no gross changes either during lung development or in adult lungs, concluding that Gstcd does not seem to have a critical role in structural development or growth of the lungs. This is despite many published examples of associations of SNPs in GSTCD with lung function measures and COPD, as well as a notable change in GSTCD expression at key lung development stages. Quantification analyses were not performed in adult samples due to variation in compression of lung tissue across lung lobes likely originating from tissue processing. Further replicates are required to demonstrate consistent lung structure before full conclusions can be drawn. Moreover, murine Gstcd knock-out did not appear to affect overall phenotypes or breeding, observed during colony maintenance, re-enforcing that Gstcd is may not be essential for development or growth in mammals.

In the future it would be useful to compare these results to the alternative Gstcd knock-out mouse generated by the European Conditional Mouse Mutagenesis Program (EUCOMM), a separate member of the IMPC to the source of the discussed Gstcd knock-out mouse (NorCOMM). $Gstcd^{EUCOMM}$ knock-out mice were generated in a similar manner as described previously for $Gstcd^{NorCOMM}$; using a gene cassette targeted to Gstcd resulting in deletion of exon 2 ($GSTCD^{NorCOMM}$ results in deletion of exons 4 and 5) (IMPC: GSTCD(EUCOMM) 2015; IMPC: GSTCD(NCOM) 2015). For the $Gstcd^{EUCOMM}$ -/- knock-out mice the respiratory system has yet to be assessed but these mice do show three significant genotype-phenotype associations compared to wild-type across an extensive range of phenotypes, these include measures such as organ morphology, clinical chemistry and sensory testing (IMPC: GSTCD 2015) (Table 18). So far significant deviations have been identified in measures of adipose tissue, cardiovascular

system and hearing/vestibular/ear assessment. Overall, the reasons behind these results are currently unknown.

However, SNPs located within *GSTCD* have also been associated with $FEV_1$ in children and with growth, measured as bronchial responsiveness development (Kreiner-Møller et al. 2014). Therefore, suggesting that GSTCD may have an early life effect with specific focus on the bronchial epithelium, despite no gross changes in lung morphology having been observed in Gstcd -/- knock-out mice.

**Table 18. Phenotype data for male and female Gstcd$^{EUCOMM}$ -/- knock-out mice assessed within the IMPC by the IMPReSS phenotyping pipeline (IMPC: GSTCD 2015).** Three significant deviations in genotype-phenotype associations (bold text) have been identified to date in preliminary statistical analysis of standard IMPC phenotype parameters (IMPC significance threshold p<0.0001). Key: - no data available.

| Phenotype Parameter | Deviance from Gstcd +/+ |
|---|---|
| Adipose Tissue | **Abnormal body fat amount (p= 4.35 x $10^{-5}$)** |
| Behaviour/Neurological | Not significant |
| Cardiovascular System | **Anomaly in an electrocardiogram measurement (p= 4.75 x $10^{-6}$)** |
| Coat, Skin and Nail Colour/ Texture/Appearance | Not significant |
| Craniofacial | Not significant |
| Embryogenesis | - |
| Fertility and Viability | Not significant |
| Forelimb and Hindlimb Grip Strength | Not significant |
| Growth/Size/Body | Not significant |
| Hearing/Vestibular/Ear | **Abnormal auditory brainstem response (p= 5.38 x $10^{-7}$)** |
| Immune System | - |
| Limbs/Digits/Tail | Not significant |
| Mortality/Aging | Not significant |
| Nervous System | Not significant |
| Pigmentation | Not significant |
| Reproductive System | Not significant |
| Respiratory System | - |

#### 4.2.7.3 Optimising Antibody Markers of Cell Growth in Mouse Lung Tissue

Two proliferation markers were initially optimised in IHC with the aim of investigating cell growth in Gstcd knock-out mouse lung tissue. Yet, staining was observed only in the positive control tissue, obtained from a separate age-matched mouse. This was thought to be attributable to the wild-type lung tissue potentially being extracted during a low proliferative stage or anomalous tissue samples. Previous research indicates approximately 3% of cells in mouse lung tissue are proliferative, many of which are alveolar epithelial type II cells (Lee et al. 2006; Pagano et al. 2007; Chess et al. 2010). This work was not continued due to evidence becoming apparent that GSTCD is appears less likely to influence actively proliferating cells, obtained from cell growth assays (Click-iT® EdU) performed alongside this experiment. Instead, cell growth in the form of total cell number and potentially apoptosis measures may be more appropriate in future studies.

#### 4.2.7.4 Gstcd Protein Expression in Knock-out Mouse Lung Tissue

For the Gstcd knock-out mice studied here loss of expression was confirmed using reverse transcriptase-PCR by Cornell University (Michael Kotlikoff group, Cornell University). However, upon examination of protein expression in the knock-out mice, staining was evident in IHC of lung tissue across genotypes. Similar results were also identified at Cornell University using a different Gstcd primary antibody (Santa Cruz, sc-138332).

The results shown here can be considered reliable as no staining was observed in any of the matched isotype controls. This is corroborated by negative regions within the stained lung tissue as well, where airway smooth muscle appears unstained. Alternatively, these results may be indicative of a degree of non-specific antibody binding in mouse tissue as the antibody immunogen is targeted to the human GSTCD protein sequence. Furthermore, due to the relatively new interest in GSTCD, all easily available GSTCD antibodies are polyclonal (antibody molecules which share overall target specificity but differ in binding region)

hence may be considered more likely to bind non-specifically, particularly in different species. Additionally, similar staining has also recently been identified in Gstcd knock-out lung samples generated from a mouse colony now at University of Nottingham (n=3 per genotype) (unpublished, Tracy Rimington, University of Nottingham), suggesting that the original staining pattern observed here was unlikely to be due to small sampling or processing errors.

Although staining in wild-type mouse lung tissue shown here (Proteintech) indicates corresponding staining to results using an additional antibody in mouse lung tissue (Santa Cruz antibody, sc-138332 (Michael Kotlikoff group, Cornell University)) and to previous results in human tissue (Abnova antibody (Obeidat et al. 2013)), the antibody remains relatively new and not thoroughly validated in mouse tissue. Thus it may be necessary to further examine specificity of this primary antibody.

On the other hand, the GSTCD primary antibody (Proteintech) may be binding to alternative GSTCD variants not limited by the knock-out cassette insertion in Gstcd -/- mice and which remain transcriptionally and translationally active. This theory is partly corroborated by the recent identification of four new potential murine Gstcd transcripts determined by automated computational analysis (Ensembl Mouse Gstcd 2015). Hence, it is possible that not all of these variants have been transcriptionally silenced in this instance. However, these transcripts are not predicted to be protein coding. Nevertheless, one transcript is predicted to be processed and, of 9 new transcripts predicted for human GSTCD, 6 are predicted to be protein coding (Ensembl Human GSTCD 2015). Thus, additional protein coding variants in mice may yet be discovered.

A potential avenue for improving specificity may be to perform IHC on lung tissue obtained from the aforementioned alternative Gstcd knock-out mouse model ($Gstcd^{EUCOMM}$), involving deletion of an exon further upstream than the knock-out examined here. Disruption of Gstcd transcription at this site may be more effective at limiting the variety of Gstcd transcripts that are expressed due to an earlier interruption to transcription. Therefore with the potential for a

variety of additional Gstcd transcripts, studying loss of Gstcd function in Gstcd -/- mice remains difficult to assess.

## 4.3 GSTCD Protein Structure

Despite significant associations between SNPs in *GSTCD* with lung function and COPD, GSTCD is a relatively unknown protein. GSTCD was named after the GST family of enzymes as both proteins contain a C-terminal α- helical domain. Typically, a shared domain present in multiple proteins indicates the same or similar function. However, the function of GSTCD is not known and appears unlikely to be associated with classical GST activity due to previous evidence demonstrating lack of GST activity in an over-expression system. Hence, the aim of this section was to identify potential GSTCD function using protein prediction methods based on protein sequence homology.

An initial step in determining protein function is often to identify similar proteins with known function based on homologous sequences. Typically, a conserved domain in the protein sequence may be revealed which is common to a particular group of proteins sharing a similar function. Consequently, it can be inferred that the function of the unknown protein may be similar to that of the homologous template protein of known function. Fortunately, due to the theory that there may be a finite number of possible domains, approximately 70% of known protein sequences are suggested to have a minimum of one domain which is measurably related to at least one known protein structure (Fiser 2010). In this instance GSTCD function was inferred from protein sequence homology of both published transcripts for which there is experimental evidence (Obeidat et al. 2013).

To address the difficulty of lack of information regarding GSTCD function, homologous protein sequences were sought using three publically available, accessible and highly rated web-based protein prediction servers (I-TASSER, SWISS-MODEL and Phyre2) as assessed by CASP. These structure modelling servers were used to infer theoretical 3D protein structures and functions of GSTCD based upon homology of human GSTCD protein sequences (NCBI references: NP_001026890.2 and NP_079027.2) with proteins of known structure and function. Three key results sections from these servers were investigated to summarise GSTCD structure and function information.

### 4.3.1 Secondary Structure Prediction Suggests a Core Region for GSTCD

Secondary structure describes local structural conformations in proteins, of which the two main forms are α-helices and β-strands. The most common type of secondary structure in proteins is the α-helix, a coil structure, in a clock-wise direction, stabilised by hydrogen bonds. Typically, β-strands are comprised of a short polypeptide chain (3-10aa) in an extended conformation. Multiple β-strands can form β-sheets, where hydrogen bonds link several β-strands in parallel or anti-parallel formation.

Human GSTCD protein sequences for variant 1 and 2 were assessed in several protein prediction servers, producing theoretical secondary and tertiary structures based on sequence homology with proteins of known structure. The secondary structure for the full GSTCD sequence was predicted by the servers: I-TASSER and Phyre2, whereas SWISS-MODEL predicts structure only for the methyltransferase domain region where strong sequence homology was identified.

Concordant results are provided across servers and between variants in relation to secondary structure, indicating regions of β-strands located at N- and C-termini and alpha helices throughout the protein sequence for both GSTCD variants (sub-sections: 4.3.1.1, 4.3.1.2 and 4.3.1.3 and Figure 28). Potential transmembrane regions were identified by Phyre2, suggested to result in localisation at the cell membrane (sub-section 4.3.1.3). Notably, a region of several β-strands can be seen in Figure 28 (A, B and C), indicating a β-sheet in a core region of the protein. The β-sheet appears to be formed from the C-terminal region of GSTCD as this structure is also evident in the SWISS-MODEL 3D structure predictions (Figure 28B).

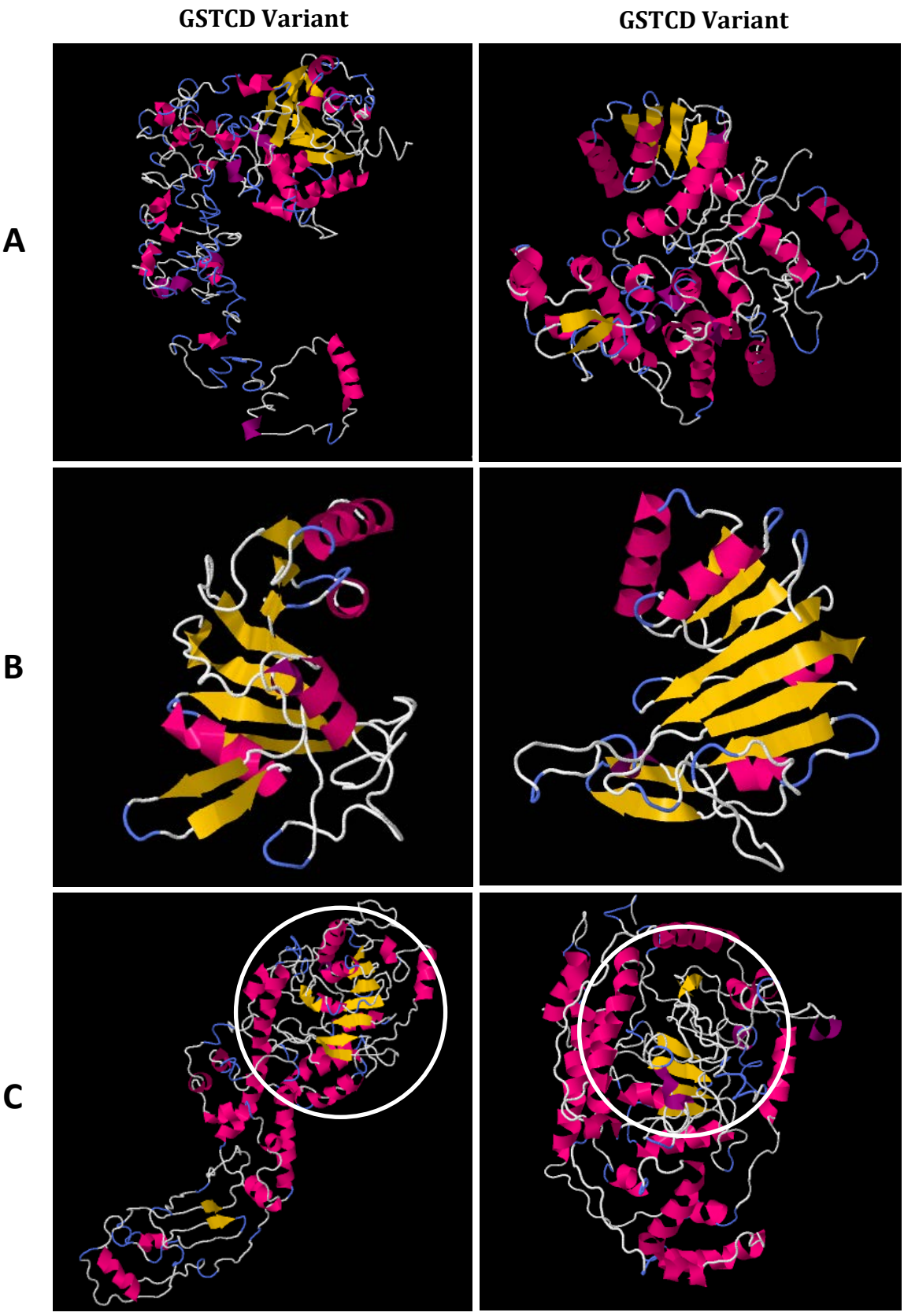



Figure continues on the following page

**Figure 28. Top 3D protein models built for human GSTCD variants 1 and 2 using the protein prediction servers: I-TASSER (A), SWISS-MODEL (B) and Phyre2 (C).** Structures indicate polypeptide backbones of amino acid sequences. I-TASSER and Phyre2 indicate potential secondary and tertiary protein structure for the full GSTCD sequence demonstrating an abundance of α-helices and few β-strands (A and C). A primary region of β-sheets indicates a core region. SWISS-MODEL demonstrates known secondary structure of the best aligned homologous template protein (Protein RdmB methyltransferase) for the C-terminal region only of both variants of GSTCD (variant 1: 441-633aa and variant 2: 351-546aa) (B). The C-terminal region identified with the highest confidence scores in Phyre2 (also the approximate location of the SWISS-MODEL C-terminal protein prediction model (B)) is shown in a white circle in Phyre2 structures (C). Colour Key: Pink coils - α-helices, Yellow arrows - β-strands, White/Blue lines - regions of undetermined structure.

#### *4.3.1.1 I-TASSER Secondary Structure Prediction*

I-TASSER predicted secondary structure results for full length GSTCD protein sequences.

Key

H – α-helix

S – β-strand

C – Coil, linking region between primary structures

0-9 – Score of confidence (0 - low and 9 - high) for secondary structure prediction at this site

Sequence rows proceed with the labels:

- Protein sequence
- Predicted secondary structure
- Confidence score for predicted structure

## I-TASSER GSTCD Variant 1

```
MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSRDS   60aa
CCCCCCCCCCCCSSSSSSSSCCCCSSSSSSHHHHHHHHHHCCCCCSSSSSSSCCCCCCCC
9853135665635999998547877876642179999998089862799996777777754

SLLRDDLIQDVEIQIISRQELPPIVQNCCLPAVVERSDNFCRAGLAVVLRHIIQKSYEAD   120aa
CSSSCCCCCCSSSSSSSHHHCCHHHHHCCCCSSSSCCCCSSSSHHHHHHHHHHHHHHHCC
0341257776048998743276678738798799669988973578999999998775417

PLKKELLELLGFKKTCLKACAEVSQWTRLCELTIPLAIENFLRESSDQPPTIPVEILQLE   180aa
CCHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHCCHHHHHHHHHCCCCCCCCCCHHHHHHH
8546589871865004330077767740200389999999737776665476999999

KKLSEPVRVHNDDKLRRQKLKQQKADGVGPPLTKGKAKSKVHTQETSEGLDSSSKSLELK   240aa
HHHCCCCSSCCHHHHHHHHHHHHHHCCCCCCCCCCCCHHHHCCCCCCCCCCCCCCHHHHH
9847972413589999999987764145654211332022111112345576430146666

VAFSKLTVQEEPATTNREPSHIRKAKASDLPPLEHVFAEGLYFTLADIVLLPCIHHFLVI   300aa
HHHHHHHHCCCCCCCCCCCCHHHHHCCCCCCHHHHHHHHCCCCCCHHHHHHHHHHHHHHH
6655621110346666662111000345674045454024322288788899999999999

ISRKFSEKLVEFPLLASWYQRIQEVPGVKTAASKCGIQFLHLPKLLTTSTEQHPNLCEVP   360aa
HHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCCCHHHHCCCCHHHHHHHHHCC
8887777510388558999999862306788886485412333425351656377661 42

GVEEQSDPLFIGGPRPTMAKLMEKGIEVMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDR   420aa
CCCCHHHHHHCCCCHHHHHHHHHHCCHHHCCCCCCCCCCCCHHHCCCCCCCCCCCCCCHH
2210011232068588998876623022126685433445722075544424465651 34

ALRKQQQLNNLVYVVTNQAKPGDRIVDFCSGGGHVGIVLAHMLPSCQVTLIENKELSLIR   480aa
HHHHHHHHHHHHHHHHHHHCCCCCSSSSSCCCCCHHHHHHHHHCCCCSSSSSSCCHHHHHH
5689999999999999764899889981799978999999977998899997988999 99

AKKRSDELGLSNIWFIQANMEYFTGMFNIGVALHACGVATDMVIEHCIKTRASFVTCPCC   540aa
HHHHHHHHCCCCCSSSSSSHHHCCCCCCSSSSSSCCCCHHHHHHHHHHHCCCCSSSSCCC
9999997037760689855521887776567610276045999999998099789945 86

YGFIQNTSKFNFPKSEQFKKTLSYKEHMILCRFADQTAVQLPPQRRLIGKQCMCLVDLDR   600aa
CHHHHHCCCCCCCCCCHHHHHHHCHHHHHHHHHHHCCCCCCCCCHHHHHHHHHHHHHHHHH
1776521323554137677552523455665203543446860111367899999999 99

ARAAEECGYSVQVISMEPESCSPKNNMIVGVPI   633aa
HHHHHHCCCSSSSSSSSCCCCCCCCCSSSSSSC
9999977997999999676657452359998 69
```

## I-TASSER - GSTCD Variant 2

```
MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSQWT   60aa
CCCCCCCCCCCCSSSSSSSSCCCCCSSCCSSSSSHHHHHHCCCCCHHHHHHCHHHHHHHH
98654666651599997423666100002012000244407863110001457765433

RLCELTIPLAIENFLRESSDQPPTIPVEILQLEKKLSEPVRVHNDDKLRRQKLKQQKADG   120aa
HHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHHHHCCCCSSSSCHHHHHHHHHHHHHCC
32521078999998753335665378899999987189748714799999999987406

VGPPLTKGKAKSKVHTQETSEGLDSSSKSLELKVAFSKLTVQEEPATTNREPSHIRKAKA   180aa
CCCCCHHHHCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCCCCCCCCCHHHHHHHCC
68740111002200310001223566521567766543000236430121245665401 0

SDLPPLEHVFAEGLYFTLADIVLLPCIHHFLVIISRKFSEKLVEFPLLASWYQRIQEVPG   240aa
CCCCCCCCSSCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHH
22675014310266465999998788999999999988878750885678999888401 3

VKTAASKCGIQFLHLPKLLTTSTEQHPNLCEVPGVEEQSDPLFIGGPRPTMAKLMEKGIE   300aa
HHHHHHHCCCSSCCHHHHCCCHHHHHHHHHHCCCCHHHHHHHHHCCCHHHHHHHHHHCCH
88888745851033433063407563876622001122212320686999876551310

VMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDRALRKQQQLNNLVYVVTNQAKPGDRIVD   360aa
HHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHCCCCCSSSS
21266852133246002645445134656411335899999999999998764899889 98

FCSGGGHVGIVLAHMLPSCQVTLIENKELSLIRAKKRSDELGLSNIWFIQANMEYFTGMF   420aa
SCCCCCHHHHHHHHHCCCCSSSSSSCCHHHHHHHHHHHHHCCCCCCSSSSSSHHHCCCCC
37999789999999779987999989899999999999971467603789876527888 7

NIGVALHACGVATDMVIEHCIKTRASFVTCPCCYGFIQNTSKFNFPKSEQFKKTLSYKEH   480aa
CSSSSCCCCCCHHHHHHHHHHCCCCSSSSCCCCHHHHHHCCCCCCCCHHHHHHHCHHHH
77887113760689999999980998999358738876311246651066776534456 7

MILCRFADQTAVQLPPQRRLIGKQCMCLVDLDRARAAEECGYSVQVISMEPESCSPKNNM   540aa
HHHHHCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHCCCSSSSSSSCCCCCCCCCCS
67653045434578632103678999999999999997799789999807555756125

IVGVPI   546aa
SSSSSC
998869
```

#### *4.3.1.2 SWISS-MODEL Secondary Structure Prediction*

SWISS-MODEL predicted secondary structure results for C-terminal region of GSTCD protein sequences, based on homology region alignments with the top homologous protein hit for each variant (GSTCD variant 1 - Protein RdmB methyltransferase and GSTCD variant 2 - Putative rRNA methylase).

