Translating lung function genome wide association study (GWAS) findings: new insights for lung biology

Abdul Kader Kheirallah*, Suzanne Miller*, Ian P. Hall and Ian Sayers**

Division of Respiratory Medicine, Queens Medical Centre, University of Nottingham, Nottingham, United Kingdom.

* These authors contributed equally

** Corresponding author

Dr Ian Sayers

Division of Respiratory Medicine,

Queen's Medical Centre, University of Nottingham,

Nottingham, NG7 2UH, UK,

Tel: 0115 82 31066, Fax: 0115 82 31059

E-mail: ian.sayers@nottingham.ac.uk

Abstract

Chronic respiratory diseases are a major cause of worldwide mortality and morbidity. Although hereditary severe deficiency of alpha 1 antitrypsin (A1AD) has been established to cause emphysema, A1AD accounts for only ~1% of Chronic Obstructive Pulmonary Disease (COPD) cases. Genome-wide association studies (GWAS) have been successful at detecting multiple loci harbouring variants predicting the variation in lung function measures and risk of COPD. However, GWAS are incapable of distinguishing causal from non-causal variants. Several approaches can be used for functional translation of genetic findings. These approaches have the scope to identify underlying alleles and pathways that are important in lung function and COPD. Computational methods aim at effective functional variant prediction by combining experimentally generated regulatory information with associated region of the human genome. Classically, GWAS association follow-up concentrated on manipulation of a single gene. However association data has identified genetic variants in >50 loci predicting disease risk or lung function. Therefore there is a clear precedent for experiments that interrogate multiple candidate genes in parallel which is now possible with genome editing technology. Gene expression profiling can been used for effective discovery of biological pathways underpinning gene function. This information may be used for informed decisions about cellular assays *post* genetic manipulation. Investigating respiratory phenotypes in human lung tissue and specific gene knockout mice is a valuable *in vivo* approach that can complement *in vitro* work. Herein, we review state-of-art *in silico*, *in vivo* and *in vitro* approaches that may be used to accelerate functional translation of genetic findings.

According to the World Health Organization, chronic respiratory diseases such as asthma or Chronic Obstructive Pulmonary Disease (COPD), are one of the leading causes of population morbidity and mortality (Mathers, 2008). Worldwide there are approximately 500 million people suffering from obstructive lung disease such as asthma and COPD. Asthma is a chronic inflammatory disorder associated with airway hyper responsiveness and reversible airway obstruction. COPD on the other hand is characterized by irreversible airway obstruction, and one or both of emphysema and chronic bronchitis. Although asthma and COPD are considered a public health problem in both developed and developing countries, most asthma and COPD related deaths occur in low income countries (Lozano et al., 2012). Both diseases are life-threatening and currently not curable. If patients are well managed and given the appropriate treatment, it can improve their quality of life and life expectancy. Nevertheless, the ultimate objective of the research carried out by the respiratory community is to be able to treat patients having chronic respiratory diseases by reversing the underlying pathophysiology. In order to do this successfully, it will be necessary to develop novel therapeutic agents and strategies targeting the underlying cascade of biological events leading to disease.

Functional genomics has the potential to accelerate the discovery of pathways involved in the pathogenesis of chronic respiratory diseases. The development of high-throughput genotyping and next generation sequencing (NGS) accompanied by development of the necessary bioinformatic analytical tools has allowed for massive and successful undertakings to identify genetic variation predicting respiratory disease status or lung function (Mardis, 2011) and has started to pave the way to understand the functional basis of some of these signals. If applied effectively it should result in the generation of novel hypotheses that can be verified experimentally. In the current review we describe historical and current genetic studies that have investigated respiratory phenotypes as well as *in silico, in vitro* and *in vivo* approaches to facilitate the biological and therapeutic translation of these findings.

Genetics of lung function and COPD

The last 10 years has seen a dramatic increase in the number of studies examining the genetic basis of lung function measures and COPD due to the development of

relatively cheap platforms for genotyping subjects on a genome wide basis with adequate coverage to permit genetic association signals to be detected. This has led to a number of high quality GWAS publications on a range of respiratory related phenotypes.

Definition of lung function and COPD phenotypes

Although it is beyond the scope of this review to describe the phenotypic manifestation of COPD in detail, some definitions need to be addressed before considering genetics. There are multiple measurements which can be made to assess lung function however the most commonly used are forced vital capacity (FVC) and forced expiratory volume in the first second (FEV_1). These are unequivocal measurements with a general consensus regarding their derivation using spirometry (M. R. Miller et al., 2005). FVC is the volume of air that can be expired forcibly after full inspiration and is reduced in conditions that either limit inspiration or cause air trapping. FEV_1 is the volume of air expelled in the first second of a maximal forced expiration from a position of full inspiration. FEV_1 is reduced when airway obstruction is present: this is usually defined as <80% of the predicted value based on age, gender and height. However, these are not independent variables and any condition that reduces vital capacity affects FEV_1. As in a healthy individual 70% of FVC is expelled in the first second airway obstruction is defined as a FEV_1/FVC ratio of less than 0.7. Therefore, reduced FEV_1/FVC defines airway obstruction, while FEV_1 grades its severity (Rabe et al., 2007). COPD has previously been defined and graded using the GOLD criteria which are summarized in Table 1 (Hurd & Pauwels, 2002) and has recently been updated in 2013 to consider symptoms and frequency of exacerbations ((GOLD), 2015). On the other hand forced expiratory flow between 25% and 75% of vital capacity ($FEV_{25-75\%}$) and $FEV_{25-75\%}$/FVC indices have been controversial in terms of value and relative diagnostic sensitivity. Some studies argued that $FEV_{25-75\%}$ is a sensitive index of airway obstruction (Lebecque, Kiakulanda, & Coates, 1993; Simon et al., 2010), while other studies suggest this index is of limited value in this regard (Ciprandi et al., 2012).