Key

H – α-helix

S – β-strand

C – Coil, linking region between primary structures

(No confidence scores provided for secondary structure predictions in these results)

Sequence rows proceed with the labels:

- Protein sequence
- Predicted secondary structure

SWISS-MODEL - GSTCD Variant 1

```
MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSRDS   60aa
------------------------------------------------------------

SLLRDDLIQDVEIQIISRQELPPIVQNCCLPAVVERSDNFCRAGLAVVLRHIIQKSYEAD   120aa
------------------------------------------------------------

PLKKELLELLGFKKTCLKACAEVSQWTRLCELTIPLAIENFLRESSDQPPTIPVEILQLE   180aa
------------------------------------------------------------

KKLSEPVRVHNDDKLRRQKLKQQKADGVGPPLTKGKAKSKVHTQETSEGLDSSSKSLELK   240aa
------------------------------------------------------------

VAFSKLTVQEEPATTNREPSHIRKAKASDLPPLEHVFAEGLYFTLADIVLLPCIHHFLVI   300aa
------------------------------------------------------------

ISRKFSEKLVEFPLLASWYQRIQEVPGVKTAASKCGIQFLHLPKLLTTSTEQHPNLCEVP   360aa
------------------------------------------------------------

GVEEQSDPLFIGGPRPTMAKLMEKGIEVMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDR   420aa
------------------------------------------------------------

ALRKQQQLNNLVYVVTNQAKPGDRIVDFCSGGGHVGIVLAHMLPSCQVTLIENKELSLIR   480aa
--------------------HCCSSSSSCCCCCHHHHHHHHHCCCCSSSSSCCCHHHHHHH

AKKRSDELGLSNIWFIQANMEYFTGMFNIGVALHACGVATDMVIEHCIKTRASFVTCPCC   540aa
HHHHHHHCCCCCSSSCCCCCCCCCCCSSSSSSSCCHHHCCHHHHHHCSSSSSSSSSSCCC

YGFIQNTSKFNFPKSEQFKKTLSYKEHMILCRFADQTAVQLPPQRRLIGKQCMCLVDLDR   600aa
CCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCHHHH

ARAAEECGYSVQVISMEPESCSPKNNMIVGVPI   633aa
HHHHHHCCSSSSSSSSSCCCCCCCSSSSSSSS
```

SWISS-MODEL - GSTCD Variant 2

```
MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSQWT   60aa
------------------------------------------------------------

RLCELTIPLAIENFLRESSDQPPTIPVEILQLEKKLSEPVRVHNDDKLRRQKLKQQKADG   120aa
------------------------------------------------------------

VGPPLTKGKAKSKVHTQETSEGLDSSSKSLELKVAFSKLTVQEEPATTNREPSHIRKAKA   180aa
------------------------------------------------------------

SDLPPLEHVFAEGLYFTLADIVLLPCIHHFLVIISRKFSEKLVEFPLLASWYQRIQEVPG   240aa
------------------------------------------------------------

VKTAASKCGIQFLHLPKLLTTSTEQHPNLCEVPGVEEQSDPLFIGGPRPTMAKLMEKGIE   300aa
------------------------------------------------------------

VMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDRALRKQQQLNNLVYVVTNQAKPGDRIVD   360aa
-----------------------------------------------------HCCCCCCSSSS

FCSGGGHVGIVLAHMLPSCQVTLIENKELSLIRAKKRSDELGLSNIWFIQANMEYFTGMF   420aa
CCCCCCHHHHHHHHHHCCCSSSSSSCCHHHHHHHHHHHHHCCCHHSSSSSCCHHHCCCCS

NIGVALHACGVATDMVIEHCIKTRASFVTCPCCYGFIQNTSKFNFPKSEQFKKTLSYKEH   480aa
SSSSSSSSSCHHHHHHHHHHSSSSSSSSSSCCCCCCCCCCCCCCHHHHHHHHHHHHHHHH

MILCRFADQTAVQLPPQRRLIGKQCMCLVDLDRARAAEECGYSVQVISMEPESCSPKNNM   540aa
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCSSSSSSSSCCCCCCCCSSS

IVGVPI   546aa
SSSS--
```

#### *4.3.1.3 Phyre2 Secondary Structure Prediction*

Predicted secondary structure results for full length GSTCD protein sequences.

Key

H – α-helix

S – β-strand

T – Transmembrane helix

C – Coil, linking region between primary structures

0-9 – Score of confidence (0 - low and 9 - high) for secondary structure prediction at this site

Sequence rows proceed with the labels:

- Protein sequence
- Predicted secondary structure
- Confidence score for predicted structure

## Phyre2 - GSTCD Variant 1

```
MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSRDS   60aa
CCCCCCCCCCCCSSSSSSSCCCCSSSSSCTTTTTTTTTTTCCCCCSSSSSSSCCCCCCCC
98765555888779999988995787555899999999987999858999967888765 4

SLLRDDLIQDVEIQIISRQELPPIVQNCCLPAVVERSDNFCRAGLAVVLRHIIQKSYEAD   120aa
CCCCCCCCCCCSSSSSSHHHCCCHHHCCCCCSSSSCCCCSSSSHHHHHHHHHHHHHHCCC
44556666775699999557875577589899968998895458999999999876678

PLKKELLELLGFKKTCLKACAEVSQWTRLCELTIPLAIENFLRESSDQPPTIPVEILQLE   180aa
CCHHHHHHHCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHH
85578999868687654566678777667765599999999986988876599999999 9

KKLSEPVRVHNDDKLRRQKLKQQKADGVGPPLTKGKAKSKVHTQETSEGLDSSSKSLELK   240aa
HHHCCCCCCCCHHHHHHHHHHHHHCCCCCCHHHCCCCCCCCCHHHCCCCCCCCCCHHHH
98658654478999999999998654567544544554445545546677776655677 7

VAFSKLTVQEEPATTNREPSHIRKAKASDLPPLEHVFAEGLYFTLADIVLLPCIHHFLVI   300aa
HHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHTTTTTTTTTTTTTTTTTTH
77654455478887766555566788776675456766779999999999999999999 9

ISRKFSEKLVEFPLLASWYQRIQEVPGVKTAASKCGIQFLHLPKLLTTSTEQHPNLCEVP   360aa
HHHHHHHHHHCCCHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHHHCCCCCCCHHHHHHC
76677887767984899998887567875467875456555677755766665566665 5

GVEEQSDPLFIGGPRPTMAKLMEKGIEVMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDR   420aa
CCCHHHHHHHHCCCHHHHHHHHCCCCCCCCCCCCCCCCCCCCHHHCCCCCCCCCCCCCCCC
55456778875576566666654686555688876555676676897777766678865 4

ALRKQQQLNNLVYVVTNQAKPGDRIVDFCSGGGHVGIVLAHMLPSCQVTLIENKELSLIR   480aa
CCCHHHHHHHHHHHHHHCCCCCSSSSSCCCCCHHHHHHHHHCCCCSSSSSSCCHHHHHHH
45589999999999997458999999999999999999999889999999899999999 9

AKKRSDELGLSNIWFIQANMEYFTGMFNIGVALHACGVATDMVIEHCIKTRASFVTCPCC   540aa
HHHHHHHCCCCCSSSSSCCHHHCCCCCCSSSSSCCCCHHHHHHHHHHHHCCCCSSSSSCC
99999986999859998897999987758765435775789999999656999999959 9

YGFIQNTSKFNFPKSEQFKKTLSYKEHMILCRFADQTAVQLPPQRRLIGKQCMCLVDLDR   600aa
HHHHCCCCCCCCCCCHHHHHCCCHHHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHHH
78955666678876556665555556789988765457775689999999999989999 9

ARAAEECGYSVQVISMEPESCSPKNNMIVGVPI   633aa
HHHHHHCCCSSSSSSSSCCCCCSSSSSSSSSSC
99999989979999998899887557898667 9
```

Phyre2 - GSTCD Variant 2

```
MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSQWT   60aa
CCCCCHHCCCCCSSSSSSSCCCCCCCCCTTTTTTTTTTHHHCCCCHHHHHHHHHHHHHH
9876555887569998557898544389999999999858875678888887777888

RLCELTIPLAIENFLRESSDQPPTIPVEILQLEKKLSEPVRVHNDDKLRRQKLKQQKADG   120aa
HHHHHHHHHHHHHHHHCCCCCCCCCCHHHHHHHHHHCCCSSSCCHHHHHHHHHHHHHHC
888888999999998687887777999999767875697343799999999999987655

VGPPLTKGKAKSKVHTQETSEGLDSSSKSLELKVAFSKLTVQEEPATTNREPSHIRKAKA   180aa
CCCHHHHHHHHHHHHHHHHCCCCCCCCCCSSSSCCCCCCCCCCCCCCHHHHHHHHHHCCC
665555545555556677656888887765454456778766777645567787765687

SDLPPLEHVFAEGLYFTLADIVLLPCIHHFLVIISRKFSEKLVEFPLLASWYQRIQEVPG   240aa
CCCCCHHHHHHHHTTTTTTTTTTTTTTTTTTTHHHHHHHHHHHCCCCHHHHHHHHHHCCC
767755788877899999999999999999999987777898665698589999877578 8

VKTAASKCGIQFLHLPKLLTTSTEQHPNLCEVPGVEEQSDPLFIGGPRPTMAKLMEKGIE   300aa
CCHHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHCCCHHHHHHHCCCCCC
655678765665556776567554567776554556777788764654556666548876

VMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDRALRKQQQLNNLVYVVTNQAKPGDRIVD   360aa
CCCCCCCCHHCCCCHHHCCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHHHCCCCCCSSSS
667888745456766667777766667788655666999999999999875689999999

FCSGGGHVGIVLAHMLPSCQVTLIENKELSLIRAKKRSDELGLSNIWFIQANMEYFTGMF   420aa
SCCCCCHHHHHHHHHCCCCSSSSSSCCHHHHHHHHHHHHHCCCCCSSSSSCCHHHCCCCC
999999999999998799899999899899999999999859999599966978898879

NIGVALHACGVATDMVIEHCIKTRASFVTCPCCYGFIQNTSKFNFPKSEQFKKTLSYKEH   480aa
CSSSSSCCCCCHHHHHHHHHHCCCCSSSSCCCHHHHHCCCCCCCCCCHHHHHCCHHHHH
999985567856899999997569999999699789555655788875677664556778

MILCRFADQTAVQLPPQRRLIGKQCMCLVDLDRARAAEECGYSVQVISMEPESCSPKNNM   540aa
HHHHHHHHHHCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHCCCSSSSSSSCCCCCSSSSS
899988776546754589999999999999999999999899799999988898774478

IVGVPI   546aa
SSSSSC
986669
```

### 4.3.2 Conserved Domains Predicted for GSTCD Suggest Methyltransferase Activity

A protein sequence common to a particular group of proteins sharing a similar function is typically referred to as a conserved domain. Consequently, if an unknown protein contains a sequence homologous to a known conserved domain, function can be inferred.

Homologous matches to the GSTCD protein sequence were searched with results indicating a high homology to methyltransferase enzymes across the servers (Table 19). Sequence homology is predominantly located across the predicted methyltransferase domain (variant 1: 423-541aa and variant 2: 336-454aa) as opposed to the GST C-terminal α-helical domain (variant 1: 251-332aa and variant 2: 164-245aa) (Figure 29) and this is consistent throughout the top homologous matches for each server (all of top 50 hits shown for I-TASSER and SWISS-MODEL and top 228 hits for Phyre2). Additionally, methyltransferase enzymes are frequently found in the top 3 structural templates listed for each server and the top 50 template results in SWISS-MODEL were entirely methyltransferase proteins, with specific focus on rRNA methyltransferase function (not shown). Alternatively, I-TASSER identified homology in this region with transport receptors, specifically importin β subunit 1 (Table 19), involved in transporting proteins into the nucleus (Andrade et al. 2001). Other top hits included: exodeoxyribonuclease (catalyses degradation of double stranded DNA), Ras guanyl-releasing protein 1 (nucleotide exchange factor specifically activating Ras, activates the extracellular signal-regulated kinases/mitogen-activated protein kinase (ERK/MAPK) cascade) and nicotinamide adenine dinucleotide binding (NADB)-Rossmann fold superfamily protein (structural motif found in proteins that bind nucleotides) (Table 19). The top 10 hits identified as having the strongest homology with GSTCD by each server are shown in Appendix section 6.6 Table 27.

Additionally, due to the potentially decreasing likelihood of a GST enzyme function for GSTCD, a control protein prediction search using a GST protein was performed to compare against these results. GST-Mu5 (GSTM5, 218aa) was selected as this was previously demonstrated as a successful positive control in

GST activity assays (Ceri Stewart, University of Nottingham). Over-expression of GSTM5 by transfection in hBECs with vectors containing GSTM5 was shown to increase typical GST activity by 2-3 times.

Results indicate high homology within the GST Mu sub-group across the full GSTM5 protein sequence (Table 20), suggesting strong conservation of structure (Figure 30) but this does not correlate with GSTCD homology results. Moreover, the predicted regions of intrinsic disorder do not correspond with the co-ordinate region for the GST domain in GSTCD. Secondary structure predictions also indicate variation, whereby GSTM5 is predicted to contain β-strands only at the N-terminus (Results section 4.3.1 and 4.3.2.1). Whereas GSTCD secondary and 3D structure appears to focus on the C-terminus and methyltransferase region.

**Table 19. Investigating GSTCD protein sequence using web-based protein structure prediction servers indicates methyltransferase homology.** Common results from three leading protein prediction servers are summarised, comparing GSTCD against proteins of known structure and function. Results include searches for both sequence variants of GSTCD (NCBI references: NP_001026890.2 and NP_079027.2). Results indicate homology with methyltransferase enzymes due to a potential shared conserved domain. Key: **N/A** denotes that this result is either not provided by this server or the prediction does not assess full length GSTCD and **%** indicates the percentage of sequence covered.

Table continues on the following page

| | Protein Structure Prediction | | | | | |
|---|---|---|---|---|---|---|
| **Protein Structure Results** | **I-TASSER** | | **SWISS-MODEL** | | **Phyre2** | |
| | **Variant 1** | **Variant 2** | **Variant 1** | **Variant 2** | **Variant 1** | **Variant 2** |
| **Predicted Secondary Structure** | • Primarily α-helices with intermittent coil structures<br>• β-strands only present at termini | • Primarily α-helices with intermittent coil structures<br>• β-strands mainly present at termini | N/A | N/A | • Primarily intermittent α-helices (45%)<br>• β-strands only present at termini (12%)<br>• Two predicted transmembrane regions (30-40aa and 281-299aa) | • Primarily intermittent α-helices (55%)<br>• β-strands only present at termini (10%)<br>• Two predicted transmembrane regions (29-38aa and 194-212aa) |
| **Regions of Primary Sequence Homology** | • Primarily across the methyltransferase domain to C-terminus<br>• Partial matches across whole protein | • Primarily across the methyltransferase domain to C-terminus<br>• Partial matches across whole protein | Only across the methyltransferase domain to C-terminus | Only across the methyltransferase domain to C-terminus | Only across the methyltransferase domain | Only across the methyltransferase domain |
| **Intrinsic disorder** | N/A | N/A | Low disorder across C-terminus | High disorder at C-terminus | Two theoretical central regions of disorder (27%):<br>• 204-270aa<br>• 325-386aa | Two theoretical central regions of disorder (26%):<br>• 118-184aa<br>• 257-299aa |
| **Top 3 Structural Templates** | 1. Putative methyltransferase<br>2. Importin β subunit<br>3. Exodeoxyribonuclease | 1. Ras guanyl-releasing protein 1<br>2. Methyltransferase<br>3. Methyltransferase | 1. Protein RdmB methyltransferase<br>2. Protein RdmB methyltransferase<br>3. Protein RdmB methyltransferase | 1. Putative rRNA methylase<br>2. Putative rRNA methylase<br>3. rRNA small subunit methyltransferase | 1. rRNA small subunit methyltransferase<br>2. Transferase<br>3. S-adenosyl-L-methionine-dependent methyltransferase | 1. Transferase<br>2. NADB-Rossmann fold superfamily protein<br>3. S-adenosyl-L-methionine-dependent methyltransferase |

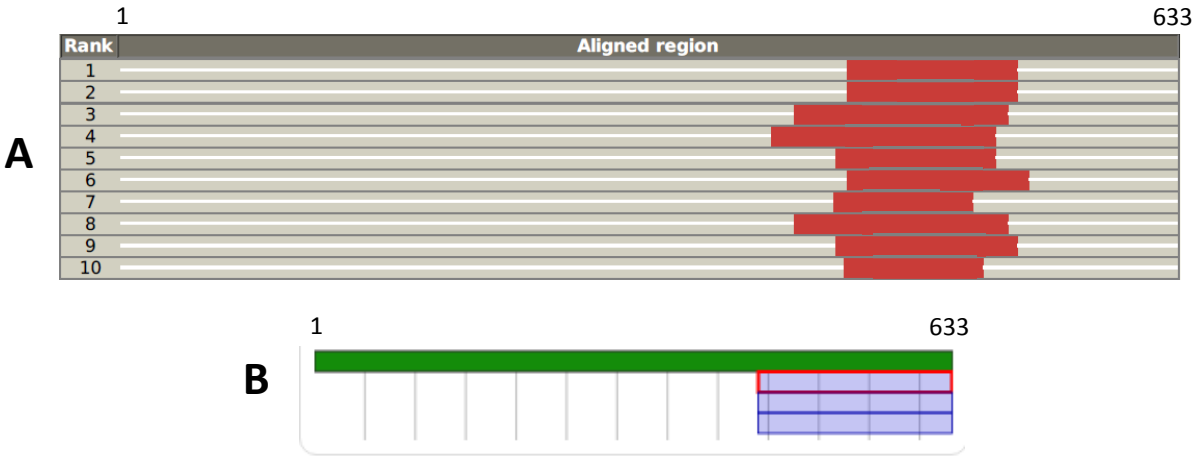



**Figure 29. Representative examples of primary homology region identified by the three protein prediction servers: I-TASSER, SWISS-MODEL and Phyre2.** Representative figure of domain analysis result from GSTCD search using Phyre2, showing top protein homology matches centred at the predicted methyltransferase domain in GSTCD. This pattern continues in this region consistently for the top 228 ranked homologous proteins (A). SWISS-MODEL pictorially indicates location of homology alignment for the top three homologous matches to GSTCD protein sequence as occurring at the C-terminus (B).

**Table 20. Investigating GSTM5 protein sequence using web-based protein structure prediction servers as control results.** Common results from three leading protein prediction servers are summarised, comparing GSTM5 against proteins of known structure and function. GSTM5 is demonstrated to have significant levels of homology across the full protein sequence with other members of the GST family of enzymes. Key: N/A denotes that this result is either not provided by this server or the prediction does not assess full length GSTCD and % indicates the percentage of sequence covered.

| | **Protein Structure Prediction** | | |
|---|---|---|---|
| **Protein Structure Results** | **I-TASSER** | **SWISS-MODEL** | **Phyre2** |
| **Predicted Secondary Structure** | • Primarily intermittent α-helices<br>• Several β-strands at N-terminus (small signal also at 147-149aa) | N/A | • Primarily intermittent α-helices (56%)<br>• β-strands only present at N-terminus (9%) |
| **Regions of Primary Sequence Homology** | Full length sequence (218aa) | Majority of protein in central region (88% protein of sequence) | Majority of protein in central region, variation only evident at termini |
| **Intrinsic disorder** | N/A | Three regions indicate relatively high intrinsic disorder (33-43aa, 115-129aa and 209-218aa) | Two theoretical terminal regions of disorder (11%):<br>• 38-41aa<br>• 212-218aa |
| **Top 3 Structural Templates** | 1. Glutathione S-Transferase Mu1a-1a<br>2. Glutathione-S-Transferase Mu2-2<br>3. Glutathione S-Transferase Mu1 | 1. Glutathione S-transferase Mu1<br>2. Glutathione S-Transferase<br>3. Glutathione S-Transferase | 1. Glutathione S-transferase Mu1a<br>2. Glutathione S-transferase Mu1-1<br>3. Glutathione S-Transferase |

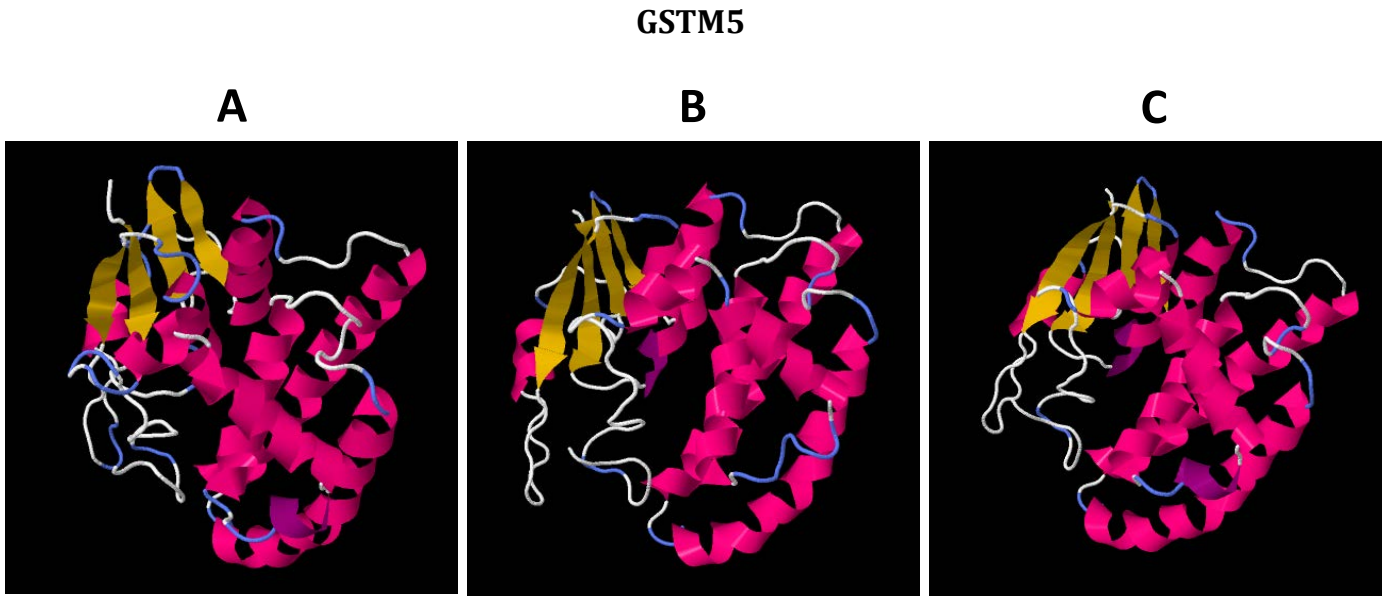



**Figure 30. Top 3D protein models built for human GSTM5 using the protein prediction servers: I-TASSER (A), SWISS-MODEL (B) and Phyre2 (C).** Structures indicate polypeptide backbones of amino acid sequences. Models indicate potential secondary and tertiary protein structure for the full GSTM5 sequence demonstrating an abundance of α-helices and few β-strands (A-C). Highly similar structures are obtained across protein prediction servers and a primary region of β-sheets indicates a core region (A-C). Colour Key: Pink coils - α-helices, Yellow arrows - β-strands, White/Blue lines - regions of undetermined structure.

#### *4.3.2.1 Secondary Structure Prediction for GSTM5*

Key

H – α-helix

S – β-strand

T – Transmembrane helix

C – Coil, linking region between primary structures

0-9 – Score of confidence (0 - low and 9 - high) for secondary structure prediction at this site

Sequence rows proceed with the labels:

- Protein sequence
- Predicted secondary structure
- Confidence score for predicted structure

I-TASSER GSTM5

```
MPMTLGYWDIRGLAHAIRLLLEYTDSSYVEKKYTLGDAPDYDRSQWLNEKFKLGLDFPNL   60aa
CCCSSSSSCCCCHHHHHHHHHHCCCCCSSSSSCCCCCCCCCHHHHHHHHHCCCCCCCCC
992899869970289999999982998489985465432336788742421158999999

PYLIDGAHKITQSNAILRYIARKHNLCGETEEEKIRVDILENQVMDNHMELVRLCYDPDF   120aa
CSSSSCCSSSSSHHHHHHHHHHHCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCH
989989999863899999998732668979999999999999999999999999708526

EKLKPKYLEELPEKLKLYSEFLGKRPWFAGDKITFVDFLAYDVLDKRIFEPKCLDAFLNL   180aa
HHHHHHHHHHHHHHHHHHHHHHHCCCCSSSCCCCCHHHHHHHHHHHHHHCCCCCCCCHHH
678999999999999999997899865589878999999999998757553100878 9

KDFISRFEGLKKISAYMKSSQFLRGLLFGKSATWNSK   117aa
HHHHHHHHHCHHHHHHHCCCCCCCCCCCCCCCCCCC
999999809889998769987899988867645899
```

## SWISS-MODEL GSTM5

```
MPMTLGYWDIRGLAHAIRLLLEYTDSSYVEKKYTLGDAPDYDRSQWLNEKFKLGLDFPNL   60aa
-CSSSSSSCCCHHHHHHHHHHHHCCCCSSSSSSCCCCCCCCCCHHHHHHHCCCCCCCCCC

PYLIDGAHKITQSNAILRYIARKHNLCGETEEEKIRVDILENQVMDNHMELVRLCYDPDF   120aa
CSSSSCCSSSSCHHHHHHHHHHHCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHHCCCH

EKLKPKYLEELPEKLKLYSEFLGKRPWFAGDKITFVDFLAYDVLDKRIFEPKCLDAFLNL   180aa
HHHHHHHHHHHHHHHHHHHHHHHHCCCCSCCSCCCCHHHHHHHHHHHHHHHCCCCCCCCSSS

KDFISRFEGLKKISAYMKSSQFLRGLLFGKSATWNSK   117aa
SSSSSSSSCCSSSSSSSCCCCCCCCCCCCCCCCCCCC
```

## Phyre2 GSTM5

```
MPMTLGYWDIRGLAHAIRLLLEYTDSSYVEKKYTLGDAPDYDRSQWLNEKFKLGLDFPNL   60aa
CCSSSSSCCCCHHHHHHHHHHHHCCCCSSSSSSCCCCCCCCCHHHHHHHHHHCCCCCCCC
99589857998778999999998899948997555545668989966556755809999 9

PYLIDGAHKITQSNAILRYIARKHNLCGETEEEKIRVDILENQVMDNHMELVRLCYDPDF   120aa
CSSSSCCSSSSCHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCH
88999989985689999999985999999899999999999999999999998599999

EKLKPKYLEELPEKLKLYSEFLGKRPWFAGDKITFVDFLAYDVLDKRIFEPKCLDAFLNL   180aa
HHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCHHHHHHHHHHHHHHHCHHHHCCCHHH
99999999999999999999865899567888699999999999999759855479778

KDFISRFEGLKKISAYMKSSQFLRGLLFGKSATWNSK   117aa
HHHHHHHHCCHHHHHHHHCCCCCCCCCCCCHHCCCCC
99999985899999975657789999984555679 9
```

### 4.3.3 Intrinsic Disorder Prediction Suggests Minimal Degree of Disorder in GSTCD Protein Structure

Intrinsic disorder assesses the degree to which a region of protein has a fixed 3D conformation. Two regions of intrinsic disorder were identified in Phyre2 (Table 19) with the second region for each variant located just downstream of the predicted GST domain. SWISS-MODEL reports level of intrinsic disorder only across the region of GSTCD with a matched homologous protein (this is approximately from the methyltransferase domain to the C-terminus) and in this instance only variant 2 shows a marginal degree of disorder.

Overall, protein prediction modelling of GSTCD indicates methyltransferase as a key domain from several database results. Although there appears to be many homologues indicating a methyltransferase domain and possibly function, results should be approached with caution. Alignment with template proteins in this instance relies upon resemblance based on similar characteristics of amino acids as opposed to extended regions of direct protein sequence alignment.

In conclusion, upon investigation of GSTCD homologues and protein structure it was identified that methyltransferase activity could be considered a potential function for GSTCD; based on this being the most prominent homology domain identified from several GSTCD protein prediction servers.

### 4.3.4 Discussion

The function of unknown proteins of interest can be inferred from known (template) proteins with sequence homology, as this often represents similarity at the structural level. Therefore the potential function of GSTCD was examined by searching protein prediction servers to identify homologous proteins of known function to infer possible functions of GSTCD. This was performed using three protein prediction servers to identify homologous proteins and, in particular, identify conserved domains within these proteins which often indicate a shared or similar function.

To address the difficulty of lack of information regarding GSTCD function, homologous protein sequences were sought using three publically available, accessible and highly rated web-based protein prediction servers (I-TASSER, SWISS-MODEL and Phyre2) as assessed by CASP. Several servers were selected to determine whether similar results would be obtained regardless of variation in bioinformatic analysis methodology.

#### *4.3.4.1 Secondary Structure Prediction Suggests a Core Region for GSTCD*

The two primary secondary structures of proteins are α-helices and β-strands. α-helices can interact with and be stabilised by water molecules, whereas β-strands typically consist of hydrophobic amino acid residues. Therefore, β-strands tend to be located within the core structure of proteins (Moran et al. 2012). In these results secondary structure predications from all three servers demonstrated a tendency of α-helices towards the centre and β-strands located towards the termini of GSTCD.

Any extended region of a common single secondary structure type can act as a recognition motif for binding to another protein (Remaut & Waksman 2006). These results show several extended regions of α-helices repeats, therefore indicating that regions important to functional activity of the protein may be centrally located in the protein sequence. However the short intermittent pattern

of β-strands across the methyltransferase domain in the GSTCD protein sequence suggests the formation of β-sheets. The β-sheets are predicted to be located in a core region of GSTCD, possibly suggesting an important role in protein function.

Potential transmembrane regions were also identified by Phyre2, suggested to result in localisation at the cell membrane, however this is not evidenced in previous lung tissue localisation research (Obeidat et al. 2013).

Overall, a complex tertiary structure indicating a potential active site has not yet been predicted due to the absence of templates with significantly high homology to GSTCD.

#### *4.3.4.2 Conserved Domains Predicted for GSTCD Suggest Methyltransferase Activity*

Results across the protein prediction servers demonstrated many homologous protein sequences with homology focussed around the C-terminus of GSTCD, specifically concentrated in the region of the predicted methyltransferase domain. Therefore, a methyltransferase function could be inferred for GSTCD.

Methyltransferases have roles in a diverse range of biological processes including interacting with DNA, RNA and proteins, transferring a methyl group to one or more of these targets, generally from the donor S-adenosylmethionine (SAM) (Struck et al. 2012; Petrossian & Clarke 2011). These enzymes often function in epigenetic regulation of gene expression and have been implicated in a wide variety of disorders (Struck et al. 2012). In this project it was also identified that potential reduction in cell viability occurs in cells where GSTCD expression is reduced, which concurs with knock-down of several methyltransferases (Takada et al. 2007; Luka et al. 2008; Scott et al. 2014). Although, this decrease in cell growth by viability is not reflected by a reduction in cell proliferation as is observed in these methyltransferase knock-down experiments.

Moreover, remarkably consistent structure can be found in a large number of the main group of methyltransferases (Class I) despite high levels of variation in

protein sequence (can be as little as 10% similarity). Therefore suggesting that the structure required for catalytic activity involving SAM is highly changeable (Struck et al. 2012) and that GSTCD may have this function even with low sequence homology. This agrees with our knowledge that 3D structure of proteins from the same family is more conserved than the amino acid sequences (Fiser 2010).

Furthermore, the methyltransferase domain is the only domain to be consistently present in GSTCD across a wide range of species, as depicted in Figure 31. This further indicates the importance of the methyltransferase domain because it is maintained across a range of vertebrate and invertebrate species.

Methyltransferase activity may also be a more promising avenue for potential function as only one of the two typical GST enzyme domains is present in GSTCD (as identified by NCBI BLAST conserved domain search). The GST family of enzymes characteristically contain N-terminal (thioredoxin-type) and C-terminal (α-helical) GST domains which are necessary for function, with the active site predicted to be located between the two domains (NCBI Conserved Protein Domain Resource 2015). Therefore theoretically it appears unlikely that GSTCD would have the same function whilst lacking a second essential GST specific domain.

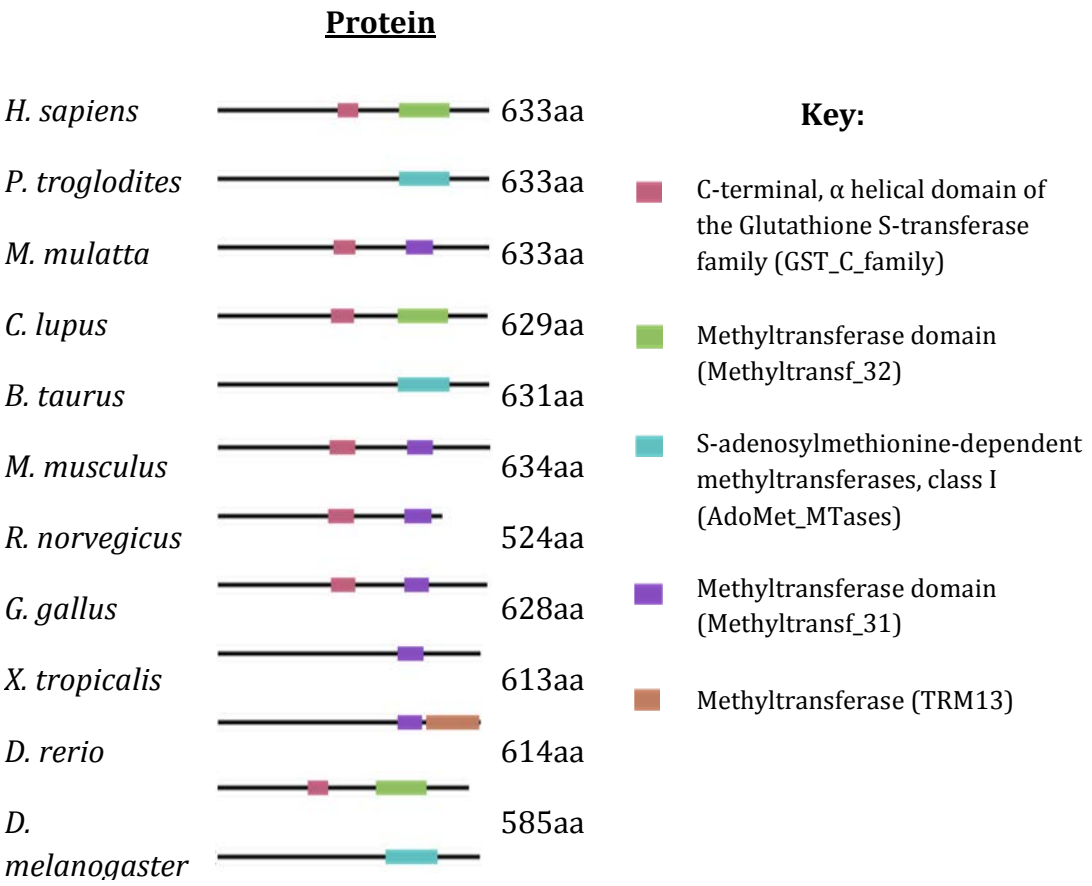



**Figure 31. Graphical representation of GSTCD conserved domain homology.** Results for GSTCD from the NCBI Homologene database comparing protein sequence and conserved domain architecture across common vertebrate and invertebrate species (adapted from: NCBI Homologene 11693). Primarily a type of methyltransferase domain is evident across species followed by GST C-terminal domain in a majority of species.

Overall, from these results and due to the wide ranging capabilities of methyltransferases the specific methyltransferase interaction that GSTCD may perform cannot yet be predicted. Alternatively, I-TASSER identified homology with transport receptors, involved in transporting proteins into the nucleus (Andrade et al. 2001). However, this could be considered a less credible theoretical function for GSTCD, based on previous evidence of general cytoplasmic localisation (Obeidat et al. 2013) as opposed to nuclear. In addition, no nuclear localisation sequences (NLS) are identified in either GSTCD variant by NLS prediction servers (NucPred - Brameier et al. 2007 and SeqNLS - Lin et al. 2012).

#### 4.3.4.3 Intrinsic Disorder Prediction Suggests Minimal Degree of Disorder in GSTCD Protein Structure

Phyre2 was the main method for identifying regions of intrinsic disorder as this server assessed the full length GSTCD sequence, in this instance two regions were detected in GSTCD for both variants, located approximately either side of the predicted GST domain. This suggests a lack of 3D structure in these regions, which generally allows a protein to exhibit structural polymorphism for binding to ligands (Wright & Dyson 2014). It has been determined that despite the majority of proteins requiring a pre-determined structure for efficient function, numerous proteins require intrinsic disorder for interactions (Dunker et al. 2002). Regions of intrinsic disorder typically mediate binding to diverse targets; hence often function as central hubs in signalling networks (Wright & Dyson 2014). Although, this does not necessarily correlate for GSTCD as no significant adverse effects to signalling pathways have been observed in Gstcd knock-out mice phenotypes. Whilst in this research intrinsic disorder may not provide definitive answers for function, it could indicate that GSTCD may be multifunctional or that a binding site may be located at the regions of intrinsic disorder.

#### 4.3.4.4 Assessment of Protein Prediction Servers

For the specific servers utilised here the main advantages include high CASP rankings, in particular I-TASSER was rated the best online protein prediction server in the last five CASP evaluations (CASP7-11) (J. Yang et al. 2014; Quashie et al. 2015). Phyre2 and SWISS-MODEL are also amongst the top performing servers in the CASP international blind trials. Phyre2 creators consider its main difference compared to some higher ranked servers to be that it is specifically designed with an emphasis on ease of use for non-experts in bioinformatics analysis (Kelley et al. 2015).

In general, high quality models can be obtained from comparative protein structure modelling methods such as these servers which are comparable with laboratory determined structures (for example medium resolution NMR and low

resolution X-ray crystallography) (Fiser 2010). Although, these approaches are limited to sequences which can be confidently mapped to known structures, described as ranging between 30-80% (Fiser 2010). Importantly, if homology cannot be detected then modelling results are presumed to be extremely unreliable, therefore emphasising the necessity of known structures for references as no reliable protein structure can be predicted entirely based on protein sequence (Kelley et al. 2015). Hence, the primary constraint for all automated protein modelling approaches is lack of available templates for these computational methods.

On the whole, as indicated by results shown here, several protein prediction servers with varied analysis methods can produce similar results for indicating possible protein function, enhancing the reliability of these results. Additionally, there is an ever increasing bank of protein structures for comparison in studies of proteins of unknown function. The leading repository of protein structures being the RCSB Protein Data Bank; collating information for over 104,000 entries (Rose et al. 2015).

In conclusion, it must be emphasised that this is preliminary research into GSTCD theoretical structure and function based on automated protein predictions. Despite similar results being obtained from several protein prediction servers here the theoretical modelling procedure is non-experimental, hence results must be considered with care. A specific point of note is that there is no human intervention during model building using an automated modelling server.

As the effectiveness of biological inferences from predicted protein models relies on the accuracy of prediction (Zhang 2009), it appears more detailed protein analysis is required for GSTCD functional classification. Future work would likely necessitate *ab initio* modelling to assess the quality of predicted models and manually curate a predicted protein structure or possibly X-ray crystallography as an experimental approach.

# 5 Discussion

Although there is currently little published data characterising GSTCD, many studies have indicated that this gene is important in the human respiratory system due to identification of several significant SNP associations in this gene in particular at locus 4q24. Where SNPs have been associated with lung function (Repapi et al. 2010; Hancock et al. 2010) and in particular with COPD (Soler Artigas, Wain, et al. 2011; Imboden et al. 2012; Li et al. 2013; Ong et al. 2013; Loth et al. 2014; Yao et al. 2014; Castaldi et al. 2011). The 4q24 locus is the site of a broad association signal with many SNPs associated with lung function which also encompass the genes: *INTS12*, *NPNT* and *FLJ20184*. Recently Wain et al. have reported two further independent signals of association at the 4q24 locus, which were located in the genes *NPNT* and *TET2* as opposed to *GSTCD* (Wain et al. 2015). This research further indicates the high level of complexity at this locus when deducing the origin and implications of these association signals amongst several genes. Additionally, the primary SNP in *GSTCD* previously associated with $FEV_1$ and COPD was not identified as an expression quantitative trait locus site for *GSTCD* in lung tissue (Obeidat et al. 2015).