Table 1: Spirometry defined GOLD standards of airway limitation severity in COPD

Classification of severity of airway limitation in COPD based on post-bronchodilator FEV_1		
In patients with $FEV_1/FVC < 0.7$		
GOLD 1	Mild	$FEV_1 \geq 80\%$ predicted
GOLD 2	Moderate	$50\% \leq FEV_1 < 80\%$ predicted
GOLD 3	Severe	$30\% \leq FEV_1 < 50\%$ predicted
GOLD 4	Very severe	$FEV_1 < 30\%$ predicted

COPD is a leading cause of death and chronic morbidity throughout the world. 3 in every 1000 people are diagnosed with COPD each year and the incidence increases rapidly with age (Afonso, Verhamme, Sturkenboom, & Brusselle, 2011). The clinical presentation of COPD is diverse (Hansen, Sun, & Wasserman, 2007; Pellegrino et al., 2005). It is a progressive disabling condition characterized by airway limitation that is not reversible. Typical symptoms include dyspnea, chronic cough or sputum production but spirometry is considered to be a gold standard method for the diagnosis of COPD (Rabe et al., 2007). This is largely due to the fact that the clinical presentation of these conditions varies greatly between individuals highlighting COPD as a heterogeneous condition. Therefore COPD is clinically defined as post-bronchodilator FEV_1/FVC less than 70% predicted and FEV_1 less than 80% predicted. FEV_1 (Percent Predicted) includes adjustments for age, gender, and height (Hankinson, Odencrantz, & Fedan, 1999). Although cigarette smoking is a major risk factor for the development of COPD, only 15 to 20% of smokers manifest clinically significant COPD (J. J. Zhou et al., 2013). Inflammatory processes of COPD are located in central airways and are connected to increased mucous production, reduced ciliary clearance and a disrupted airspace epithelial barrier. Emphysema (parenchymal destruction) is a sub-phenotype of COPD and is characterized by enlargement of distal airspaces due to destruction of the airway walls (Hemminki, Li, Sundquist, & Sundquist, 2008b).

Familial aggregation of spirometry measures and COPD in families

Familial aggregation studies provide diverse but not completely consistent evidence implicating genetic factors in lung function phenotypes. The original twin study by

Hubert et al. demonstrated that lung function measures show heritability estimates to be above 70%, which suggested that most of the variation observed in the studied population is caused by genetic factors (Hubert, Fabsitz, Feinleib, & Gwinn, 1982). Redline et al. reported that monozygotic twins reared together showed intra-pair correlations of pulmonary function ranging from 0.5 to 0.7, while dizygotic twins reared together had correlations approximately one-half the magnitude of those for the monozygotic twins suggesting the presence of a significant genetic component (Redline et al., 1987). Subsequent cross-sectional studies report heritability ranging as high as 85% for FEV_1, 91% for FVC, and 45% for FEV_1/FVC ratio (Coultas, Hanis, Howard, Skipper, & Samet, 1991; Lewitter, Tager, McGue, Tishler, & Speizer, 1984; McClearn, Svartengren, Pedersen, Heller, & Plomin, 1994; Ober, Abney, & McPeek, 2001; Wilk et al., 2000). Moreover, heritability of lung function measures was also found to be consistent through time (Lewitter et al., 1984). A more recent study by Hukkinen et al. revealed heritability estimates of 32% and 36% for FEV_1, 41% and 37% for FVC, while 46% and 16% for FEV_1/FVC ratio at baseline and at later follow-up, respectively (Hukkinen et al., 2011). The same group also found that differences in environmental effects explained ~60 to ~70% of observed variation suggesting spirometry measures to be complex phenotypes, where the individual variation is strongly affected by environmental effects.

Silverman et al. have shown that the risk of COPD is ~2-3 higher in smokers who have a first degree relative affected by COPD suggesting the presence of hereditary factors contributing to COPD pathogenesis (Silverman et al., 1998). In agreement with these estimations McCloskey et al. found that the odds ratio of having COPD if a sibling has the disease is approximately five (McCloskey et al., 2001). Hemminki et al. reported that singleton siblings and twins have much higher risks of emphysema and chronic bronchitis than their parents (Hemminki, Li, Sundquist, & Sundquist, 2008a). Considering the fact that both siblings and partners usually share roughly the same environment, the study was able to provide genetic epidemiological evidence for a heritable aetiology in COPD. The heritability of chronic bronchitis which is one of the main conditions underlying COPD, was evaluated at 40% (Hallberg et al., 2008). Recently, Zhou et al. reported the first estimate of emphysema heritability at 25% (J. J. Zhou et al., 2013). Taken together these studies demonstrate a significant familial aggregation of lung function, and other

COPD related phenotypes which have motivated research efforts to identify genetic variants predisposing to airway obstruction.

Standing on the shoulders of giants: a brief historical overview of molecular genetics and functional studies in pulmonary physiology

The discovery of α-1-antitrypsin (A1AT) deficiency as a cause of emphysema

The first gene linked to emphysema was *SERPINA1* encoding serine protease α1-antitrypsin (A1AT). A1AT is a member of the serine protease inhibitor superfamily (SERPINS) and phylogenetic analyses indicate its evolutionary conservation in higher animals, nematodes, insects, plants, and viruses (Irving, Pike, Lesk, & Whisstock, 2000). The path that led to the discovery of A1AT deficiency as a risk factor for emphysema has a long history. It began with studies by Fermi and Pernossi in 1894 and later by Pugliese and Coggi in 1897 that noted protease inhibitor activity of the human plasma due to its preventative action upon trypsin. It took half a century to isolate the main inhibitor responsible for antiprotease activity which was named A1AT because of its location in the α1-globulin fraction and its ability to inhibit trypsin (for a review describing these initial discoveries see (Janciauskiene et al., 2011)). In 1963, Laurell and Eriksson reported that patients with pulmonary lesions suffering from severe respiratory deficiency had markedly reduced levels of A1AT (Laurell & Eriksson, 2013). They noted that some patients were relatives and attributed their clinical pathology to potential ‘inborn error’. At a later date, Eriksson gathered a substantial collection of A1AT cases including their families providing comprehensive evidence of the link between A1AT deficiency and emphysema (Eriksson, 1965). Subsequently, Lieberman showed that serum deficiency of A1AT is greater in homozygotes and heterozygotes with the susceptibility allele than in individuals with the normal “healthy” allele (Lieberman, 1969). He concluded that Z allele predisposes to pulmonary emphysema. The plasma levels of A1AT in individuals that have at least one copy of the Z allele is approximately 10 to 15% of the normal levels (Eriksson, 1965). Taking all these studies together it became accepted that A1AT homoeostasis is necessary for pulmonary health and that A1AT imbalance may lead to pathological decline in lung function due to excessive protease activity in the airways. Many genetic variants of