Although, it has been reported that GSTCD expression levels vary in the lung across development (Obeidat et al. 2013) and *GSTCD* has also been associated with $FEV_1$ in children and with growth (Kreiner-Møller et al. 2014), therefore suggesting a potential role in lung formation and/or growth. Specifically, GSTCD expression was noted to be increased during the pseudoglandular stage of lung development (Obeidat et al. 2013), when branching morphogenesis begins and proximal airway epithelial and mesenchymal differentiation occurs (Bancalari & Polin 2012; Koopman 2010; Warburton et al. 2010) hence possibly indicating an important role for GSTCD in lung development.

So far GSTCD has been classed as a member of the GST family of enzymes involved in cellular detoxification. However, GST activity has not been detected in this protein. Therefore the purpose of this study was to further investigate the potential reasons behind the significant association of *GSTCD* with lung function and COPD. This was performed using knock-down and knock-out approaches in human cells and a mouse model respectively, to examine cellular and phenotypic effects caused by loss of this protein.

## 5.1 Conclusions

*GSTCD* was initially identified as a plausible candidate gene due to significant associations at locus 4q24 with lung function measures and COPD but the reasons behind these associations were not known.

In conclusion this study has provided novel and important contributions to a relatively unknown protein. Knock-down of GSTCD was demonstrated to result in reduced total cell numbers, reflecting some of the typical methyltransferase knock-down consequences identified in previous research. Therefore, suggesting a possible role for GSTCD in cell viability (potentially apoptosis) as opposed to proliferation. This may be relevant to COPD pathogenesis as increases in apoptosis in endothelial and alveolar epithelial cells from COPD patients have been detected, which have not been counterbalanced by an increase in proliferation of these structural cells (Demedts et al. 2006; Siganaki et al. 2010; Vijayan 2013). Hence, these changes can lead to destruction of the lung tissue and the development of emphysema (Demedts et al. 2006).

Additionally, analysis into a loss of function model for GSTCD in a murine model suggests that GSTCD does not have a critical role in mammalian cells as no gross adverse effects have so far been observed in Gstcd knock-out mice. However, more detailed analysis into cell growth in the Gstcd -/- mice is necessary to uncover potential subtle differences in cell viability.

Recently, investigation into potential GSTCD signalling pathways has been performed by RNA sequencing in GSTCD knock-down treated hBECs. RNA sequencing quantifies gene expression levels from RNA samples. Changes in gene expression in treatment conditions (for example siRNA knock-down) may indicate biologically associated proteins and infer potential signalling pathways. In preliminary analysis of GSTCD siRNA knock-down samples this has identified significant changes in expression of only 7 genes. This is important as this suggests a very limited active role for GSTCD in the cell which corresponds with the lack of phenotype evident in Gstcd knock-out mice. Although no GST or methyltransferase enzymes were identified in this initial analysis.

Finally, investigating GSTCD protein structure in this study has added to the value of the predicted methyltransferase domain in this protein sequence, as these results have also suggested homology primarily with methyltransferases. Yet the role of this potential methyltransferase activity for GSTCD in specific cells in the lung is presently unknown as methyltransferases interact with a wide range of molecules. Hence, it is challenging to elucidate a specific role that GSTCD may have with this activity in the lungs.

## 5.2 Limitations

Developing tools for protein expression analyses such as primary GSTCD antibodies was time consuming in this study and difficulties were encountered using the polyclonal antibodies as the detailed specificity of these is still unknown.

During this project low replicate numbers of Gstcd knock-out mice were available due to on-going colony amplification. In addition, it was not possible to extract lung tissue at University of Nottingham as the colony was based at our collaborating institution, Cornell University, America. Therefore initially there have been difficulties in standardising extraction and delivery of tissue. The mouse colony has since been imported to University of Nottingham where future work will involve a standard method of lung inflation and extraction to allow for improved comparison between genotypes. Additionally, no lacZ staining was performed in Gstcd +/+ mice as a definitive negative control, however there is evidence of specific staining due to internal negative controls in the majority of lung tissue shown.

## 5.3 Future Directions

One way to improve confidence in GSTCD antibody specificity would be to utilise a monoclonal GSTCD antibody. One monoclonal antibody option is currently commercially available through Abnova, although the primary limitation in this instance remains substantial cost.

Furthermore, investigation of GSCTD expression in a type II pneumocyte (alveolar epithelial) model such as A549 human cells may be of interest for comparison against the previously assessed lung specific cells (hBECs, human airway smooth muscle cells and whole lung tissue) (Obeidat et al. 2013). This would be of relevance based on the localisation pattern seen in specific cells across the alveolar tissue in GSTCD IHC results, which could be considered highly similar to the typical localisation pattern of alveolar epithelial type II cell marker Pro-surfactant protein C (ProSP-C).

In the future, as a significant reduction in total cell number was identified in GSTCD siRNA knock-down treated hBECs using a cell growth assay, it would be useful to image these cells to further assess whether general morphology is also affected between conditions. To further interrogate the potential reason behind the reduction in total cell number rates of apoptosis could be examined using an assay such as Apo-ONE® Homogeneous Caspase-3/7 Assay to measure enzyme markers of apoptosis.

During this research methyltransferase activity has become a strong possibility for GSTCD function but the potential activity of GSTCD has yet to be experimentally explored. Future work may involve studying isolated GSTCD protein in a methyltransferase activity assay, assessing the ability of the protein to transfer the biological co-factor SAM to various acceptor biomolecules. For example the EPIgeneous™ Methyltransferase Assay kit could be utilised, a universal solution for all methyltransferases and substrates which allows direct detection of the converted SAM product, S-adenosylhomocysteine (SAH). However, based on the lack of detailed information available for GSTCD localisation, function and interactions, experimental structural determination of GSTCD by X-ray crystallography or NMR spectroscopy may be the optimal

method for clearly identifying function or homologous proteins to infer function and active sites.

Moreover, since no notable differences have so far been observed in Gstcd knock-out mice, future exploration may include general characterisation of cell types within the lungs, to determine whether these are altered in the Gstcd -/- mice relative to Gstcd +/+ mice. This may focus on markers for the lung cell types where Gstcd appears to be localising based on IHC results. For example: bronchial epithelial cells (e.g. cytokeratin 19), alveolar epithelial type II cells (e.g. ProSP-C) and macrophages (e.g. CD68) (based on localisation pattern seen from IHC in mouse lung tissue). Finally, it would be pertinent to isolate lung cells from Gstcd -/- mice in future work to investigate whether effects cell growth and cell viability reflect results observed for GSTCD knock-down in human cells.

# 6 Appendix

# 6.1 Experimental Reagents

## 6.1.1 PCR

**Table 21. PCR reagents and volumes used for amplification of cDNA.**

| Reagent | Suppliers | Catalogue Number | X1 Sample (µl) |
|---|---|---|---|
| 10X PCR buffer | Bioline | BIO-21040 | 2.5 |
| $Mg^{2+}$ (50mM) | Bioline | BIO-21040 | 0.75 |
| dNTPs (5mM) | Bioline | BIO 39026 | 1 |
| GAPDH Forward | Sigma | - | 1.25 |
| GAPDH Reverse | Sigma | - | 1.25 |
| *Taq* DNA polymerase | Bioline | BIO-21040 | 0.5 |
| cDNA | - | - | 1 |
| Water | Baxter | UKF7114 | 16.75 |
| **Total** | **-** | **-** | **25** |

### 6.1.2 SDS-PAGE and Western Blotting

**Table 22. Buffer reagents and volumes used for SDS-PAGE gel electrophoresis and Western blotting of protein samples.**

| Buffer | Reagents | Supplier | Catalogue Number | Volume/Weight |
|---|---|---|---|---|
| **5X reducing gel loading buffer** | DTT | Sigma | D9779 | 1g |
| | 1M Tris-HCl (pH6.8) | Invitrogen | 15504-020 | 5.2ml |
| | Sodium dodecyl sulphate (SDS) | Sigma | L3771 | 1.3g |
| | Glycerol | Fisher | BPE299-1 | 6.5ml |
| | 10% Bromophenol blue | VWR | 44305 | 130µl |
| **10X Running buffer (1:10 dilution up to 1L used to create 1X working solution)** | Tris | Invitrogen | 15504-020 | 30.3g |
| | Glycine | Sigma | G-8898 | 144g |
| | Sodium dodecyl sulphate (SDS) | Sigma | L3771 | 20g |
| | Distilled sterile water | Baxter | UKF7114 | Up to 1L |
| **10X Transfer buffer (1:10 dilution up to 1L used to create 1X working solution)** | Tris | Invitrogen | 15504-020 | 30.3g |
| | Glycine | Sigma | G-8898 | 144g |
| | Distilled sterile water | Baxter | UKF7114 | Up to 1L |
| **PBS-T buffer** | Phosphate buffered saline (PBS) tablets | Oxoid | BR0014G | 10 tablets |
| | TWEEN® 20 | Sigma | P-7949 | 2ml |
| | Distilled sterile water | Baxter | UKF7114 | Make up to 1L |

| | | | | |
|---|---|---|---|---|
| **Blocking buffer** | PBS-T | (as above) | - | 100ml |
| | Marvel Dried Milk | Premier International Foods | - | 5g |
| **Developer solution** | Carestream® Kodak® autoradiography GBX developer/ replenisher | Sigma | P7042 | 200ml |
| | Distilled sterile water | Baxter | UKF7114 | 800ml |
| **Fixer solution** | Carestream® Kodak® autoradiography GBX fixer/ replenisher | Sigma | P7167 | 200ml |
| | Distilled sterile water | Baxter | UKF7114 | 800ml |

**Table 23. SDS PAGE gel reagents combined in sequence to produce two gel layers.** First the resolving gel was added to gel apparatus followed by a top layer of stacking gel.

| Buffer | Reagents | Supplier | Catalogue Number | Volume |
|---|---|---|---|---|
| **Resolving Gel** | Distilled sterile water | Baxter | UKF7114 | 4ml |
| | 30% (w/v) acrylamide | Sigma | A3699 | 3.3ml |
| | 10% Sodium dodecyl sulphate (SDS) | Sigma | L3771 | 100µl |
| | 1.5M Tris-HCl (pH8.8) | Invitrogen | 15504-020 | 2.5ml |
| | 10% Ammonium Persulphate (APS) | Promega | V3131 | 100µl |
| | N,N,N′,N′-Tetramethylethylenediamine (TEMED) | Sigma | T9281 | 4µl |
| **Stacking Gel** | Distilled sterile water | Baxter | UKF7114 | 3.4ml |
| | 30% (w/v) acrylamide | Sigma | A3699 | 830µl |
| | 10% Sodium dodecyl sulphate (SDS) | Sigma | L3771 | 50µl |
| | 1M Tris-HCl (pH6.8) | Invitrogen | 15504-020 | 630µl |
| | 10% Ammonium Persulphate (APS) | Promega | V3131 | 50µl |
| | N,N,N′,N′-Tetramethylethylenediamine (TEMED) | Sigma | T9281 | 5µl |

### 6.1.3 Immunohistochemistry

**Table 24. Buffer reagents and volumes used for IHC.**

| Buffer | Reagents | Supplier | Catalogue Number | Volume |
|---|---|---|---|---|
| **PBS** | Phosphate buffered saline (PBS) tablets | Oxoid | BR0014G | 10 tablets |
| | Distilled sterile water | Baxter | UKF7114 | Up to 1L |
| **PBS-BT** | Bovine serum albumin (BSA) (1%) | Millipore | 82-045-2 | 0.1g |
| | Triton™ X-100 (10%) | Sigma | 9002-93-1 | 100µl |
| | PBS | (As above) | - | Up to 10ml |
| **Citrate buffer** | Sodium citrate | Sigma | S4641 | 1.47g |
| | Distilled sterile water | Baxter | UKF7114 | Up to 500ml |

## 6.2 Sequence Alignments

### 6.2.1 GSTCD Antibody Immunogen Alignments

Immunogen regions of the two GSTCD antibodies utilised in this research are shown aligned to GSTCD variant 1 protein sequence to demonstrate location of predicted binding. Sequences were assessed using the DNA alignment server, Clustal Omega - Multiple Alignment tool (1.2.1) (European Molecular Biology Laboratory - European Bioinformatics suite (EMBL-EBI)). Abnova antibody immunogen is designed against full length GSTCD protein and Proteintech antibody immunogen is homologous to the region of 283aa - 633aa of GSTCD. A single amino acid is absent in both antibody immunogens compared to GSTCD (position 330aa), believed to be due to the GSTCD sequence being updated after antibody design and production.

**Key**

hGSTCDv1: human GSTCD variant 1 protein sequence

Abnova: Abnova GSTCD antibody immunogen protein sequence

Proteintech: Proteintech GSTCD antibody immunogen protein sequence

```
hGSTCDv1    MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSRDS  60
Abnova      MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSRDS  60
Proteintech ------------------------------------------------------------  0

hGSTCDv1    SLLRDDLIQDVEIQIISRQELPPIVQNCCLPAVVERSDNFCRAGLAVVLRHIIQKSYEAD  120
Abnova      SLLRDDLIQDVEIQIISRQELPPIVQNCCLPAVVERSDNFCRAGLAVVLRHIIQKSYEAD  120
Proteintech ------------------------------------------------------------  0

hGSTCDv1    PLKKELLELLGFKKTCLKACAEVSQWTRLCELTIPLAIENFLRESSDQPPTIPVEILQLE  180
Abnova      PLKKELLELLGFKKTCLKACAEVSQWTRLCELTIPLAIENFLRESSDQPPTIPVEILQLE  180
Proteintech ------------------------------------------------------------  0

hGSTCDv1    KKLSEPVRVHNDDKLRRQKLKQQKADGVGPPLTKGKAKSKVHTQETSEGLDSSSKSLELK  240
Abnova      KKLSEPVRVHNDDKLRRQKLKQQKADGVGPPLTKGKAKSKVHTQETSEGLDSSSKSLELK  240
Proteintech ------------------------------------------------------------  0

hGSTCDv1    VAFSKLTVQEEPATTNREPSHIRKAKASDLPPLEHVFAEGLYFTLADIVLLPCIHHFLVI  300
Abnova      VAFSKLTVQEEPATTNREPSHIRKAKASDLPPLEHVFAEGLYFTLADIVLLPCIHHFLVI  300
Proteintech ----------------------------------------FTLADIVLLPCIHHFLVI  18

hGSTCDv1    ISRKFSEKLVEFPLLASWYQRIQEVPGVKTAASKCGIQFLHLPKLLTTSTEQHPNLCEVP  360
Abnova      ISRKFSEKLVEFPLLASWYQRIQEVPGVKT-ASKCGIQFLHLPKLLTTSTEQHPNLCEVP  359
Proteintech ISRKFSEKLVEFPLLASWYQRIQEVPGVKT-ASKCGIQFLHLPKLLTTSTEQHPNLCEVP  77

hGSTCDv1    GVEEQSDPLFIGGPRPTMAKLMEKGIEVMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDR  420
Abnova      GVEEQSDPLFIGGPRPTMAKLMEKGIEVMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDR  419
Proteintech GVEEQSDPLFIGGPRPTMAKLMEKGIEVMFSPHPCPTWTLDWNVLPAAVSPKEGKMSSDR  137

hGSTCDv1    ALRKQQQLNNLVYVVTNQAKPGDRIVDFCSGGGHVGIVLAHMLPSCQVTLIENKELSLIR  480
Abnova      ALRKQQQLNNLVYVVTNQAKPGDRIVDFCSGGGHVGIVLAHMLPSCQVTLIENKELSLIR  479
Proteintech ALRKQQQLNNLVYVVTNQAKPGDRIVDFCSGGGHVGIVLAHMLPSCQVTLIENKELSLIR  197

hGSTCDv1    AKKRSDELGLSNIWFIQANMEYFTGMFNIGVALHACGVATDMVIEHCIKTRASFVTCPCC  540
Abnova      AKKRSDELGLSNIWFIQANMEYFTGMFNIGVALHACGVATDMVIEHCIKTRASFVTCPCC  539
Proteintech AKKRSDELGLSNIWFIQANMEYFTGMFNIGVALHACGVATDMVIEHCIKTRASFVTCPCC  257

hGSTCDv1    YGFIQNTSKFNFPKSEQFKKTLSYKEHMILCRFADQTAVQLPPQRRLIGKQCMCLVDLDR  600
Abnova      YGFIQNTSKFNFPKSEQFKKTLSYKEHMILCRFADQTAVQLPPQRRLIGKQCMCLVDLDR  599
Proteintech YGFIQNTSKFNFPKSEQFKKTLSYKEHMILCRFADQTAVQLPPQRRLIGKQCMCLVDLDR  317

hGSTCDv1    ARAAEECGYSVQVISMEPESCSPKNNMIVGVPI  633
Abnova      ARAAEECGYSVQVISMEPESCSPKNNMIVGVPI  632
Proteintech ARAAEECGYSVQVISMEPESCSPKNNMIVGVPI  350
```

### 6.2.2 GSTCD Sequence Alignment to GST tag

Whilst optimising GSTCD antibodies in hBEC protein in Western blotting a recombinant GSTCD protein was used as a positive control. This recombinant protein included a GST tag, used to isolate the protein during production. Therefore using a recombinant protein including this tag could be considered inappropriate for use with GSTCD antibodies, as this could produce a false positive signal. However, the GST tag and GSTCD protein sequences were assessed using the DNA alignment server, Clustal Omega - Multiple Alignment tool (1.2.1) (EMBL-EBI)) which revealed notably low levels of homology (percentage homology to GST tag: GSTCD variant 1 - 20% and GSTCD variant 2 - 21%), with no extended regions of direct sequence homology. This low level of alignment suggests that GSTCD targeted antibodies would be unlikely to bind to GST proteins.

**Key**

hGSTCDv1: human GSTCD variant 1 protein sequence

GST-tag: sequence of the human GST protein, used to isolate recombinant GSTCD during manufacturing production

* indicates positions which have a single, fully conserved amino acid

: indicates conservation between amino acids of strongly similar properties

. indicates conservation between amino acids of weakly similar properties

```
hGSTCDv1  MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTKEVSRDS  60
hGSTCDv2  MKAIKKSLTEEEYLYLDFSHQTEGCIFPLHTSVTLFLLSYCDCKIFKICLVVTK------  54
GST_tag   ------------------------------------------------------------  0


hGSTCDv1  SLLRDDLIQDVEIQIISRQELPPIVQNCCLPAVVERSDNFCRAGLAVVLRHIIQKSYEAD  120
hGSTCDv2  ------------------------------------------------------------  54
GST_tag   ------------------------------------------------------------  0


hGSTCDv1  PLKKELLELLGFKKTCLKACAEVSQWTRLCELTIPLAIENFLRESSDQPPTIPVEILQLE  180
hGSTCDv2  --------------------EVSQWTRLCELTIPLAIENFLRESSDQPPTIPVEILQLE  93
GST_tag   -----------------MESPILGYWKIK--------------GLVQPTRLLLEY--LE  26
                             :. *.                .  **  : :*   **

hGSTCDv1  KKLSEPVRV-HNDDKLRRQKLKQQKADGVGPPLTKGKAKSKV---------HTQETSEGL  230
hGSTCDv2  KKLSEPVRV-HNDDKLRRQKLKQQKADGVGPPLTKGKAKSKV---------HTQETSEGL  143
GST_tag   EKYEEHLYERDEGDKWRNKKFELGLEFPNLPYYIDGDVKLTQSMAIIRYIADKHNMLGGC  86
          :* .* :   .: ** *.:*::        *   .*..* .          ..::   *

hGSTCDv1  D--------SSSKSLELKVAFSKLTVQEEPATTNREPSHIRKAKASDLPPLEHVFAEGL-  281
hGSTCDv2  D--------SSSKSLELKVAFSKLTVQEEPATTNREPSHIRKAKASDLPPLEHVFAEGL-  194
GST_tag   PKERAEISMLEGAVLDIRYGVSRIAYSK-------DFETLKVDFLSKLPEMLKMFEDRLC  139
                  ..  *::: ..*::: .:       : . ::    *.** : ::* : *

hGSTCDv1  --------YFTLADIVLLPCIHHFLVIISRKFSEKLVEFPLLASWYQRIQEVPGVKTAAS  333
hGSTCDv2  --------YFTLADIVLLPCIHHFLVIISRKFSEKLVEFPLLASWYQRIQEVPGVKTAAS  246
GST_tag   HKTYLNGDHVTHPDFMLYDALDVVLY----MDPMCLDAFPKLVCFKKRIEAIPQIDKYLK  195
                 :.*  *::*  .:. .*          *  ** *..: :**: :* :..  .

hGSTCDv1  KCGIQFLHLPKLLTTSTEQHPNLCEVPGVEEQSDPLFIGGPRPTMAKLMEKGIEVMFSPH  393
hGSTCDv2  KCGIQFLHLPKLLTTSTEQHPNLCEVPGVEEQSDPLFIGGPRPTMAKLMEKGIEVMFSPH  306
GST_tag   SSK--YIAWPL---------------Q----GWQATFGGGDHPPKSD-----LEVLFQ--  227
          ..   ::  *                      :  * ** :*  :.     :**:*.

hGSTCDv1  PCPTWTLDWNVLPAAVSPKEGKMSSDRALRKQQQLNNLVYVVTNQAKPGDRIVDFCSGGG  453
hGSTCDv2  PCPTWTLDWNVLPAAVSPKEGKMSSDRALRKQQQLNNLVYVVTNQAKPGDRIVDFCSGGG  366
GST_tag   --------------------------------GPLETSLYKKAGT---------------  240
                                          *:. :*  :

hGSTCDv1  HVGIVLAHMLPSCQVTLIENKELSLIRAKKRSDELGLSNIWFIQANMEYFTGMFNIGVAL  513
hGSTCDv2  HVGIVLAHMLPSCQVTLIENKELSLIRAKKRSDELGLSNIWFIQANMEYFTGMFNIGVAL  426
GST_tag   ------------------------------------------------------------  240


hGSTCDv1  HACGVATDMVIEHCIKTRASFVTCPCCYGFIQNTSKFNFPKSEQFKKTLSYKEHMILCRF  573
hGSTCDv2  HACGVATDMVIEHCIKTRASFVTCPCCYGFIQNTSKFNFPKSEQFKKTLSYKEHMILCRF  486
GST_tag   ------------------------------------------------------------  240


hGSTCDv1  ADQTAVQLPPQRRLIGKQCMCLVDLDRARAAEECGYSVQVISMEPESCSPKNNMIVGVPI  633
hGSTCDv2  ADQTAVQLPPQRRLIGKQCMCLVDLDRARAAEECGYSVQVISMEPESCSPKNNMIVGVPI  546
GST_tag   ------------------------------------------------------------  240
```

### 6.2.3 GSTCD siRNA Sequence Alignments

Locations of the siRNA sequences used to knock-down expression of GSTCD are shown demonstrated by using the DNA alignment server, Clustal Omega - Multiple Alignment tool (EMBL-EBI). All siRNAs align to both published transcripts of human GSTCD; hence translation of both transcripts is theoretically reduced.

**Key**

hGSTCDv[1]: human GSTCD variant 1

GSTCD siRNA sequences:

siRNA A: AGAAUGUGGAUACUCCGUUCAAGTG (exon 12)

siRNA B: GGAUGUUUAAUAUUGGAGUAGCATT (exon 8/9)

siRNA C: GCCUGUUAGAGUGCAUAAUGAUGAT (exon 3)

Start codon

Stop codon

Exon boundary

### 6.2.3.1 Human GSTCD variant 1

```
hGSTCDv1    AGATGATTCCCTGATTCTCCAGAGAGATTACACACTTCGTTTGTGGCTAAGGTGACCCAA    60
hGSTCDv1    TGAAAGAAGAAAATGAAAGCCATAAAGAAAAGTCTTACAGAAGAAGAATACCTGTACCTG    120
hGSTCDv1    GACTTTTCTCACCAAACAGAAGGATGCATCTTTCCTCTTCATACATCTGTAACTTTATTT    180
hGSTCDv1    CTGTTATCTTACTGTGACTGTAAAATCTTTAAAATTTGCTTAGTTGTCACCAAAGAGGTG    240
hGSTCDv1    AGTAGAGATAGTTCACTACTAAGAGATGACCTGATCCAGGATGTTGAAATACAGATTATT    300
hGSTCDv1    TCAAGGCAGGAGCTCCCACCAATAGTCCAAAATTGCTGTTTGCCTGCAGTAGTAGAACGA    360
hGSTCDv1    TCAGACAATTTTTGTAGAGCAGGACTTGCTGTTGTATTGAGACACATAATCCAGAAATCC    420
hGSTCDv1    TATGAAGCAGACCCCTTAAAGAAGGAACTTTTGGAACTTCTGGGCTTTAAAAAGACTTGC    480
hGSTCDv1    TTGAAAGCCTGTGCTGAAGTTAGTCAGTGGACCAGGCTATGTGAACTCACCATCCCTTTG    540
hGSTCDv1    GCTATTGAGAATTTTCTCAGAGAATCTTCTGACCAGCCCCCAACTATACCTGTAGAAATA    600
hGSTCDv1    CTACAGCTAGAGAAAAAGCTTAGTGAGCCTGTTAGAGTGCATAATGATGATAAACTCCGC    660
hGSTCDv1    AGGCAGAAGCTCAAGCAACAGAAGGCTGATGGAGTTGGGCCTCCCCTTACTAAGGGAAAG    720
hGSTCDv1    GCAAAGAGCAAGGTCCACACACAGGAAACATCTGAAGGGTTGGATTCTTCATCCAAGAGT    780
hGSTCDv1    CTGGAACTGAAAGTGGCATTCTCAAAGCTCACAGTACAGGAAGAACCAGCTACTACCAAC    840
hGSTCDv1    AGAGAGCCTTCTCACATCAGAAAAGCAAAAGCCTCCGACCTTCCACCTCTGGAGCATGTG    900
hGSTCDv1    TTTGCAGAAGGGCTTTACTTCACTCTGGCAGATATTGTGCTCTTGCCCTGTATCCATCAT    960
hGSTCDv1    TTCTTGGTAATTATCAGCAGGAAATTTTCTGAGAAGCTGGTAGAATTTCCATTGCTAGCC    1020
hGSTCDv1    TCTTGGTACCAGAGGATTCAGGAAGTGCCAGGAGTAAAAACAGCAGCTTCTAAGTGTGGG    1080
hGSTCDv1    ATCCAATTTCTCCATTTACCAAAGTTGTTGACAACCTCAACTGAACAGCATCCAAACTTA    1140
hGSTCDv1    TGTGAAGTCCCAGGTGTAGAAGAGCAAAGCGATCCTTTATTTATAGGAGGACCAAGACCA    1200
hGSTCDv1    ACCATGGCCAAGTTAATGGAAAAGGGCATTGAAGTGATGTTTTCTCCCCACCCTTGCCCT    1260
hGSTCDv1    ACTTGGACTCTTGATTGGAATGTTCTCCCTGCAGCAGTCAGCCCAAAGGAAGGTAAAATG    1320
hGSTCDv1    TCCAGTGATCGAGCTTTGAGGAAGCAGCAACAGTTGAACAACCTTGTCTATGTGGTAACA    1380
hGSTCDv1    AATCAGGCCAAACCTGGTGACAGAATTGTGGATTTCTGCAGCGGTGGGGGCCATGTTGGA    1440
hGSTCDv1    ATTGTCCTTGCTCACATGCTGCCATCATGTCAGGTTACATTAATAGAAAACAAGGAATTA    1500
hGSTCDv1    TCATTAATTCGTGCTAAGAAGAGAAGTGATGAACTGGGTTTAAGCAACATTTGGTTCATT    1560
hGSTCDv1    CAAGCAAATATGGAATATTTTACTGGGATGTTTAATATTGGAGTAGCATTGCATGCTTGT    1620
hGSTCDv1    GGAGTGGCAACAGACATGGTGATTGAGCACTGTATCAAAACACGGGCTTCCTTCGTCACA    1680
hGSTCDv1    TGCCCTTGCTGTTATGGTTTCATTCAGAACACCTCAAAGTTCAATTTTCCAAAAAGTGAA    1740
hGSTCDv1    CAATTCAAGAAAACTTTATCATACAAGGAACACATGATTCTGTGCAGATTTGCAGACCAG    1800
hGSTCDv1    ACAGCTGTCCAGCTCCCACCCCAACGAAGGCTCATAGGAAAACAGTGCATGTGCTTGGTG    1860
hGSTCDv1    GATCTGGATCGAGCAAGAGCTGCAGAAGAATGTGGATACTCCGTTCAAGTGATATCCATG    1920
```

```
hGSTCDv1   GAGCCAGAGAGCTGCTCTCCCAAAAATAACATGATTGTGGGAGTCCCCATTTAAAATGAG   1980
hGSTCDv1   ATATTCACATTTGATATTTTGCCATCCGTCTTCACGTTGGATGCCCAGTGGCATTCAGGA   2040
hGSTCDv1   GGTTCTTGGCATAACTAGGAAACAGCATTAGCCATCTTGAACCTATTGTGCTCAGGAAGG   2100
hGSTCDv1   AAAGCAACAGGGAAATCTTGGAAGTAAAGGCTCCCTGCAAAGCATCACAAATGCTTTACA   2160
hGSTCDv1   ATGTGTGATTAAAGGACTGCTGGTTTTTATAGTGAGAATCCCCTGAAGTCTCAGCTGCCA   2220
hGSTCDv1   AAAGAATAATACTAGTGGAGGTTCCATCACAGGAATAACAATTTCCTTCTCACTTCACAA   2280
hGSTCDv1   GCTCCTCTGATCTTGCCAATGTGATGTTTAATTCAAAACAGCGCACTTACAGCCTTTATT   2340
hGSTCDv1   TTATACTTATTTAGATATTCACATGGGAACTGAAAGTAGACTCAGTGGAGTTTTTTGTTT   2400
hGSTCDv1   TTTATATTTTTAACAATTGTACTTTTTCTATCACTTTAAAAAAATCTAGTCAGGTAATTA   2460
hGSTCDv1   TAGTACAGTTATTTTTAAGCTGGGGTTAGTTGAACTTGTATAGCATTTAATATACTCCTT   2520
hGSTCDv1   TTAAATATATTTCTGCCAGGCATGGTGGCTCATGCCTATAATCCCAGCACTTTGAGAGGC   2580
hGSTCDv1   CAAGGTGGGAGGATCATTTGAGCCCAGGATATCGAGACCAGCCTGCACTACAAAAAAATC   2640
hGSTCDv1   CAAAAAATGAGGTGGGAGGATCACTTGAGCCCAGGAAGTCAAGGCTGCAGTGAGCCTTGA   2700
hGSTCDv1   TAGCGCCACTGCACTCCTGCCTGGGAAAATGAGACCCAGACTCAAAATAAATCAATGAAT   2760
hGSTCDv1   GAATTAATAAACAAATAAATAAATAAATAAATTTCCCATAAATTACCAAATGAGATTGTC   2820
hGSTCDv1   ACTGGTTCAAATTATTAAATAATTGCCATAAATTTCCTATTACGGGAATGTGGAATTTGG   2880
hGSTCDv1   ACATACATTTTAACAATAAGTACATTGATTAATACATTTAAGTTTATAAGTAATTATTGA   2940
hGSTCDv1   GGATCCAGTATTTTCAAGACCTTGGCCTAGTTTCTTTTCAATTCAAAGATGAGTTGCAAA   3000
hGSTCDv1   AGACATAATGTTTTGCCTATGGAAGTAGATCATCTAGAAGAAAAAGATTAAAGAATTAGT   3060
hGSTCDv1   TAAGCTTTTTTTTTTTTTTTTTTTGAGACAGAGTTTCACTCTTGTTGCCCAGGCTGGAA   3120
hGSTCDv1   TTCAGTGGCTCGATCTTGGCTCACCACAACCTCTGCATTCTCCTGCCTCAGCCTCCCGAG   3180
hGSTCDv1   TAGCTAGGATTATAGGCACCTGCCACCATGCCCGGCTAATTTTGTTTGTAATTTTAGTAG   3240
hGSTCDv1   AGATGGGGTTTCTCCATGTTGGCCAGGCTGGTTTCGAACTCCTGACCTCAGGTGATCTAC   3300
hGSTCDv1   CCGCCTCAGCCTCCCAAAGTGCTGAAATTACAGGCGTGAGCCACCTCACCTGGCCATGTT   3360
hGSTCDv1   CTTTATGGAATACTACAATGTGTTGTCACAGGATTTGTTATTCAGATGTCAATCATTTTT   3420
hGSTCDv1   ATTTATATATTTCTGGGTAAAACTTATTAGTGGTGTGAGGAGTGCAGAATATGGAAGCTT   3480
hGSTCDv1   ACCTAAATTCTTTTTCTGCCAGCCAACATTTTAAAAGACAAATGATGCATCCTTTCTCAA   3540
hGSTCDv1   TGCTGCTAATTCTTTAAATTTTACATGTCCCAATTTATTCCATTTGAGATTTATTTTAAT   3600
hGSTCDv1   GTTTTCAACCCCTGTGTAACATGTATCATGAGGAGAAAGCACTATCATTGATCATCACTA   3660
hGSTCDv1   TGGATTAAAATATAGGAATATTTTACAGGGCAGTATGGGTGGGGGTGGGAGACATATTGG   3720
hGSTCDv1   AATCGTGTTTTCTGGGATAAAATTTTTAAAAAGAATTAATCATTAGCCCTGAGTATTTTA   3780
hGSTCDv1   GTCATTTGTGCCTATAGAAAACAACATGCTCTTTTATTGTGGACTTTCTTCTTTTAAGTG   3840
hGSTCDv1   AGGTTTAATTTATATACAATACAATAACAGATTTTTAAGTATTCATTTGAGATGTTTTGA   3900
hGSTCDv1   CAACTGTATATATACATGTAACCATCACTGGAATCAGATATAGAATATTCCCCTTAAATA   3960
```

```
hGSTCDv1    TTGTCAGGTAAACACATTGAGAATATTTTTTCCAGTCTGTGGCTTACCTACTTATTTTCA    4020
hGSTCDv1    TAATGAATGTTTTACTTGTGATGAAGATTTAACTTACCCTTTTACTTTAATGATTAATAT    4080
hGSTCDv1    TTTCTGTGTTCTTTCTAATAAATTTTACGAACTCTGAGGTTA                      4122
```

### 6.2.3.2 *Human GSTCD variant 2*

```
hGSTCDv2    GTTACTTCCCTTTTTCCGGTCCGCCGGATTATGAATGACGGCCGGCGCGAGTATTTTCCA    60
hGSTCDv2    CATAAGGTGGCTGTCGTTTTTCTCCTGGCGTCTGTGGAGGCGAGTGGTCTGCGGGCAGCA    120
hGSTCDv2    GCTCCCAGAGGCAGCCTTGGAATTCCAGCTCGGACTGGGCGGGAAGGCGCAGGCGGCCCA    180
hGSTCDv2    GGTCGCCGACACGCTCACGCACCCTCCCTGCCTGGCCGCGCCTCTGCGACCAGGTGACCC    240
hGSTCDv2    AATGAAAGAAGAAAATGAAAGCCATAAAGAAAAGTCTTACAGAAGAAGAATACCTGTACC    300
hGSTCDv2    TGGACTTTTCTCACCAAACAGAAGGATGCATCTTTCCTCTTCATACATCTGTAACTTTAT    360
hGSTCDv2    TTCTGTTATCTTACTGTGACTGTAAAATCTTTAAAATTTGCTTAGTTGTCACCAAAGAGG    420
hGSTCDv2    TTAGTCAGTGGACCAGGCTATGTGAACTCACCATCCCTTTGGCTATTGAGAATTTTCTCA    480
hGSTCDv2    GAGAATCTTCTGACCAGCCCCCAACTATACCTGTAGAAATACTACAGCTAGAGAAAAAGC    540
hGSTCDv2    TTAGTGAGCCTGTTAGAGTGCATAATGATGATAAACTCCGCAGGCAGAAGCTCAAGCAAC    600
hGSTCDv2    AGAAGGCTGATGGAGTTGGGCCTCCCCTTACTAAGGGAAAGGCAAAGAGCAAGGTCCACA    660
hGSTCDv2    CACAGGAAACATCTGAAGGGTTGGATTCTTCATCCAAGAGTCTGGAACTGAAAGTGGCAT    720
hGSTCDv2    TCTCAAAGCTCACAGTACAGGAAGAACCAGCTACTACCAACAGAGAGCCTTCTCACATCA    780
hGSTCDv2    GAAAAGCAAAAGCCTCCGACCTTCCACCTCTGGAGCATGTGTTTGCAGAAGGGCTTTACT    840
hGSTCDv2    TCACTCTGGCAGATATTGTGCTCTTGCCCTGTATCCATCATTTCTTGGTAATTATCAGCA    900
hGSTCDv2    GGAAATTTTCTGAGAAGCTGGTAGAATTTCCATTGCTAGCCTCTTGGTACCAGAGGATTC    960
hGSTCDv2    AGGAAGTGCCAGGAGTAAAAACAGCAGCTTCTAAGTGTGGGATCCAATTTCTCCATTTAC    1020
hGSTCDv2    CAAAGTTGTTGACAACCTCAACTGAACAGCATCCAAACTTATGTGAAGTCCCAGGTGTAG    1080
hGSTCDv2    AAGAGCAAAGCGATCCTTTATTTATAGGAGGACCAAGACCAACCATGGCCAAGTTAATGG    1140
hGSTCDv2    AAAAGGGCATTGAAGTGATGTTTTCTCCCCACCCTTGCCCTACTTGGACTCTTGATTGGA    1200
hGSTCDv2    ATGTTCTCCCTGCAGCAGTCAGCCCAAAGGAAGGTAAAATGTCCAGTGATCGAGCTTTGA    1260
hGSTCDv2    GGAAGCAGCAACAGTTGAACAACCTTGTCTATGTGGTAACAAATCAGGCCAAACCTGGTG    1320
hGSTCDv2    ACAGAATTGTGGATTTCTGCAGCGGTGGGGGCCATGTTGGAATTGTCCTTGCTCACATGC    1380
hGSTCDv2    TGCCATCATGTCAGGTTACATTAATAGAAAACAAGGAATTATCATTAATTCGTGCTAAGA    1440
hGSTCDv2    AGAGAAGTGATGAACTGGGTTTAAGCAACATTTGGTTCATTCAAGCAAATATGGAATATT    1500
hGSTCDv2    TTACTGGGATGTTTAATATTGGAGTAGCATTGCATGCTTGTGGAGTGGCAACAGACATGG    1560
hGSTCDv2    TGATTGAGCACTGTATCAAAACACGGGCTTCCTTCGTCACATGCCCTTGCTGTTATGGTT    1620
```

```
hGSTCDv2    TCATTCAGAACACCTCAAAGTTCAATTTTCCAAAAAGTGAACAATTCAAGAAAACTTTAT    1680
hGSTCDv2    CATACAAGGAACACATGATTCTGTGCAGATTTGCAGACCAGACAGCTGTCCAGCTCCCAC    1740
hGSTCDv2    CCCAACGAAGGCTCATAGGAAAACAGTGCATGTGCTTGGTGGATCTGGATCGAGCAAGAG    1800
hGSTCDv2    CTGCAGAAGAATGTGGATACTCCGTTCAAGTGATATCCATGGAGCCAGAGAGCTGCTCTC    1860
hGSTCDv2    CCAAAAATAACATGATTGTGGGAGTCCCCATTTAAAATGAGATATTCACATTTGATATTT    1920
hGSTCDv2    TGCCATCCGTCTTCACGTTGGATGCCCAGTGGCATTCAGGAGGTTCTTGGCATAACTAGG    1980
hGSTCDv2    AAACAGCATTAGCCATCTTGAACCTATTGTGCTCAGGAAGGAAAGCAACAGGGAAATCTT    2040
hGSTCDv2    GGAAGTAAAGGCTCCCTGCAAAGCATCACAAATGCTTTACAATGTGTGATTAAAGGACTG    2100
hGSTCDv2    CTGGTTTTTATAGTGAGAATCCCCTGAAGTCTCAGCTGCCAAAAGAATAATACTAGTGGA    2160
hGSTCDv2    GGTTCCATCACAGGAATAACAATTTCCTTCTCACTTCACAAGCTCCTCTGATCTTGCCAA    2220
hGSTCDv2    TGTGATGTTTAATTCAAAACAGCGCACTTACAGCCTTTATTTTATACTTATTTAGATATT    2280
hGSTCDv2    CACATGGGAACTGAAAGTAGACTCAGTGGAGTTTTTTGTTTTTTATATTTTTAACAATTG    2340
hGSTCDv2    TACTTTTTCTATCACTTTAAAAAAATCTAGTCAGGTAATTATAGTACAGTTATTTTTAAG    2400
hGSTCDv2    CTGGGGTTAGTTGAACTTGTATAGCATTTAATATACTCCTTTTAAATATATTTCTGCCAG    2460
hGSTCDv2    GCATGGTGGCTCATGCCTATAATCCCAGCACTTTGAGAGGCCAAGGTGGGAGGATCATTT    2520
hGSTCDv2    GAGCCCAGGATATCGAGACCAGCCTGCACTACAAAAAAATCCAAAAAAATGAGGTGGGAGG    2580
hGSTCDv2    ATCACTTGAGCCCAGGAAGTCAAGGCTGCAGTGAGCCTTGATAGCGCCACTGCACTCCTG    2640
hGSTCDv2    CCTGGGAAAATGAGACCCAGACTCAAAATAAATCAATGAATGAATTAATAAACAAATAAA    2700
hGSTCDv2    TAAATAAATAAATTTCCCATAAATTACCAAATGAGATTGTCACTGGTTCAAATTATTAAA    2760
hGSTCDv2    TAATTGCCATAAATTTCCTATTACGGGAATGTGGAATTTGGACATACATTTTAACAATAA    2820
hGSTCDv2    GTACATTGATTAATACATTTAAGTTTATAAGTAATTATTGAGGATCCAGTATTTTCAAGA    2880
hGSTCDv2    CCTTGGCCTAGTTTCTTTTCAATTCAAAGATGAGTTGCAAAAGACATAATGTTTTGCCTA    2940
hGSTCDv2    TGGAAGTAGATCATCTAGAAGAAAAAGATTAAAGAATTAGTTAAGCTTTTTTTTTTTTT    3000
hGSTCDv2    TTTTTTGAGACAGAGTTTCACTCTTGTTGCCCAGGCTGGAATTCAGTGGCTCGATCTTGG    3060
hGSTCDv2    CTCACCACAACCTCTGCATTCTCCTGCCTCAGCCTCCCGAGTAGCTAGGATTATAGGCAC    3120
hGSTCDv2    CTGCCACCATGCCCGGCTAATTTTGTTTGTAATTTTAGTAGAGATGGGGTTTCTCCATGT    3180
hGSTCDv2    TGGCCAGGCTGGTTTCGAACTCCTGACCTCAGGTGATCTACCCGCCTCAGCCTCCCAAAG    3240
hGSTCDv2    TGCTGAAATTACAGGCGTGAGCCACCTCACCTGGCCATGTTCTTTATGGAATACTACAAT    3300
hGSTCDv2    GTGTTGTCACAGGATTTGTTATTCAGATGTCAATCATTTTTATTTATATATTTCTGGGTA    3360
hGSTCDv2    AAACTTATTAGTGGTGTGAGGAGTGCAGAATATGGAAGCTTACCTAAATTCTTTTTCTGC    3420
hGSTCDv2    CAGCCAACATTTTAAAAGACAAATGATGCATCCTTTCTCAATGCTGCTAATTCTTTAAAT    3480
hGSTCDv2    TTTACATGTCCCAATTTATTCCATTTGAGATTTATTTTAATGTTTTCAACCCCTGTGTAA    3540
hGSTCDv2    CATGTATCATGAGGAGAAAGCACTATCATTGATCATCACTATGGATTAAAATATAGGAAT    3600
hGSTCDv2    ATTTTACAGGGCAGTATGGGTGGGGGTGGGAGACATATTGGAATCGTGTTTTCTGGGATA    3660
```

```
hGSTCDv2    AAATTTTTAAAAAGAATTAATCATTAGCCCTGAGTATTTTAGTCATTTGTGCCTATAGAA    3720
hGSTCDv2    AACAACATGCTCTTTTATTGTGGACTTTCTTCTTTTAAGTGAGGTTTAATTTATATACAA    3780
hGSTCDv2    TACAATAACAGATTTTTAAGTATTCATTTGAGATGTTTTGACAACTGTATATATACATGT    3840
hGSTCDv2    AACCATCACTGGAATCAGATATAGAATATTCCCCTTAAATATTGTCAGGTAAACACATTG    3900
hGSTCDv2    AGAATATTTTTTCCAGTCTGTGGCTTACCTACTTATTTTCATAATGAATGTTTTACTTGT    3960
hGSTCDv2    GATGAAGATTTAACTTACCCTTTTACTTTAATGATTAATATTTTCTGTGTTCTTTCTAAT    4020
hGSTCDv2    AAATTTTACGAACTCTGAGGTTA       4043
```