A1AT were identified some of which altered the plasma levels of A1AT while others were structural in nature (for a review see (DeMeo & Silverman, 2004)).

Later studies revealed that although most patients with A1AT deficiency suffer from emphysema, this deficiency occurs in only 1 to 3 % of the COPD population (Stoller & Aboussouan, 2005). Therefore despite the unprecedented genetic, molecular and mechanistic advances in the understanding of A1AT deficiency as causative in emphysema, it is still not clear what the underlying biological processes giving rise to COPD are in the majority of patients. Current therapeutic strategies to treat A1AT deficiency include preventative measures (smoking cessation) and, in some countries, A1AT replacement therapy.

Genetic mapping of lung function genes: linkage analyses

Genetic mapping is the process of localization of genomic loci harbouring genetic variation which can contribute to the phenotypic variation of either a continuous trait or dichotomous state. Perhaps the biggest advantage of genetic mapping is the fact that it can be performed in a hypothesis free fashion without any prior knowledge about the gene's biological functions. Therefore it allows the unbiased discovery of candidate disease susceptibility genes. The underlying principle of genetic mapping is the identification of association between a recognized genetic marker (i.e. polymorphic variant whose genomic location is known) and the phenotype. If a particular marker is showing correlated segregation with a trait it is said that this marker is in linkage with the 'causative variant' associated with the trait under study. Typically linkage studies for human traits involve genotyping families that contain multiple affected individuals for 300-400 microsatellite markers, such as short tandem repeats (STR), that span the whole genome and testing for co-segregation of a trait and marker alleles (Lander & Schork, 1994).

The proximity of a marker to a gene can be estimated by measuring the number of recombination events between them, measured as a recombination fraction (θ). The closer two loci are, the lower the probability that they will be separated during meiosis. The relationship between recombination fraction (θ) and map distance is that θ equal to 0.1 corresponds to 10cM and, although variable, 1cM roughly corresponds to one megabase of DNA in the human genome. The statistical

significance of the linkage is commonly measured by the LOD score, which is the log of the ratio of the data's likelihood given linkage to the likelihood of no linkage or a P value. A LOD score of 3.3 corresponds to a P value of 5×10^{-5}, which is the recommended threshold for genome wide scans (5% false positives at this stringency). A LOD score of 2.2 ($P = 7 \times 10^{-4}$) is suggestive of linkage, 3.6 ($P = 2 \times 10^{-5}$) corresponds to significant linkage and a score of 5.4 ($P = 3 \times 10^{-7}$) is highly significant linkage.

There have been several studies that performed genome-wide linkage scans to reveal susceptibility loci for airway obstruction and these studies focused on both lung function measures as well as COPD diagnosis. The first study to do linkage analysis for COPD related phenotypes was by Silverman et al. (Silverman, Mosley, et al., 2002). These analyses were performed on pedigrees ascertained through severe and early-onset COPD without A1AT deficiency. Following the criterion of significant linkage as LOD score above 3.3 no loci showed significant linkage. However another study by Silverman et al. in the same year, focused exclusively on spirometry measures and significant evidence for association to FEV_1/FVC was demonstrated on chromosome 2q with LOD score of 4.12 at 222 cM (Silverman, Palmer, et al., 2002). Restricting the analysis to smokers increased the statistical significance of linkage suggesting gene-by-smoking interaction as contributing to disease development. None of the other markers tested for association with FEV_1/FVC had a LOD score above 3.3. FEV_1 did not show any evidence of linkage (based on LOD score). Again, restricting the analysis to smokers increased the LOD scores suggesting gene-by-smoking interaction as contributing to disease development. Other linkage studies for lung function and COPD phenotypes are summarized in Table 2 below.

Overall, linkage studies have had a limited success in investigating association of genetic variants to lung function and COPD. This is probably due to a fact that linkage analyses, although highly effective in studying monogenic disorders (such as cystic fibrosis), are not optimal and do not have the power to identify genetic variants for complex, multifactorial diseases and traits. Importantly, the late onset of COPD makes it difficult to perform family based studies in large numbers of subjects limiting this kind of study design and approach to gene identification.

Table 2: Summary of main genetic linkage studies for lung function and COPD phenotypes (Celedon et al., 2004; DeMeo & Silverman, 2003; Palmer et al., 2003; Wilk et al., 2003).