## 6.3 Additional GSTCD Localisation in Mouse Lung Tissue

Immunostaining results performed by collaborators at Cornell University using an alternative GSTCD antibody (Santa Cruz, sc-138332) (Figure 32) concur with localised staining achieved in this project with Proteintech (17502-1-AP) GSTCD antibody in wild-type (C57BL/6) mouse lung tissue (Results section 4.2.2 Figure 22) and with lacZ staining in Gstcd knock-out lung tissue Results section 4.2.1 Figure 21).

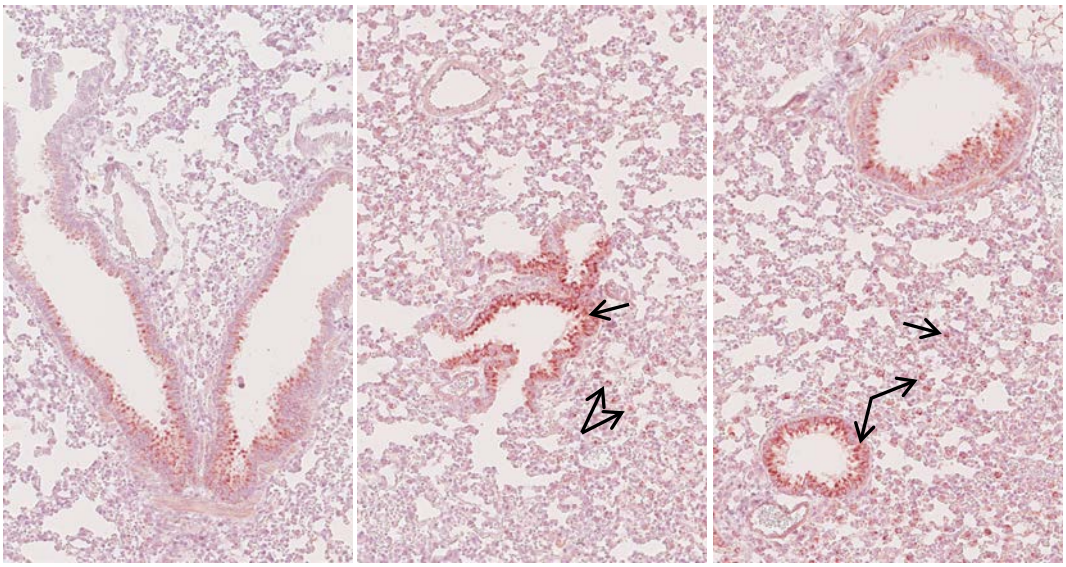


**Figure 32. Immunostaining for Gstcd in mouse lung tissue using an alternative primary antibody (Santa Cruz, sc-138332) performed by Michael Kotlikoff research group, Cornell University.** Staining identifies Gstcd localisation at the bronchial epithelium and in selected cells across the alveolar tissue (indicated by arrows).

## 6.4 Gstcd Knock-out Mice

### 6.4.1 Gstcd Knock-out Genotyping Confirmation

Gstcd -/- mice were genotyped by collaborators at Cornell University (Jane Lee), using reverse transcriptase PCR methods to amplify a region downstream of the knock-out cassette insertion (Figure 33).

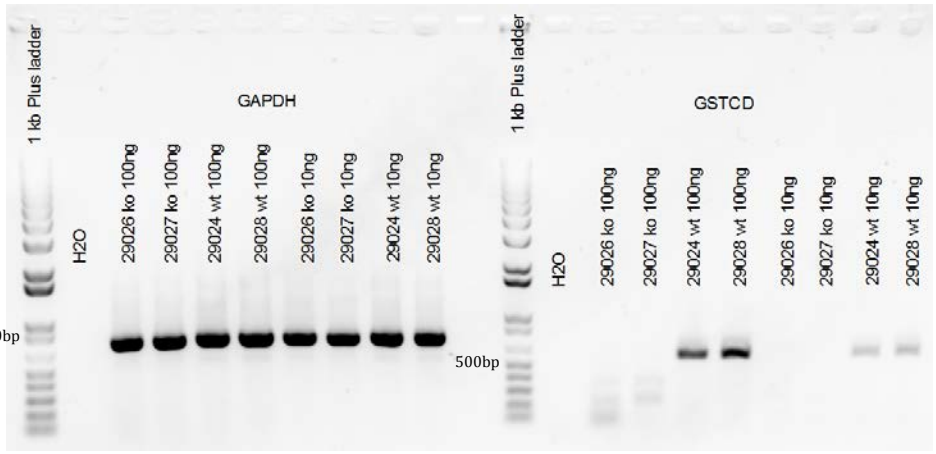



**Figure 33. Genotyping results for Gstcd +/+ and Gstcd -/- mice.** Two amounts of DNA were run from the same samples (100ng and 10ng) for each mouse genotyped. Reverse transcriptase PCR was used to amplify a region of Gstcd (557bp), this is expected to be present in the Gstcd +/+ mice and absent in Gstcd -/- mice where this region has been disrupted by insertion of the knock-out cassette. Samples were then run on an agarose electrophoresis gel. Bands are only evident in Gstcd +/+ mice. GAPDH was used as a positive control and demonstrated even bands across samples (761bp). Two amounts of DNA were run from the same samples (100ng and 10ng) for each mouse genotyped. Key: ko – Gstcd -/- mouse, wt – Gstcd +/+ mouse, H2O – water negative control.

### 6.4.2 Necropsy Analysis of Gstcd Knock-out Mice

During expansion of the Gstcd knock-out mouse colony, -/- mice were observed to reach adulthood and breed normally, therefore a necropsy analysis investigating major organs, bones and glands in two male and female mice was performed (Teresa Southard, Cornell University). This also found no overt significant differences between Gstcd +/+ and Gstcd -/- mice.

**Table 25. Report summarising findings of necropsy analysis of major organs, bones and glands from Gstcd +/+ and Gstcd -/- knock-out mice performed by collaborators at Cornell University (Teresa Southard).** Mild necrotizing hepatitis was observed in the livers of three of the mice, this was described as a very common finding in mouse husbandry rooms that are not helicobacter negative. Key: NSF - No significant findings, NT - No tissue, N/A - Not applicable.

| | Gstcd +/+ | | Gstcd -/- | |
|---|---|---|---|---|
| **Organ/Tissue** | **Male** | **Female** | **Female** | **Male** |
| **Heart** | NSF | NSF | NSF | NSF |
| **Tongue** | NSF | NSF | Focal area of edema and mast cells in the skeletal muscle | NSF |
| **Thymus** | Scattered apoptotic cells in cortex | Scattered apoptotic cells in cortex | Scattered apoptotic cells in the cortex | Few apoptotic cells in cortex |
| **Lung** | NSF | Small numbers of perivascular lymphocytes | NSF | NSF |
| **Trachea** | NSF | NSF | NT | NSF |
| **Oesophagus** | NSF | NSF | NSF | Superficial layer of bacteria on epithelial surface |
| **Thyroid** | NSF | NSF | NT | NSF |
| **Salivary Glands** | NSF | NSF | NSF | NSF |
| **Lymph nodes** | NSF | NSF | NSF | NSF |
| **Kidney** | NSF | NSF | Focal small interstitial infiltrate of lymphocytes and plasma cells | NSF |
| **Adrenal** | NT | Mild subcapsular cell hyperplasia | NSF | NT |
| **Pancreas** | NSF | NSF | NSF | NSF |
| **Lymph nodes** | Mild sinus histiocytosis | NSF | NSF | NSF |

| | | | | |
|---|---|---|---|---|
| **Stomach** | NSF | NSF | Focal area of squamous metaplasia in the glandular portion | NSF |
| **Small Intestine** | NSF | NSF | NSF | NSF |
| **Large intestine** | NSF | NSF | NSF | NSF |
| **Cecum** | NSF | NSF | Small numbers of neutrophils in the lamina propria | NSF |
| **Liver** | Mild multifocal necrotizing hepatitis | Few foci of extramedullary hematopoiesis, rare foci of necrotizing hepatitis | Few foci of necrotizing hepatitis | NSF |
| **Spleen** | Red:White = 1:1 | Red:White = 1:1 | Red:White = 1:1 | Red:White = 1:1; NSF |
| **Ovary** | N/A | Many follicles, one corpus luteum | Many follicles in varying stages of development, one CL | N/A |
| **Uterus** | N/A | NSF | NSF | N/A |
| **Bladder** | NSF, protein coagulum in urethra | NT | NSF | NSF, protein coagulum in urethra |
| **Testis** | Few multinucleate spermatidic giant cells | N/A | N/A | Few multinucleate spermatidic giant cells |
| **Accessory Sex Glands** | NSF | N/A | N/A | NSF |
| **Haired skin** | NSF | Focal area of furunculosis | NSF | NSF |
| **Clitoral/preputial gland** | NSF | NT | NT | NT |
| **Mammary gland** | NT | NSF | NSF | NT |
| **Bone** | NSF | NSF | NSF | NSF |
| **Bone Marrow** | Myeloid:Erythroid = 1:1 | Myeloid:Erythroid = 1:1 | Myeloid:Erythroid = 2:1 | Myeloid:Erythroid = 1:1 |

| | | | | |
|---|---|---|---|---|
| **Brain** | NSF | NSF | NSF | NSF |
| **Eye** | NSF | NSF | NSF | NSF |
| **Nasal Cavities** | NSF | NSF | NSF | NSF |
| **Teeth** | NSF | NSF | NSF | NSF |
| **Pituitary Gland** | NSF | NSF | NSF | NSF |
| **Blood vessels** | NSF | NSF | Focally there are a small number of lymphocytes around a facial blood vessel | |
| **Ear** | NSF | Small amount of proteinaceous fluid and few macrophages in one middle ear with some neutrophils in the surrounding connective tissue | NSF | Small amount of proteinaceous fluid in one middle ear |

## 6.5 Western Blotting Results

### 6.5.1 Quantification of GSTCD Knock-down hBEC samples in Western Blotting

**Table 26. Quantification of Western blotting band intensities reveals a significant reduction of GSTCD protein in siRNA B and siRNA C treated hBECs relative to controls.** Light intensity values for GSTCD bands are normalised against equivalent β-actin values (Relative Density) and proportional changes relative to untransfected control calculated. Key: **Untrans** - untransfected control, **Scrambled** - transfection with scrambled siRNA control, **siRNA B/C** - transfection with GSTCD knock-down siRNA.

| Replicate | Samples | GSTCD | β-actin | Relative Density | % Relative to Untrans |
|---|---|---|---|---|---|
| **siRNA B** | | | | | |
| n1 | Untrans | 76.794 | 145.108 | 0.529 | 100 |
| | Scrambled | 62.047 | 139.454 | 0.445 | 84 |
| | siRNA B | 19.386 | 130.097 | 0.149 | 28 |
| n2 | Untrans | 96.52 | 139.458 | 0.692 | 100 |
| | Scrambled | 88.701 | 133.066 | 0.667 | 96 |
| | siRNA B | 45.09 | 121.264 | 0.372 | 54 |
| n3 | Untrans | 70.825 | 123.886 | 0.572 | 100 |
| | Scrambled | 77.47 | 142.653 | 0.543 | 95 |
| | siRNA B | 19.186 | 113.254 | 0.169 | 30 |
| **siRNA C** | | | | | |
| n1 | Untrans | 79.888 | 129.104 | 0.619 | 100 |
| | Scrambled | 83.611 | 132.445 | 0.631 | 102 |
| | siRNA C | 27.956 | 139.807 | 0.200 | 32 |
| n2 | Untrans | 102.351 | 139.264 | 0.735 | 100 |
| | Scrambled | 94.672 | 114.914 | 0.824 | 112 |
| | siRNA C | 25.435 | 137.782 | 0.185 | 25 |
| n3 | Untrans | 93.601 | 100.489 | 0.931 | 100 |
| | Scrambled | 66.392 | 115.327 | 0.576 | 62 |
| | siRNA C | 18.378 | 93.3 | 0.197 | 21 |

## 6.6 Inferred GSTCD Homologous Proteins

### 6.6.1 GSTCD Top 10 Homologous Proteins Inferred from Protein Prediction Servers

To infer potential function of GSTCD, three protein prediction servers: I-TASSER, SWISS-MODEL and Phyre2, were used to identify proteins sharing the highest levels of homology in protein sequence with both transcript variants of GSTCD. The top 10 homologous results for each server are listed in Table 27.

**Table 27. Top 10 homology templates against the protein sequences of human GSTCD variants 1 and 2 as assessed by web-based protein structure prediction servers.**

| Ranked Homology Template | Protein Structure Prediction Servers | | | | | |
|---|---|---|---|---|---|---|
| | I-TASSER | | SWISS-MODEL | | Phyre2 | |
| | Variant 1 | Variant 2 | Variant 1 | Variant 2 | Variant 1 | Variant 2 |
| 1 | Putative methyltransferase | RAS guanyl-releasing protein 1 | Protein RdmB methyltransferase | Putative rRNA methylase | rRNA small subunit methyltransferase | Transferase |
| 2 | Importin beta subunit | Methyltransferase | Protein RdmB methyltransferase | Putative rRNA methylase | Transferase | NADB-Rossmann fold superfamily protein |
| 3 | Exodeoxyribonuclease | Methyltransferase | Protein RdmB methyltransferase | rRNA small subunit methyltransferase | S-adenosyl-L-methionine-dependent methyltransferase | S-adenosyl-L-methionine-dependent methyltransferase |
| 4 | Ubiquitin ligase | SAM dependent methyltransferase | Putative rRNA methylase | rRNA small subunit methyltransferase | Transferase | S-adenosyl-L-methionine-dependent methyltransferase |
| 5 | rRNA small subunit methyltransferase | rRNA 5-methyluridine methyl-transferase | Putative rRNA methylase | rRNA small subunit methyltransferase | S-adenosyl-L-methionine-dependent methyltransferase | Transferase |
| 6 | Importin beta subunit | Protein phosphatase | rRNA methylase | Protein methyltransferase hemK | Methyltransferase | Uncharacterized protein, (duf633) |
| 7 | N5-glutamine methyltransferase | Methyltransferase | rRNA methylase | Protein methyltransferase hemK | Methyltransferase | S-adenosyl-L-methionine-dependent methyltransferase |

| | | | | | | |
|---|---|---|---|---|---|---|
| **8** | Fanconi-associated nuclease 1 | rRNA 5-methyluridine methyltransferase | Possible methyltransferase | rRNA large subunit methyltransferase | Methyltransferase | Transferase |
| **9** | rRNA small subunit methyltransferase | Methyltransferase | Uncharacterised RNA methyltransferase | rRNA large subunit methyltransferase | Transferase | Transferase |
| **10** | Ubiquitin ligase | tRNA (uracil)-methyltransferase | Uncharacterised RNA methyltransferase | rRNA large subunit methyltransferase | Transferase | S-adenosyl-L-methionine-dependent methyltransferase |

# 7 References

Abbey, D.E. et al., 1998. Long-term Particulate and Other Air Pollutants and Lung Function in Nonsmokers. *American Journal of Respiratory and Critical Care Medicine*, 158(1), pp.289–298. Available at: http://www.atsjournals.org/doi/abs/10.1164/ajrccm.158.1.9710101#.Vh4uBuxVhHw [Accessed October 14, 2015].

Abnova Technologies, 2015. Wheat Germ Protein - Technologies - Support - Abnova. Available at: http://www.abnova.com/support/technologies.asp?switchfunctionid={539614B1-7CBB-4068-A93B-1B44DC6F2801} [Accessed October 8, 2015].

Adnan, H., 2012. Role of Human Glutathione S-Transferase Alpha in Modulating Cellular Stress and Cell Phase Transitions. Available at: https://atrium.lib.uoguelph.ca/xmlui/handle/10214/3965 [Accessed November 17, 2015].

Althoff, K. et al., 2013. MiR-137 functions as a tumor suppressor in neuroblastoma by downregulating KDM1A. *International journal of cancer. Journal international du cancer*, 133(5), pp.1064–73. Available at: http://doi.wiley.com/10.1002/ijc.28091 [Accessed October 27, 2015].

Andrade, M.A., Perez-Iratxeta, C. & Ponting, C.P., 2001. Protein repeats: structures, functions, and evolution. *Journal of structural biology*, 134(2-3), pp.117–31. Available at: http://www.sciencedirect.com/science/article/pii/S1047847701943928 [Accessed July 21, 2015].

Artigas, M.S. et al., 2015. Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nature communications*, 6, p.8658. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26635082 [Accessed December 7, 2015].

Aswal, A.P.S. et al., 2008. Expression stability of two housekeeping genes (18S rRNA and G3PDH) during in vitro maturation of follicular oocytes in buffalo (Bubalus bubalis). *Animal reproduction science*, 103(1-2), pp.164–71. Available at: http://www.sciencedirect.com/science/article/pii/S0378432007001455

[Accessed November 13, 2015].

Bancalari, E. & Polin, R.A., 2012. *The Newborn Lung: Neonatology Questions and Controversies: Expert Consult-Online and Print*, Elsevier Health Sciences.

Bas, A. et al., 2004. Utility of the housekeeping genes 18S rRNA, beta-actin and glyceraldehyde-3-phosphate-dehydrogenase for normalization in real-time quantitative reverse transcriptase-polymerase chain reaction analysis of gene expression in human T lymphocytes. *Scandinavian journal of immunology*, 59(6), pp.566–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15182252 [Accessed November 13, 2015].

Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, 72(1-2), pp.248–254. Available at: http://www.sciencedirect.com/science/article/pii/0003269776905273 [Accessed July 9, 2014].

Bradley, A. et al., 2012. The mammalian gene function resource: the International Knockout Mouse Consortium. *Mammalian genome : official journal of the International Mammalian Genome Society*, 23(9-10), pp.580–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3463800&tool=pmcentrez&rendertype=abstract [Accessed November 11, 2015].