Locus	Measure	LOD score	Reference
Chr.12 at 35cM	$FEV_{25\%-75\%}$	5.03	DeMeo et al. (2004)
Chr.6 at 184cM	FEV_1	5	Wilk et al. (2003)
Chr.2q	FEV_1/FVC	4.42	Palmer et al. (2003)
Chr.2 at 229cM	FEV_1/FVC	4.13	DeMeo et al. (2004)
Chr.2 at 221cM	$FEV_{25\%-75\%}$/FVC	4.12	DeMeo et al. (2004)
Chr.4 at 28cM	FEV_1/FVC	3.5	Wilk et al. (2003)
Chr.12 at 35cM	$FEV_{25\%-75\%}$/FVC	3.46	DeMeo et al. (2004)
Chr.19q	FEV_1	3.3	Celedon et al. (2004)
Chr.8p	FEV_1	3.3	Palmer et al. (2003)
Chr.12 at 36cM	FEV_1/FVC	3.26	DeMeo et al. (2004)

Candidate gene studies

DeMeo et al. performed a follow-up study to identify the most likely causative gene behind the linkage signal on chromosome 2 (Demeo et al., 2006). This was achieved by interrogating the transcriptomic profiling with genetic approaches. Although it was not explicitly mentioned in the publication, DeMeo et al. hypothesised that genes that appear to be differentially expressed at different stages of embryonic lung development would have a role in lung embryogenesis, which would in turn explain the observed linkage peak for chromosome 2. The limitation of this approach is that a gene that is differentially expressed during lung development does not show this gene to play a *per se* role in lung development. DeMeo et al. used a mouse microarray dataset to measure the differential expression of genes located within the linkage interval. The serpin peptidase inhibitor, clade E, member 2 (*SERPINE2*) gene was found to have the greatest change in expression across the developmental time series. Therefore *SERPINE2* was taken forward for further investigation. Researchers also had other reasons to pursue this path, including the fact that *SERPINE2* encodes a cellular and extracellular matrix-associated serine protease inhibitor known to be involved in coagulation, fibrinolysis and protease homeostasis

which is also true for A1AT. Leveraging a lung microarray dataset from a population of COPD subjects and healthy controls, *SERPINE2* expression was found to be significantly correlated with various respiratory parameters such as lung hyperinflation and post bronchodilator FEV_1 (Demeo et al., 2006). *SERPINE2* expression was only moderately increased in COPD (1.25 fold difference) and the observed effect did not meet the 5% false discovery rate. Immunohistochemistry (IHC) was used to demonstrate SERPINE2 expression in both mouse and human lung tissue. In human lung positive staining was demonstrated in healthy, emphysematous, and asthmatic lungs. Crucially, although Zhu et al. provided an independent and strong replication of genetic association of *SERPINE2* as a susceptibility gene for COPD, Chappell et al. did not replicate an association of 5 single nucleotide polymorphisms (SNPs) of the *SERPINE2* gene with COPD (Chappell et al., 2006; Zhu et al., 2007). *SERPINE2* is ~64kb in size and some of the 5 genotyped SNPs are not in strong linkage disequilibrium (LD). This suggests the presence of homologous recombination hotspots within the *SERPINE2* gene (Chappell et al., 2006). Therefore relying on 5 variants to replicate a genetic association is limiting since it may miss those SNPs that are driving the observed association but are not in linkage with genotyped SNPs. Zhu et al. used 25 SNPs in their replication of *SERPINE2* association with COPD and this highlighted the need to use a sufficient number of genotyped variants in order to properly examine a given gene (Zhu et al., 2007). What can be learned from these studies is the fact that although common haplotypes may appear to be associated with a given trait, in different populations the same SNPs may not be associated with the same phenotypes. This complex pattern of association is not surprising in multifaceted diseases or traits such as COPD and lung function measures, and points towards the importance of conducting functional studies aiming at assessing the effect of SNP variation on gene function or expression. It is particularly intriguing that, as is the case for *SERPINE2*, a region of the gene shows association with a given phenotype but another region of the same gene may not be associated at all. This further emphasizes the need to perform functional studies upon candidate SNPs to confirm relevant causative genes. More recently, in a GWAS of airway wall thickness (AWT) it was identified that SNPs within *SERPINE2* were associated with AWT and reduced levels of *SERPINE2* in human lung tissue (Dijkstra, Postma, et al., 2015).

The starting point for another study by DeMeo et al. was a publicly available microarray dataset of differentially expressed probe sets in human lung tissue stratified by lung function measures. Genomic regions appearing as differentially expressed were LD tagged and 889 SNPs from identified haplotypes were selected for association testing with COPD. Among these, 71 SNPs were significant at a nominal level (i.e. without correction for multiple comparisons) and taken forward for replication in a separate population. A stringent threshold of significance was established and only SNPs present on the iron regulatory protein 2 (*IREB2*) gene met statistical significance. Finally, *IREB2* mRNA and protein expression were shown to be significantly increased in lung tissue samples from COPD subjects in comparison to healthy controls implicating *IREB2* as a COPD susceptibility gene. Therefore, DeMeo et al. firstly combined transcriptomics as well as genomics to inform the candidate COPD SNPs selection, and secondly followed this by a genetic association study, finally showing up-regulation of the putative gene in a disease state. Although *IREB2* may act as a marker for COPD, at this stage it is not clear whether its levels are causal in relation to COPD pathogenesis or whether it is simply an epiphenomenon of other COPD causal mechanisms.

In addition to candidate genes from linkage regions there have been a large number of candidate gene studies in COPD. Many of these suffered from limited coverage of the genetic variation in target genes and small sample sizes thus limiting interpretation due to the lack of replication. These studies have been reviewed in detail elsewhere (Hardin M, 2014). Of note, a well powered study (8,300+ subjects) using a candidate gene approach identified association between SNPs in the matrix metalloprotease 12 gene (*MMP12*) and both FEV_1 and COPD risk (Hunninghake et al., 2009).