Brameier, M., Krings, A. & MacCallum, R.M., 2007. NucPred Predicting nuclear localization of proteins. *Bioinformatics*, 23(9), pp.1159–1160. Available at: http://bioinformatics.oxfordjournals.org/content/23/9/1159.long [Accessed October 3, 2015].

Brehm, J.M. et al., 2015. A Genome-Wide Association Study of Post-bronchodilator Lung Function in Children with Asthma. *American journal of respiratory and critical care medicine*, 192(5), pp.634–7. Available at: http://www.atsjournals.org/doi/full/10.1164/rccm.201501-0047LE#.Vj9gbtLhDct [Accessed November 8, 2015].

Buck, S. et al., 2008. Detection of S-phase cell cycle progression using 5-ethynyl-2′-deoxyuridine incorporation with click chemistry, an alternative to using

5-bromo-2′-deoxyuridine antibodies. *BioTechniques*, 44(7), pp.927–929. Available at: http://europepmc.org/abstract/med/18533904 [Accessed October 22, 2015].

Cappella, P. et al., 2008. A novel method based on click chemistry, which overcomes limitations of cell cycle analysis by classical determination of BrdU incorporation, allowing multiplex antibody staining. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 73(7), pp.626–36. Available at: http://doi.wiley.com/10.1002/cyto.a.20582 [Accessed September 10, 2015].

Caruana, G. & Bernstein, A., 2001. Craniofacial dysmorphogenesis including cleft palate in mice with an insertional mutation in the discs large gene. *Molecular and cellular biology*, 21(5), pp.1475–83. Available at: http://mcb.asm.org/content/21/5/1475.full [Accessed November 11, 2015].

Castaldi, P.J. et al., 2011. The association of genome-wide significant spirometric loci with chronic obstructive pulmonary disease susceptibility. *American journal of respiratory cell and molecular biology*, 45(6), p.1147.

Celli, B.R. et al., 2004. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. *European Respiratory Journal*, 23(6), pp.932–946. Available at: http://erj.ersjournals.com/content/23/6/932.short [Accessed August 4, 2015].

Chang, D.R. et al., 2013. Lung epithelial branching program antagonizes alveolar differentiation. *Proceedings of the National Academy of Sciences*, 110(45), pp.18042–18051.

Chehrehasa, F. et al., 2009. EdU, a new thymidine analogue for labelling proliferating cells in the nervous system. *Journal of neuroscience methods*, 177(1), pp.122–30. Available at: http://www.sciencedirect.com/science/article/pii/S0165027008005785 [Accessed July 28, 2015].

Chess, P.R. et al., 2010. Murine mechanical ventilation stimulates alveolar

epithelial cell proliferation. *Experimental lung research*, 36(6), pp.331–41. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4342053&tool=pmcentrez&rendertype=abstract [Accessed November 18, 2015].

Choi, J.-S. & Berdis, A.J., 2015. Visualizing nucleic acid metabolism using non-natural nucleosides and nucleotide analogs. *Biochimica et biophysica acta*. Available at: http://www.sciencedirect.com/science/article/pii/S1570963915001508 [Accessed September 2, 2015].

Christofferson, D.E. et al., 2012. A novel role for RIP1 kinase in mediating TNFα production. *Cell Death and Disease*, 3(6), p.e320. Available at: http://dx.doi.org/10.1038/cddis.2012.64 [Accessed September 30, 2015].

Collins, F.S., Rossant, J. & Wurst, W., 2007. A mouse for all reasons. *Cell*, 128(1), pp.9–13.

Cooper, K. & Mitchell, P., 2003. Procedure for the assessment of lung function with spirometry. *Nursing times*, 99(23), pp.57–8. Available at: http://europepmc.org/abstract/med/12838651 [Accessed October 14, 2015].

Cozzetto, D. et al., 2009. Evaluation of template-based models in CASP8 with standard measures. *Proteins*, 77 Suppl 9, pp.18–28. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4589151&tool=pmcentrez&rendertype=abstract [Accessed October 13, 2015].

Crystal, R.G. et al., 2012. Airway Epithelial Cells. *Proceedings of the American Thoracic Society*. Available at: http://www.atsjournals.org/doi/full/10.1513/pats.200805-041HR#.Vkmn-dLhCUk [Accessed November 16, 2015].

Cuthbert, P.C. et al., 2007. Synapse-associated protein 102/dlgh3 couples the NMDA receptor to specific plasticity pathways and learning strategies. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 27(10), pp.2673–82. Available at: http://www.jneurosci.org/content/27/10/2673.short [Accessed November

11, 2015].

Dahl, J.P. et al., 2003. Characterization of the WAVE1 Knock-Out Mouse: Implications for CNS Development. *J. Neurosci.*, 23(8), pp.3343–3352. Available at: http://www.jneurosci.org/content/23/8/3343.short [Accessed November 11, 2015].

Dai, L. & Zhou, Y., 2011. Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. *Journal of molecular biology*, 408(3), pp.585–95. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3075335&tool=pmcentrez&rendertype=abstract [Accessed October 15, 2015].

Danahay, H. et al., 2002. Interleukin-13 induces a hypersecretory ion transport phenotype in human bronchial epithelial cells. *American journal of physiology. Lung cellular and molecular physiology*, 282(2), pp.L226–36. Available at: http://ajplung.physiology.org/content/282/2/L226.abstract [Accessed October 28, 2015].

Dasgupta, P. et al., 2009. Nicotine induces cell proliferation, invasion and epithelial-mesenchymal transition in a variety of human cancer cell lines. *International journal of cancer. Journal international du cancer*, 124(1), pp.36–45. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2826200&tool=pmcentrez&rendertype=abstract [Accessed November 12, 2015].

Davis, R.M. & Novotny, T.E., 1989. The epidemiology of cigarette smoking and its impact on chronic obstructive pulmonary disease. *The American review of respiratory disease*, 140(3 Pt 2), pp.S82–4. Available at: http://www.atsjournals.org/doi/abs/10.1164/ajrccm/140.3_Pt_2.S82?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%3dpubmed#.VjuNIdLhCUk [Accessed November 5, 2015].

Dean, C., Bingle, C. & Hind, M., 2013. Delivering and phenotyping mouse models for the respiratory community: a report on the Biochemical Society Workshop. *Clinical science (London, England : 1979)*, 125(10), pp.495–500.

Available at: http://www.clinsci.org/content/125/10/495.abstract [Accessed November 11, 2015].

Demedts, I.K. et al., 2006. Role of apoptosis in the pathogenesis of COPD and pulmonary emphysema. *Respiratory research*, 7, p.53. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1501017&tool=pmcentrez&rendertype=abstract [Accessed November 17, 2015].

DeMeo, D.L. & Silverman, E.K., 2003. Genetics of chronic obstructive pulmonary disease. *Seminars in respiratory and critical care medicine*, 24(2), pp.151–60. Available at: http://europepmc.org/abstract/med/16088534 [Accessed July 13, 2015].

DeMeo, D.L. & Silverman, E.K., 2004. α1-Antitrypsin deficiency 2: Genetic aspects of α1-antitrypsin deficiency: phenotypes and genetic modifiers of emphysema risk. *Thorax*, 59(3), pp.259–264. Available at: http://thorax.bmj.com/content/59/3/259.full [Accessed November 6, 2015].

Dunker, A.K. et al., 2002. Intrinsic disorder and protein function. *Biochemistry*, 41(21), pp.6573–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12022860 [Accessed August 20, 2015].

Ensembl Human GSTCD, 2015. Ensembl genome browser 81: Homo sapiens - Summary - Gene: GSTCD (ENSG00000138780). Available at: http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000138780;r=4:105708778-105847728 [Accessed September 10, 2015].

Ensembl Mouse Gstcd, 2015. Ensembl genome browser 81: Mus musculus - Summary - Gene: Gstcd (ENSMUSG00000028018). Available at: http://www.ensembl.org/Mus_musculus/Gene/Summary?db=core;g=ENSMUSG00000028018;r=3:132981752-133092033 [Accessed September 10, 2015].

Ensembl: GSTCD Domains and Features., 2014. Ensembl genome browser 81: Homo sapiens - Domains & features - Transcript: GSTCD-002 (ENST00000360505). Available at:

http://www.ensembl.org/Homo_sapiens/Transcript/Domains?db=core;g=ENSG00000138780;r=4:105708778-105847728;t=ENST00000360505 [Accessed September 17, 2015].

Fehrenbach, H., 2001. Alveolar epithelial type II cell: defender of the alveolus revisited. *Respiratory research*, 2(1), p.33.

Feng, L.-X. et al., 2011. Clarifying the signal network of salvianolic acid B using proteomic assay and bioinformatic analysis. *Proteomics*, 11(8), pp.1473–85. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21365754 [Accessed October 22, 2015].

Ferranti, S., 2014. Pulmonary Critical Care. *Health Tap*. Available at: https://www.healthtap.com/user_questions/308811 [Accessed November 6, 2015].

Fire, A. et al., 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669), pp.806–11. Available at: http://dx.doi.org/10.1038/35888 [Accessed July 17, 2014].

Fiser, A., 2010. Template-based protein structure modeling. *Methods in molecular biology (Clifton, N.J.)*, 673, pp.73–94. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4108304&tool=pmcentrez&rendertype=abstract [Accessed September 9, 2015].

GOLD, 2015. *Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease*, Available at: http://www.goldcopd.org/.

Grant, S.G. et al., 1992. Impaired long-term potentiation, spatial learning, and hippocampal development in fyn mutant mice. *Science (New York, N.Y.)*, 258(5090), pp.1903–10. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1361685 [Accessed November 11, 2015].

Grimm, D., 2006. Mouse genetics. A mouse for every gene. *Science (New York, N.Y.)*, 312(5782), pp.1862–6. Available at: http://www.sciencemag.org/content/312/5782/1862.full [Accessed July

29, 2015].

Haines, K.M., Feldman, E.L. & Lentz, S.I., 2010. Visualization of mitochondrial DNA replication in individual cells by EdU signal amplification. *Journal of visualized experiments : JoVE*, (45). Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159597&tool=pmcentrez&rendertype=abstract [Accessed September 1, 2015].

Hall, I.P., 2009. How will genetic approaches assist in the management of respiratory diseases? *Current opinion in pharmacology*, 9(3), pp.256–261.

Hallberg, J. et al., 2010. Genetic and environmental influence on lung function impairment in Swedish twins. *Respiratory research*, 11(1), p.92. Available at: http://respiratory-research.com/content/11/1/92 [Accessed September 18, 2015].

Hancock, D.B. et al., 2012. Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS genetics*, 8(12), p.e1003098. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3527213&tool=pmcentrez&rendertype=abstract [Accessed August 12, 2014].

Hancock, D.B. et al., 2010. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nature genetics*, 42(1), pp.45–52.

Hardyman, M.A. et al., 2013. TNF-α-mediated bronchial barrier disruption and regulation by src-family kinase activation. *The Journal of allergy and clinical immunology*, 132(3), pp.665–675.e8. Available at: http://www.sciencedirect.com/science/article/pii/S0091674913004259 [Accessed October 28, 2015].

Harrison, M.J., Tang, Y.H. & Dowhan, D.H., 2010. Protein arginine methyltransferase 6 regulates multiple aspects of gene expression. *Nucleic Acids Research*, 38(7), pp.2201–2216. Available at: http://nar.oxfordjournals.org/content/38/7/2201.abstract?ijkey=3c5057ea971ec6789222dfc4858e2adadcb63626&keytype2=tf_ipsecsha [Accessed November 17, 2015].

Hayes, J.D., Flanagan, J.U. & Jowsey, I.R., 2005. Glutathione transferases. *Annu. Rev. Pharmacol. Toxicol.*, 45, pp.51–88.

Herrera, I. et al., 2013. Matrix metalloproteinase (MMP)-1 induces lung alveolar epithelial cell migration and proliferation, protects from apoptosis, and represses mitochondrial oxygen consumption. *The Journal of biological chemistry*, 288(36), pp.25964–75. Available at: http://www.jbc.org/content/288/36/25964.short [Accessed November 12, 2015].

Herriges, M. & Morrisey, E.E., 2014. Lung development: orchestrating the generation and regeneration of a complex organ. *Development*, 141(3), pp.502–513.

Hirschhorn, J.N. & Daly, M.J., 2005. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2), pp.95–108.

Huang, Q. et al., 2009. A novel function of poly(ADP-ribose) polymerase-1 in modulation of autophagy and necrosis under oxidative stress. *Cell death and differentiation*, 16(2), pp.264–77. Available at: http://dx.doi.org/10.1038/cdd.2008.151 [Accessed November 11, 2015].

Huh, J.Y. et al., 2013. A novel function of adipocytes in lipid antigen presentation to iNKT cells. *Molecular and cellular biology*, 33(2), pp.328–39. Available at: http://mcb.asm.org/content/33/2/328.full [Accessed November 11, 2015].

Ignatius, M., 2009. Cell Proliferation and Cell Counting. *Invitrogen*. Available at: http://webcache.googleusercontent.com/search?q=cache:avYJOHp3fsQJ:www.researchgate.net/publictopics.PublicPostFileLoader.html%3Fid%3D544a6155d5a3f2ab148b45a1%26key%3D9c7cd8c7-241a-439b-b677-4a7972945bcf+&cd=1&hl=en&ct=clnk&gl=uk [Accessed September 3, 2015].

Ijiri, T.W. et al., 2011. Identification and validation of mouse sperm proteins correlated with epididymal maturation. *Proteomics*, 11(20), pp.4047–62. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3517136&tool=pmcentrez&rendertype=abstract [Accessed November 16, 2015].

Imboden, M. et al., 2012. Genome-wide association study of lung function decline in adults with and without asthma. *Journal of Allergy and Clinical Immunology*, 129(5), pp.1218–1228.

IMPC: GSTCD, 2015. Gene details for glutathione S-transferase, C-terminal domain containing | International Mouse Phenotyping Consortium. Available at: https://www.mousephenotype.org/data/genes/MGI:1914803 [Accessed August 27, 2015].

IMPC: GSTCD(EUCOMM), 2015. Allele details | International Mouse Phenotyping Consortium. Available at: https://www.mousephenotype.org/data/alleles/MGI:1914803/tm1a(EUCOMM)Wtsi/ [Accessed August 27, 2015].

IMPC: GSTCD(NCOM), 2015. Allele details - NorCOMM | International Mouse Phenotyping Consortium. Available at: http://www.mousephenotype.org/data/alleles/MGI:1914803/L1L2_GOHANU/237043/ [Accessed August 26, 2015].

IMPReSS, 2015. The Adult and Embryonic Phenotype Pipeline. , (29/10/2015). Available at: https://www.mousephenotype.org/impress/.

Jackson-Informatics, 2014. Vertebrate Homology. , (12/02/2014). Available at: www.informatics.jax.org/homology.shtml.

Jeffery, P.K., 2001. Remodeling in asthma and chronic obstructive lung disease. *American journal of respiratory and critical care medicine*, 164(10 Pt 2), pp.S28–38. Available at: http://www.atsjournals.org/doi/abs/10.1164/ajrccm.164.supplement_2.2106061#.Vkt46dLhCUk [Accessed October 5, 2015].

Jin, Y. et al., 2011. A novel function for programmed death ligand-1 regulation of angiogenesis. *The American journal of pathology*, 178(4), pp.1922–9. Available at: http://www.sciencedirect.com/science/article/pii/S0002944011000150 [Accessed November 11, 2015].

Jirkovsky, E. et al., 2012. Chronic anthracycline cardiotoxicity: molecular and

functional analysis with focus on nuclear factor erythroid 2-related factor 2 and mitochondrial biogenesis pathways. *The Journal of pharmacology and experimental therapeutics*, 343(2), pp.468–78. Available at: http://jpet.aspetjournals.org/content/343/2/468.full [Accessed November 10, 2015].

Johnson, P.R. et al., 2001. Airway smooth muscle cell proliferation is increased in asthma. *American journal of respiratory and critical care medicine*, 164(3), pp.474–7. Available at: http://www.atsjournals.org/doi/abs/10.1164/ajrccm.164.3.2010109#.VkSUmdLhCUk [Accessed October 26, 2015].

Jones, L.J. et al., 2001. Sensitive determination of cell number using the CyQUANT cell proliferation assay. *Journal of immunological methods*, 254(1-2), pp.85–98. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11406155 [Accessed January 5, 2016].

Kelley, L.A. et al., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6), pp.845–858. Available at: http://dx.doi.org/10.1038/nprot.2015.053 [Accessed May 8, 2015].

Khalil, A.M. et al., 2008. A novel RNA transcript with antiapoptotic function is silenced in fragile X syndrome. *PloS one*, 3(1), p.e1486. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0001486 [Accessed November 11, 2015].

Kihara, D. & Skolnick, J., 2003. The PDB is a covering set of small protein structures. *Journal of molecular biology*, 334(4), pp.793–802. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14636603 [Accessed September 18, 2015].

Kohansal, R. et al., 2009. The natural history of chronic airflow obstruction revisited: an analysis of the Framingham offspring cohort. *American journal of respiratory and critical care medicine*, 180(1), pp.3–10. Available at: http://www.atsjournals.org/doi/abs/10.1164/rccm.200901-0047OC#.Vjy3HtLhDcs [Accessed November 6, 2015].

Koopman, P., 2010. *Organogenesis in Development*, Elsevier.

Kopp, J. et al., 2007. Assessment of CASP7 predictions for template-based modeling targets. *Proteins*, 69 Suppl 8, pp.38–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17894352 [Accessed August 10, 2015].

Kreiner-Møller, E., Bisgaard, H. & Bønnelykke, K., 2014. Prenatal and postnatal genetic influence on lung function development. *The Journal of allergy and clinical immunology*. Available at: http://www.sciencedirect.com/science/article/pii/S0091674914005211 [Accessed August 25, 2014].

Kryshtafovych, A. et al., 2014. Assessment of the assessment: Evaluation of the model quality estimates in CASP10. *Proteins: Structure, Function, and Bioinformatics*, 82, pp.112–126. Available at: http://doi.wiley.com/10.1002/prot.24347 [Accessed October 5, 2015].

Kuchipudi, S. V et al., 2012. 18S rRNA is a reliable normalisation gene for real time PCR based on influenza virus infected cells. *Virology journal*, 9, p.230. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3499178&tool=pmcentrez&rendertype=abstract [Accessed November 13, 2015].

Lapperre, T.S. et al., 2007. Smoking cessation and bronchial epithelial remodelling in COPD: a cross-sectional study. *Respiratory research*, 8, p.85. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2214729&tool=pmcentrez&rendertype=abstract [Accessed November 17, 2015].

Lee, J. et al., 2006. Contribution of proliferation and DNA damage repair to alveolar epithelial type 2 cell recovery from hyperoxia. *American journal of physiology. Lung cellular and molecular physiology*, 290(4), pp.L685–L694. Available at: http://ajplung.physiology.org/content/290/4/L685 [Accessed November 18, 2015].

Leigh, M.W. et al., 1995. Cell proliferation in bronchial epithelium and submucosal glands of cystic fibrosis patients. *American journal of respiratory cell and molecular biology*, 12(6), pp.605–12. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/7766425 [Accessed November 12, 2015].

Lentz, S.I. et al., 2009. Mitochondrial DNA (mtDNA) Biogenesis: Visualization and Duel Incorporation of BrdU and EdU Into Newly Synthesized mtDNA In Vitro. *Journal of Histochemistry and Cytochemistry*, 58(2), pp.207–218. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2803709&tool=pmcentrez&rendertype=abstract [Accessed October 5, 2015].

Levenkova, N., Gu, Q. & Rux, J.J., 2004. Gene specific siRNA selector. *Bioinformatics (Oxford, England)*, 20(3), pp.430–2. Available at: http://bioinformatics.oxfordjournals.org/content/20/3/430.short [Accessed September 1, 2015].

Levinsson, A. et al., 2014. Interaction Effects of Long-Term Air Pollution Exposure and Variants in the GSTP1, GSTT1 and GSTCD Genes on Risk of Acute Myocardial Infarction and Hypertension: A Case-Control Study. *PloS one*, 9(6), p.e99043.

Lewitter, F.I. et al., 1984. Genetic and Environmental Determinants of Level of Pulmonary Function. *Am. J. Epidemiol.*, 120(4), pp.518–530. Available at: http://aje.oxfordjournals.org/content/120/4/518.short [Accessed July 13, 2015].

Li, X. et al., 2013. Genome-wide association study identifies TH1 pathway genes associated with lung function in asthmatic patients. *The Journal of allergy and clinical immunology*, 132(2), pp.313–20.e15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3746327&tool=pmcentrez&rendertype=abstract [Accessed August 12, 2014].

Lin, J.-R. et al., 2012. Minimalist ensemble algorithms for genome-wide protein localization prediction. *BMC bioinformatics*, 13(1), p.157. Available at: http://www.biomedcentral.com/1471-2105/13/157 [Accessed October 21, 2015].

Løkke, A. et al., 2006. Developing COPD: a 25 year follow up study of the general population. *Thorax*, 61(11), pp.935–9. Available at:

http://thorax.bmj.com/content/61/11/935.abstract [Accessed October 14, 2015].

Loth, D.W. et al., 2014. Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nature genetics*, 46(7), pp.669–77. Available at: http://dx.doi.org/10.1038/ng.3011 [Accessed August 30, 2014].

Lozano, R. et al., 2012. 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. *Lancet*, 380(9859), pp.2095–128. Available at: http://www.sciencedirect.com/science/article/pii/S0140673612617280 [Accessed July 10, 2014].

Luisetti, M. & Seersholm, N., 2004. α1-Antitrypsin deficiency 1: Epidemiology of α1-antitrypsin deficiency. *Thorax*, 59(2), pp.164–169. Available at: http://thorax.bmj.com/content/59/2/164.full [Accessed November 5, 2015].

Luka, Z., Loukachevitch, L. V & Wagner, C., 2008. Acetylation of N-terminal valine of glycine N-methyltransferase affects enzyme inhibition by folate. *Biochimica et biophysica acta*, 1784(9), pp.1342–6. Available at: http://www.sciencedirect.com/science/article/pii/S1570963908001337 [Accessed November 16, 2015].

Majumder, S. et al., 2006. Involvement of arginine methyltransferase CARM1 in androgen receptor function and prostate cancer cell viability. *The Prostate*, 66(12), pp.1292–1301. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16705743 [Accessed November 17, 2015].

Mariani, V. et al., 2011. Assessment of template based protein structure predictions in CASP9. *Proteins*, 79 Suppl 1, pp.37–58. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22002823 [Accessed October 15, 2015].

Mathers, C., Fat, D. & Boerma, J., 2008. *The Global Burden of Disease: 2004 update*, Available at:

http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/index.html.

McElroy, M.C. & Kasper, M., 2004. The use of alveolar epithelial type I cell-selective markers to investigate lung injury and repair. *The European respiratory journal*, 24(4), pp.664–73. Available at: http://erj.ersjournals.com/content/24/4/664 [Accessed November 11, 2015].

Metzger, R.J. et al., 2008. The branching programme of mouse lung development. *Nature*, 453(7196), pp.745–750.

Moran, L. et al., 2012. *Principles of Biochemistry*, Boston: Pearson Education. Available at: https://library.wur.nl/WebQuery/clc/1974613.

Morrisey, E.E. & Hogan, B.L.M., 2010. Preparing for the first breath: genetic and cellular mechanisms in lung development. *Developmental cell*, 18(1), pp.8–23.

Morrison, G. et al., 2014. Therapeutic potential of the dual EGFR/HER2 inhibitor AZD8931 in circumventing endocrine resistance. *Breast cancer research and treatment*, 144(2), pp.263–72. Available at: http://link.springer.com/article/10.1007/s10549-014-2878-x/fulltext.html [Accessed October 22, 2015].

Nagy, A., Mar, L. & Watts, G., 2009. Creation and use of a cre recombinase transgenic database. *Methods in molecular biology (Clifton, N.J.)*, 530, pp.365–78. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19266338 [Accessed November 13, 2015].

NCBI Conserved Protein Domain Resource, 2015. NCBI Conserved Protein Domain GST_C_family. Available at: http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?ascbin=8&maxaln=10&seltype=2&uid=cl02776 [Accessed November 18, 2015].

NCBI HomoloGene, 2014. NCBI HomoloGene:11693 Pairwise Alignment Scores. Available at: http://www.ncbi.nlm.nih.gov/homologene?cmd=Retrieve&dopt=Alignment

Scores&list_uids=11693 [Accessed November 11, 2015].

Nebert, D.W. & Vasiliou, V., 2004. Analysis of the glutathione S-transferase (GST) gene family. *Hum Genomics*, 1(6), pp.460–464.

Needham, M., 2004. 1-Antitrypsin deficiency * 3: Clinical manifestations and natural history. *Thorax*, 59(5), pp.441–445. Available at: http://thorax.bmj.com/content/59/5/441.full [Accessed November 6, 2015].

NHS, 2014. Chronic Obstructive Pulmonary Disease. , (01/10/2014). Available at: http://www.nhs.uk/Conditions/chronic-obstructive-pulmonary-disease/Pages/Introduction.aspx [Accessed November 5, 2015].

Nogawa, H., Morita, K. & Cardoso, W. V, 1998. Bud formation precedes the appearance of differential cell proliferation during branching morphogenesis of mouse lung epithelium in vitro. *Developmental dynamics : an official publication of the American Association of Anatomists*, 213(2), pp.228–35. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9786423 [Accessed November 12, 2015].

Obeidat, M. et al., 2013. GSTCD and INTS12 Regulation and Expression in the Human Lung. *PloS one*, 8(9), p.e74630.

Obeidat, M. et al., 2015. Molecular mechanisms underlying variations in lung function: a systems genetics analysis. *The Lancet Respiratory Medicine*, 3(10), pp.782–795. Available at: http://www.thelancet.com/article/S221326001500380X/fulltext [Accessed September 22, 2015].

Obeidat, M. & Hall, I.P., 2011. Genetics of complex respiratory diseases: implications for pathophysiology and pharmacology studies. *British journal of pharmacology*, 163(1), pp.96–105.

Okamura, T. et al., 2009. Tyrosine phosphorylation of the human glutathione S-transferase P1 by epidermal growth factor receptor. *The Journal of biological chemistry*, 284(25), pp.16979–89. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2719335&tool

=pmcentrez&rendertype=abstract [Accessed November 6, 2015].

Ong, B.A. et al., 2013. Gene network analysis in a pediatric cohort identifies novel lung function genes. *PloS one*, 8(9), p.e72899. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3759429&tool=pmcentrez&rendertype=abstract [Accessed August 12, 2014].

Pagano, A. et al., 2007. Poly(ADP-ribose) polymerase-1 (PARP-1) controls lung cell proliferation and repair after hyperoxia-induced lung damage. *American journal of physiology. Lung cellular and molecular physiology*, 293(3), pp.L619–29. Available at: http://ajplung.physiology.org/content/293/3/L619 [Accessed November 18, 2015].

Pajtler, K.W. et al., 2013. The KDM1A histone demethylase is a promising new target for the epigenetic therapy of medulloblastoma. *Acta neuropathologica communications*, 1, p.19. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3893444&tool=pmcentrez&rendertype=abstract [Accessed October 27, 2015].

Palmer, L.J. et al., 2001. Familial aggregation and heritability of adult lung function: results from the Busselton Health Study. *Eur. Respir. J.*, 17(4), pp.696–702. Available at: http://erj.ersjournals.com/content/17/4/696.short [Accessed July 13, 2015].

Pauwels, R.A. et al., 2001. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: National Heart, Lung, and Blood Institute and World Health Organization Global Initiative for Chronic Obstructive Lung Disease (GOLD): executive summar. *Respiratory care*, 46(8), pp.798–825. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11463370 [Accessed May 20, 2015].

Penzkofer, D. et al., 2014. Phenotypic characterization of miR-92a-/- mice reveals an important function of miR-92a in skeletal development. *PloS one*, 9(6), p.e101153. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0101153

[Accessed November 11, 2015].

Peters, A.H.F.M. et al., 2001. Loss of the Suv39h Histone Methyltransferases Impairs Mammalian Heterochromatin and Genome Stability. *Cell*, 107(3), pp.323–337. Available at: http://www.sciencedirect.com/science/article/pii/S0092867401005426 [Accessed November 17, 2015].

Petrilli, A. et al., 2014. LIM domain kinases as potential therapeutic targets for neurofibromatosis type 2. *Oncogene*, 33(27), pp.3571–82. Available at: http://dx.doi.org/10.1038/onc.2013.320 [Accessed October 22, 2015].

Petrossian, T.C. & Clarke, S.G., 2011. Uncovering the human methyltransferasome. *Molecular & cellular proteomics : MCP*, 10(1), p.M110.000976. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013446&tool=pmcentrez&rendertype=abstract [Accessed August 12, 2014].

Postma, D.S. & Boezen, H.M., 2004. Rationale for the Dutch hypothesis. Allergy and airway hyperresponsiveness as genetic factors and their interaction with environment in the development of asthma and COPD. *Chest*, 126(2 Suppl), p.96S–104S; discussion 159S–161S. Available at: http://journal.publications.chestnet.org/article.aspx?articleid=1082740 [Accessed September 18, 2015].

Poujol, F. et al., 2014. Flow cytometric evaluation of lymphocyte transformation test based on 5-ethynyl-2'deoxyuridine incorporation as a clinical alternative to tritiated thymidine uptake measurement. *Journal of immunological methods*, 415, pp.71–9. Available at: http://www.sciencedirect.com/science/article/pii/S0022175914002920 [Accessed September 2, 2015].

Primhak, R.A. & Tanner, M.S., 2001. Alpha-1 antitrypsin deficiency. *Archives of Disease in Childhood*, 85(1), pp.2–5. Available at: http://adc.bmj.com/content/85/1/2.full [Accessed October 14, 2015].

Quashie, P.K. et al., 2015. Structural Studies of the HIV-1 Integrase Protein: Compound Screening and Characterization of a DNA-Binding Inhibitor D. Harrich, ed. *PLOS ONE*, 10(6), p.e0128310. Available at:

/pmc/articles/PMC4457863/?report=abstract [Accessed September 7, 2015].

Ranganathan, P.N., Whalen, R. & Boyer, T.D., 2005. Characterization of the molecular forms of glutathione S-transferase P1 in human gastric cancer cells (Kato III) and in normal human erythrocytes. *The Biochemical journal*, 386(Pt 3), pp.525–33. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1134871&tool=pmcentrez&rendertype=abstract [Accessed November 16, 2015].

Remaut, H. & Waksman, G., 2006. Protein-protein interaction through beta-strand addition. *Trends in biochemical sciences*, 31(8), pp.436–44. Available at: http://www.sciencedirect.com/science/article/pii/S0968000406001691 [Accessed July 30, 2015].

Repapi, E. et al., 2010. Genome-wide association study identifies five loci associated with lung function. *Nature genetics*, 42(1), pp.36–44.

Reznik, S.I. et al., 2000. Anti-HLA antibody binding to HLA class I molecules induces proliferation of airway epithelial cells: A potential mechanism for bronchiolitis obliterans syndrome. *The Journal of Thoracic and Cardiovascular Surgery*, 119(1), pp.39–45. Available at: http://www.sciencedirect.com/science/article/pii/S0022522300702157 [Accessed November 12, 2015].

Rose, F. et al., 2002. Hypoxic pulmonary artery fibroblasts trigger proliferation of vascular smooth muscle cells: role of hypoxia-inducible transcription factors. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 16(12), pp.1660–1. Available at: http://www.fasebj.org/content/16/12/1660.full [Accessed November 12, 2015].

Rose, P.W. et al., 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic acids research*, 43(Database issue), pp.D345–56. Available at: http://nar.oxfordjournals.org/content/43/D1/D345.full [Accessed August

10, 2015].

Rowe, J.D., Nieves, E. & Listowsky, I., 1997. Subunit diversity and tissue distribution of human glutathione S-transferases: interpretations based on electrospray ionization-MS and peptide sequence-specific antisera. *The Biochemical journal*, 325 ( Pt 2, pp.481–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1218585&tool=pmcentrez&rendertype=abstract [Accessed November 16, 2015].

Rudolph, M., 2010. National Clinical Guideline Centre, Chronic obstructive pulmonary disease: management of chronic obstructive pulmonary disease in adults in primary and secondary care. *London: National Clinical Guideline Centre*.

Salic, A. & Mitchison, T.J., 2008. A chemical method for fast and sensitive detection of DNA synthesis in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 105(7), pp.2415–20. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2268151&tool=pmcentrez&rendertype=abstract [Accessed October 22, 2015].

Santa Cruz Biotechnology RNAi, 2015. RNA Interference. Available at: https://www.scbt.com/gene_silencers.html [Accessed October 15, 2015].

Schindelin, J. et al., 2012. Fiji: an open-source platform for biological-image analysis. *Nature methods*, 9(7), pp.676–82. Available at: http://dx.doi.org/10.1038/nmeth.2019 [Accessed July 9, 2014].

Scott, A. et al., 2014. Regulation of protein stability of DNA methyltransferase 1 by post-translational modifications. *Acta biochimica et biophysica Sinica*, 46(3), pp.199–203. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3932833&tool=pmcentrez&rendertype=abstract [Accessed November 16, 2015].

Shahab, L. et al., 2006. Prevalence, diagnosis and relation to tobacco dependence of chronic obstructive pulmonary disease in a nationally representative population sample. *Thorax*, 61(12), pp.1043–7. Available at: http://thorax.bmj.com/content/61/12/1043.abstract [Accessed October 14, 2015].

Sherratt, P.J. & Hayes, J.D., 2001. Glutathione S-transferases. *Enzyme systems that metabolise drugs and other xenobiotics*, 319.

Siganaki, M. et al., 2010. Deregulation of apoptosis mediators' p53 and bcl2 in lung tissue of COPD patients. *Respiratory research*, 11, p.46. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2873932&tool=pmcentrez&rendertype=abstract [Accessed November 17, 2015].

Silverman, E.K. et al., 1998. Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease: risk to relatives for airflow obstruction and chronic bronchitis. *American journal of respiratory and critical care medicine*, 157(6), pp.1770–1778.

Silverman, E.K. et al., 2002. Genomewide linkage analysis of quantitative spirometric phenotypes in severe early-onset chronic obstructive pulmonary disease. *The American Journal of Human Genetics*, 70(5), pp.1229–1239.

Skarnes, W.C. et al., 2011. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351), pp.337–342.

Söding, J., Biegert, A. & Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(Web Server issue), pp.W244–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1160169&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Soler Artigas, M., Wain, L. V, et al., 2011. Effect of five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function. *American journal of respiratory and critical care medicine*, 184(7), p.786. Available at: http://www.atsjournals.org/doi/full/10.1164/rccm.201102-0192OC#.U-nhW5RdWSo.

Soler Artigas, M., Loth, D.W., et al., 2011. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nature genetics*, 43(11), pp.1082–1090.

Stein, C. et al., 2012. The arginine methyltransferase PRMT6 regulates cell proliferation and senescence through transcriptional repression of tumor suppressor genes. *Nucleic acids research*, 40(19), pp.9522–33. Available at: http://nar.oxfordjournals.org/content/40/19/9522.full#ref-12 [Accessed November 17, 2015].

Strange, R.C. et al., 2001. Glutathione-S-transferase family of enzymes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 482(1-2), pp.21–26. Available at: http://www.sciencedirect.com/science/article/pii/S0027510701002068 [Accessed November 9, 2015].

Struck, A.-W. et al., 2012. S -Adenosyl-Methionine-Dependent Methyltransferases: Highly Versatile Enzymes in Biocatalysis, Biosynthesis and Other Biotechnological Applications. *ChemBioChem*, 13(18), pp.2642–2655. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23180741 [Accessed September 10, 2015].

Takada, I. et al., 2007. A histone lysine methyltransferase activated by non-canonical Wnt signalling suppresses PPAR-gamma transactivation. *Nature cell biology*, 9(11), pp.1273–85. Available at: http://dx.doi.org/10.1038/ncb1647 [Accessed November 16, 2015].

Tamari, M., Tomita, K. & Hirota, T., 2011. Genome-wide association studies of asthma. *Allergology International*, 60(3), p.247.

The Human Protein Atlas, 2015. The Human Protein Atlas - GSTCD. , (11/03/2014). Available at: http://www.proteinatlas.org/ENSG00000138780/tissue.

Tsien, J.Z., Huerta, P.T. & Tonegawa, S., 1996. The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory. *Cell*, 87(7), pp.1327–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8980238 [Accessed November 11, 2015].

Tzortzaki, E.G. et al., 2006. Laboratory markers for COPD in “susceptible” smokers. *Clinica chimica acta; international journal of clinical chemistry*,

364(1-2), pp.124–38. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16139829 [Accessed November 5, 2015].

Vestbo, J. et al., 2013. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *American journal of respiratory and critical care medicine*, 187(4), pp.347–365.

Vijayan, V.K., 2013. Chronic obstructive pulmonary disease. *The Indian journal of medical research*, 137(2), pp.251–69. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3657849&tool=pmcentrez&rendertype=abstract [Accessed August 12, 2014].

Wain, L. V et al., 2015. 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*, 3(10), pp.769–781. Available at: http://www.thelancet.com/article/S2213260015002830/fulltext [Accessed September 30, 2015].

Warburton, D. et al., 2006. Lung development and susceptibility to chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society*, 3(8), pp.668–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17065371 [Accessed November 5, 2015].

Warburton, D. et al., 2010. Lung Organogenesis. *Current topics in developmental biology*, 90, pp.73–158.

Weimann, M. et al., 2013. A Y2H-seq approach defines the human protein methyltransferase interactome. *Nature methods*, 10(4), pp.339–42. Available at: http://dx.doi.org/10.1038/nmeth.2397 [Accessed October 7, 2014].

Wilk, J.B. et al., 2009. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS genetics*, 5(3), p.e1000429. Available at: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.100

0429 [Accessed August 31, 2015].

Wright, P.E. & Dyson, H.J., 2014. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, 16(1), pp.18–29. Available at: http://dx.doi.org/10.1038/nrm3920 [Accessed December 22, 2014].

Xia, M. et al., 2010. Novel function and intracellular localization of methionine adenosyltransferase 2beta splicing variants. *The Journal of biological chemistry*, 285(26), pp.20015–21. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2888413&tool=pmcentrez&rendertype=abstract [Accessed October 22, 2015].

Yamaoka, T. & Kitamura, Y., 2015. Characterization of a highly sensitive and selective novel trapping reagent, stable isotope labeled glutathione ethyl ester, for the detection of reactive metabolites. *Journal of pharmacological and toxicological methods*, 76, pp.83–95. Available at: http://www.sciencedirect.com/science/article/pii/S105687191500252X [Accessed November 10, 2015].

Yang, H. et al., 2014. miR-329 suppresses the growth and motility of neuroblastoma by targeting KDM1A. *FEBS letters*, 588(1), pp.192–7. Available at: http://www.sciencedirect.com/science/article/pii/S0014579313008806 [Accessed November 17, 2015].

Yang, J. et al., 2014. The I-TASSER Suite: protein structure and function prediction. *Nature methods*, 12(1), pp.7–8. Available at: http://dx.doi.org/10.1038/nmeth.3213 [Accessed December 31, 2014].

Yang, Y. et al., 2011. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics (Oxford, England)*, 27(15), pp.2076–82. Available at: http://bioinformatics.oxfordjournals.org/content/27/15/2076.abstract [Accessed September 15, 2015].

Yao, T.-C. et al., 2014. Genome-wide association study of lung function phenotypes in a founder population. *The Journal of allergy and clinical immunology*, 133(1), pp.248–55.e1–10. Available at: http://www.sciencedirect.com/science/article/pii/S009167491300986X [Accessed August 12, 2014].

Yokota, S. et al., 2012. Implication of Antigenic Conversion of Helicobacter pylori Lipopolysaccharides That Involve Interaction with Surfactant Protein D B. A. McCormick, ed. *Infection and Immunity*, 80(8), pp.2956–2962. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3434594&tool=pmcentrez&rendertype=abstract [Accessed October 22, 2015].

Yoshimatsu, M. et al., 2011. Dysregulation of PRMT1 and PRMT6, Type I arginine methyltransferases, is involved in various types of human cancers. *International journal of cancer. Journal international du cancer*, 128(3), pp.562–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20473859 [Accessed November 17, 2015].

Yu, Y. et al., 2009. EdU incorporation is an alternative non-radioactive assay to [(3)H]thymidine uptake for in vitro measurement of mice T-cell proliferations. *Journal of immunological methods*, 350(1-2), pp.29–35. Available at: http://www.sciencedirect.com/science/article/pii/S0022175909002294 [Accessed October 28, 2015].

Zeng, C. et al., 2010. Evaluation of 5-ethynyl-2'-deoxyuridine staining as a sensitive and reliable method for studying cell proliferation in the adult nervous system. *Brain research*, 1319, pp.21–32. Available at: http://www.sciencedirect.com/science/article/pii/S000689931000003X [Accessed September 11, 2015].

Zhai, G. et al., 2007. The Interaction of Genes and Smoking on Forced Expiratory Volume *. *CHEST Journal*, 132(6), p.1772. Available at: http://journal.publications.chestnet.org/article.aspx?articleid=1085610 [Accessed September 18, 2015].

Zhang, Y. et al., 2006. On the origin and highly likely completeness of single-

domain protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8), pp.2605–10. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1413790&tool=pmcentrez&rendertype=abstract [Accessed October 15, 2015].

Zhang, Y., 2009. Protein structure prediction: when is it useful? *Current opinion in structural biology*, 19(2), pp.145–55. Available at: http://www.sciencedirect.com/science/article/pii/S0959440X09000256 [Accessed July 6, 2015].