Genome-wide association (GWA) studies

The advent of genome-wide association (GWA) studies is attributable to advances in genotyping technology (Syvanen, 2005), the Human Genome Project (Venter et al., 2001) and the completion of the HapMap project (International HapMap, 2005). In GWAS, hundreds of thousands of SNPs in large populations are assayed to determine the co-occurrence of these variants with disease symptoms or with certain trait distribution (Pearson & Manolio, 2008). Importantly these SNPs are selected to

capture the maximum information on the human genome by using optimised panels that tag haplotype blocks which is made possible by our improved understanding of the human genome, thanks to the initiatives such as HapMap (International HapMap, 2005). Since GWAS is a population-based approach, most GWAS have concentrated on looking for association with common variants (>5%) and they are generally less well designed to evaluate low allele frequency variants (Hirschhorn & Daly, 2005). This is in contrast to family-based linkage approaches which are ideally suited for detecting rare genetic variants of large phenotypic effect. However GWAS generally offer greater resolution and more power in association mapping. GWAS rely on appropriate reconstruction of haplotypes based on population data however results may be misleading if this reconstruction is erroneous. This is because investigators may use one SNP (also known as tag SNP) as a proxy for a number of other SNPs present on the same haplotype. Importantly, the boundaries of haplotype blocks vary between populations of different ancestries which complicates cross-sectional comparison of studies that leveraged different ethnic populations (International HapMap, 2005). As in linkage scans, GWAS can be conducted in a hypothesis free fashion without any prior knowledge about trait or gene function. Nevertheless, as in any association mapping, they can only identify SNPs in LD with causal SNPs but cannot pinpoint the causal SNP or gene (Hirschhorn & Daly, 2005). GWAS typically examine association with 500,000+ common polymorphisms spanning the entire genome in cases and controls which therefore requires very stringent statistical thresholds (e.g. $P < 5 \times 10^{-8}$) to limit the risk of type I error.

The landscape of GWAS for lung function and COPD

Meta-analyses of FEV_1

Individual GWAS of lung function measures have identified a number of candidate SNPs potentially involved with human lung function measures FEV_1 and FEV_1/FVC. Notably, between 2006 - 2010 there were four small GWAS utilizing high throughput SNP genotyping with traits; lung function (Wilk et al., 2009; Wilk, Walter, Laramie, Gottlieb, & O'Connor, 2007) and COPD (Cho et al., 2010; Pillai et al., 2009). These studies identified several genetic loci underlying these traits including e.g. Hedgehog Interacting Protein (*HHIP*), nicotinic acetylcholine receptor 3/5 (*CHRNA3/5*) and Family with sequence similarity 13, member A (*FAM13A*) (Cho et al., 2010; Pillai et

al., 2009; Wilk et al., 2009; Wilk et al., 2007). The *CHRNA3/5* signal is likely to be driven by tobacco addiction. Importantly, while these studies demonstrated the potential to identify novel lung function and COPD loci using GWA approaches it was clear that greater statistical power was required to identify genes with confidence indicating the need for very large population sizes.

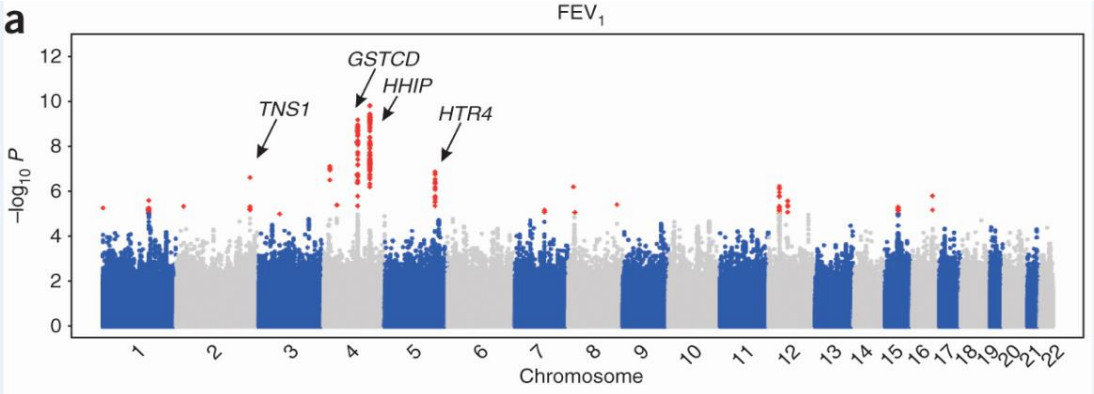
This led to the use of meta-analyses, i.e. analysing the results of many separate GWAS together to increase study power for novel gene discovery. A key component of these meta-analyses is the use of imputation whereby genetic variation not directly genotyped on the specific genotyping platform can be inferred with a measurable degree of confidence using reference genomes now available from the HapMap project and subsequently the 1000 and 10,000 genomes initiatives (Marchini & Howie, 2010). This approach makes possible the combining of genotyping data generated on a diverse number of genotyping platforms from individual studies.

The first two of these studies was the SpiroMeta and Charge consortium studies that investigate FEV_1 and FEV_1/FVC and were published as back to back papers in 2010 (Hancock et al., 2010; Repapi et al., 2010). These studies had large discovery and replication samples. In Spirometa the sample sizes were 20,288 in the discovery population and 21,209 in the replication population: imputation resulted in testing for 2.5 million genotyped and imputed SNPs (Repapi et al., 2010).

Forced Expiratory Volume in 1 second

In the SpiroMeta study four loci were reported as reaching genome wide significance for FEV_1 including common variants at novel loci; 2q35 in Tensin 1 (*TNS1)*, 4q24 in Glutathione S-Transferase, C-Terminal Domain Containing (*GSTCD)*, and 5q33 in 5-Hydroxytryptamine (Serotonin) Receptor 4 (*HTR4)* and in the previously reported loci at 4q31 in *HHIP* (see Figures 1 and 2).

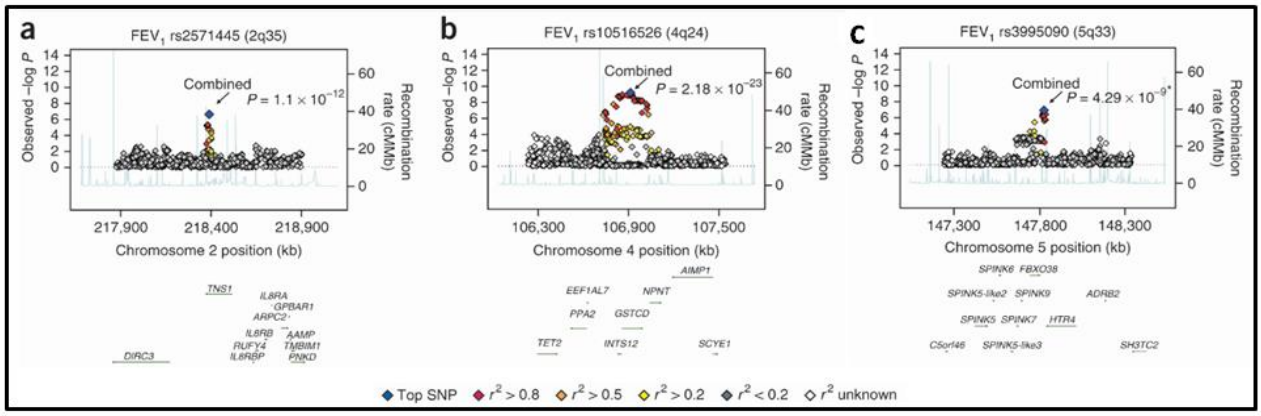
Figure 1: Manhattan plot of association results for FEV_1 from the SpiroMeta Study. Plot ordered by chromosome position. SNPs with −log10 P > 5 are indicated in red. The four loci indicated by arrows showed association with FEV_1 ($P < 5 \times 10^{-8}$) in the meta-analysis. Figure reproduced from (Repapi et al., 2010). Reprinted by permission from Macmillan Publishers Ltd: [Nature Genetics] (January; 42(1):36–44, doi:10.1038/ng.501), copyright (2010).



Focussing in on specific loci associated with FEV_1 shows the sentinel SNPs involved at loci 2q35, 4q24 and 5q33, potentially involving *TNS1*, *GSTCD* and *HTR4*, respectively (Figure 2). These data illustrate typical genetic association signal observed in GWAS as the signal can be relatively broad spanning many potential candidate genes and involving extended linkage disequilibrium making identification of the causative variant and gene a challenge.

Importantly, the Charge consortium again using a large discovery population of 20,890 subjects also showed genome wide association with FEV_1 at the *INTS12/GSTCD/NPNT* locus as well as describing additional signals (Hancock et al., 2010). This region encompasses genes Intergrator Complex Subunit 12 (*INTS12*), *GSTCD* and nephronectin (*NPNT*). In Repapi et al the most significant association was with rs10516526 within an intron of *GSTCD* (P = 2.18 x 10^{-23}), whereas Hancock and colleagues study found the most significant association to involve SNP rs17331332 located near *NPNT* (P = 4.00 x 10^{-10}) (Table 1) (Hancock et al., 2010; Repapi et al., 2010).

Figure 2: Regional association plots of three FEV_1-associated loci in the SpiroMeta study. 2q35 (a), 4q24 (b) and 5q33 (c). Statistical significance of each SNP on the −log10 scale as a function of chromosome position (NCBI build 36) in the meta-analysis. The sentinel SNP at each locus is shown in blue; the correlations (r^2) of each of the surrounding SNPs to the sentinel SNP are shown in the indicated colours. Adapted from (Repapi et al., 2010) by permission from Macmillan Publishers Ltd: [Nature Genetics] (January; 42(1):36–44, doi:10.1038/ng.501), copyright (2010).



To date, seven meta-analyses of GWAS have now been completed and identified SNP associations with FEV_1 and/or FEV_1 decline (Hancock et al., 2012; Hancock et al., 2010; Repapi et al., 2010; Soler ArtigasLoth, et al., 2011; Soler Artigas M, 2015; Tang et al., 2014; Wain L, 2015).

Sixteen genetic loci have been identified associated with FEV_1 and 18 loci have been associated with either FEV_1 decline, rate of FEV_1 change or an extreme of FEV_1 using meta-analyses approaches. The most reproducible locus associated with FEV_1 from lung function GWAS meta-analyses across all studies is the 4q24 locus (Table 3).

Table 3: Summary of meta-analyses of genome-wide associations with cross sectional and longitudinal FEV_1. All samples were of European ancestry.

Gene	Locus	Sentinel SNP	Measure	Reference	Discovery cohort	Replication cohort
ST3GAL3	1p34.1	rs121374475	FEV_1 decline	Tang et al 2014	27,249	10,476
NFIA	1p31.3	rs766488	FEV_1 decline	Tang et al 2014	27,249	10,476
ENSA	1q21.3	rs6681426	FEV_1	Soler-Artigas et al 2015	37944	54301
ESRRG/GPATCH2	1q41	rs17698444	FEV_1 decline	Tang et al 2014	27,249	10,476
BAZ2B	2q24.2	rs12692550	FEV_1 decline and #Rate of FEV_1 change	Tang et al 2014	27,249	10,476
FOSL2/PLB1	2p23.2	rs10209501	#Rate of FEV_1 change	Tang et al 2014	27,249	10,476
TNS1	2q35	rs2571445	FEV_1	Repapi et al 2010	20,288	21,209
HDAC4	2q37.3	rs12477314	FEV_1	Soler-Artigas et al 2011	48,201	46,411
MECOM	3q26.2	rs1344555	FEV_1	Soler-Artigas et al 2011	48,201	46,411
FLJ25363/MIR4445	3q13.13	rs1729588	#Rate of FEV_1 change	Tang et al 2014	27,249	10,476
INTS12	4q24	rs1172189	FEV_1	Hancock et al 2010	20,890	20,288
GSTCD	4q24	rs10516526	FEV_1	Repapi et al 2010	20,288	21,209
NPNT	4q24	rs34712979	Low vs High FEV_1 in never smokers	Wain et al 2015	50,000	34866
TET2	4q24	rs2047409	Low vs High FEV_1 in never smokers	Wain et al 2015	50,000	34866
HHIP	4q31	rs12604628	FEV_1	Repapi et al 2010	20,288	21,209
HTR4	5q33	rs3995090	FEV_1	Repapi et al 2010	20,288	21,209
ZKSCAN3	6p22.1	rs6903823	FEV_1	Soler-Artigas et al 2011	48,201	46,411
HLA-DQB1/HLA-DQA2	6p21.3	rs9274600	Low vs High FEV_1 in never smokers	Wain et al 2015	50,000	34866
CDC123	10p13	rs7068966	FEV_1	Soler-Artigas et al 2011	48,201	46,411
C10orf112/MALRD1	10p12.31	rs10764053	#Rate of FEV_1 change	Tang et al 2014	27,249	10,476
C10orf11	10q22.3	rs11001819	FEV_1	Soler-Artigas et al 2011	48,201	46,411
ME3	11q14.2	rs507211	#Rate of FEV_1 change	Tang et al 2014	27,249	10,476
TBX3	12q24.21	rs10850377	FEV_1	Soler-Artigas et al 2015	37268	52722
RBM19/TBX5	12q24.21	12:114743533	Low vs High FEV_1 in heavy smokers	Wain et al 2015	50,000	34866
TMCO3	13q34	rs2260722	FEV_1 decline	Tang et al 2014	27,249	10,476
TRIP11	14q31-q32	rs7155279	FEV_1	Soler-Artigas et al 2015	37691	54471
RIN3	14q32.12	rs117068593	FEV_1	Soler-Artigas et al 2015	34496	52572
SV2B	15q26.1	rs8027498	FEV_1 decline	Tang et al 2014	27,249	10,476
MYH11	16p13.11	rs8051319	FEV_1 decline	Tang et al 2014	27,249	10,476
KANSL1	17q21.31	rs2532349	Low vs High FEV_1 in never smokers	Wain et al 2015	50,000	34866
CACNG4	17q24.2	rs740557	FEV_1 decline	Tang et al 2014	27,249	10,476
KCNJ2/SOX9	17q24.3	rs11654749	FEV_1	Hancock et al 2012	50,047	48,201
TSEN54	17q25.1	rs7218675	Low vs High FEV_1 in never smokers	Wain et al 2015	50,000	34866
MN1	22q12.1	rs134041	FEV_1	Soler-Artigas et al 2015	37669	52770

Legend: #: participants had ≥ 3 measurements taken for FEV_1 and the rate of change (mL/year) calculated.

Meta-analyses of FEV_1/FVC

In an analogous manner to the FEV_1 analyses, SNPs associated with FEV_1/FVC have also been identified using GWAS meta-analyses. Again in the SpiroMeta and Charge consortia papers, several loci met conventional genome wide significance including variants in *HHIP* (Hedgehog interacting protein, 4q31), *AGER* (Advanced glycosylation end product-specific receptor, 6p21) and *THSD4* (Thrombospondin, Type I, Domain Containing 4, 15q25) *HTR4*, ADAM Metallopeptidase Domain 19 (*ADAM19*), Adhesion G Protein-Coupled Receptor G6 (*GPR126*) and *THSD4*. (Figures 3 and 4, Table 4) Hedgehog-interacting protein is a known morphogen shown to be involved in many developmental processes, inhibiting hedgehog signalling. SNP rs2070600 in exon 3 of the Advanced Glycosylation End Product

Receptor gene was found associated with FEV_1/FVC in 2 meta-analyses (Hancock et al., 2010; Repapi et al., 2010). AGER encodes RAGE, a multiligand receptor implicated in homeostasis, development, inflammation and diseases including diabetes, Alzheimer's and COPD (Bierhaus & Nawroth, 2009; Cheng et al., 2013; Ferhani et al., 2010; Hofmann et al., 2002; Hori et al., 1995; Sparvero et al., 2009; Wu, Ma, Nicholson, & Black, 2011; Yan et al., 1996). Associations were also identified for *GPR126* encoding an adhesion G protein coupled receptor, whilst Thrombospondin, Type I, Domain Containing 4 (*THSD4*) encodes a potential metalloproteinendopeptidase. Recently, a new meta-analyses by Soler-Artigas and colleagues has revealed 6 further novel loci and 2 novel signals associated with FEV_1/FVC (Soler Artigas M, 2015; Wain L, 2015). Novel signals in both *NPNT* and *GPR126* were identified.

Figure 3: Manhattan plot of association results for FEV_1/FVC from the SpiroMeta Study. Plot ordered by chromosome position. SNPs with −log10 *P* > 5 are indicated in red. The four loci indicated by arrows showed association with FEV_1 ($P < 5 \times 10^{-8}$) in the meta-analysis. Figure reproduced from (Repapi et al., 2010). Reproduced by permission from Macmillan Publishers Ltd: [Nature Genetics] (January; 42(1):36–44, doi:10.1038/ng.501), copyright (2010).

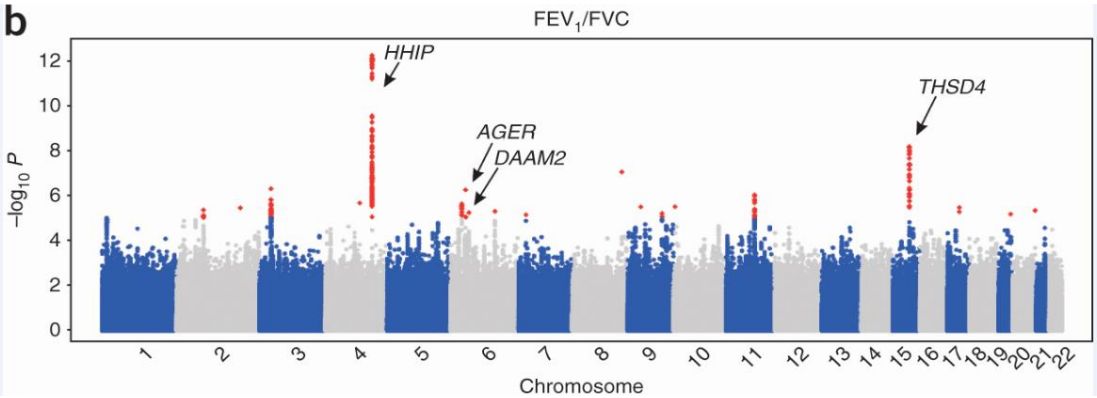


Figure 4: Regional association plots of three FEV_1/FVC-associated loci, 4q31 (a), 6p21 (b) and 15q23 (c). Statistical significance of each SNP on the −log10 scale as a function of chromosome position (NCBI build 36) in the meta-analysis. The sentinel SNP at each locus is shown in blue; the correlations (r^2) of each of the surrounding SNPs to the sentinel SNP are shown in the indicated colours. Adapted by permission from Macmillan Publishers Ltd: [Nature Genetics] (January; 42(1):36–44, doi:10.1038/ng.501), copyright (2010).

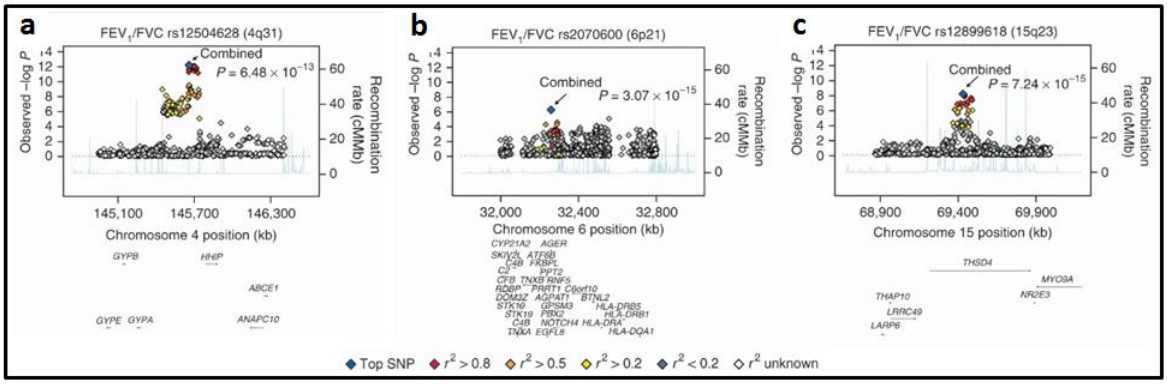


Table 4: Summary of meta-analyses of genome-wide associations with FEV_1/FVC. All samples were of European ancestry.

Gene	Locus	SNP	Measure	Reference	Discovery	Replication
RNU5F	1p34.1	rs201204531	FEV_1/FVC	Soler-Artigas 2015	34866	53760
MFAP2	1p36.1-p35	rs2284746	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
TGFB2	1q41	rs993925	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
KCNS3	2p24	rs61067109	FEV_1/FVC	Soler-Artigas 2015	37416	54341
PID1	2q36.3	rs1435867	FEV_1/FVC	Hancock 2010	20,890	20,288
DNER	2q36.3	rs7594321	FEV_1/FVC	Hancock 2012	50,047	48,201
HDAC4	2q37.3	rs12477314	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
RARB	3q24.2	rs1529672	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
FAM13A	4q22.1	rs2869967	FEV_1/FVC	Hancock 2010	20,890	20,288
NPNT	4q24	rs6856422	FEV_1/FVC	Soler-Artigas 2015	31446	49026
HHIP	4q31.21	rs12504628	FEV_1/FVC	Repapi 2010	20,288	21,209
HHIP	4q31.21	rs1980057	FEV_1/FVC	Hancock 2010	20,890	20,288
SPATA9	5q15	rs153916	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
HTR4	5q32.33	rs11168048	FEV_1/FVC	Hancock 2010	20,890	20,288
ADAM19	5q33.3	rs2277027	FEV_1/FVC	Hancock 2010	20,890	20,288
HLA-DQB1/HLA-DQA2	6p21.32	rs7764819	FEV_1/FVC	Hancock 2012	50,047	48,201
PPT2	6p21.3	rs10947233	FEV_1/FVC	Hancock 2010	20,890	20,288
NCR3	6p21.3	rs2857595	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
DAAM2	6p21.2	rs2395730	FEV_1/FVC	Repapi 2010	20,288	21,209
AGER	6p21	rs2070600	FEV_1/FVC	Repapi 2010 and Hancock 2010	20,288 and 20890	21,209 and 20288
ARMC2	6q21	rs2798641	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
GPR126	6q24.1	rs148274477	FEV_1/FVC	Soler-Artigas 2015	30398	50047
GPR126	6q24.1	rs3817928	FEV_1/FVC	Hancock 2010	20,890	20,288
PTCH1	9q22.3	rs16909898	FEV_1/FVC	Hancock 2010	20,890	20,288
ASTN2	9q33.1	rs34886460	FEV_1/FVC	Soler-Artigas 2015	37567	53920
CDC123	10p13	rs7068966	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
LRP1	12q13.3	rs11172113	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
CCDC38	12q23.1	rs1036429	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
THSD4	15q23	rs12899618	FEV_1/FVC	Repapi 2010	20,288	21,209
TEKT5	16p13.13	rs12149828	FEV_1/FVC	Soler-Artigas 2015	33999	50807
MMP15	16q13	rs12447804	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
CFDP1	16q22.2-q22.3	rs2865531	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
LTBP4	19q13.1-q13.2	rs113473882	FEV_1/FVC	Soler-Artigas 2015	32207	52907
KCNE2	21q22.12	rs9978142	FEV_1/FVC	Soler-Artigas 2011	48,201	46,411
AP1S2	Xp22.2	rs7050036	FEV_1/FVC	Soler-Artigas 2015	32285	50530

Interestingly, SNPs in *HHIP* and *HTR4* were found to be significantly associated with both FEV_1 and FEV_1/FVC. SNP rs12604628 was associated with both lung function measures (Repapi et al., 2010) whilst for *HTR4*, different SNPs were identified in each case (Hancock et al., 2010; Repapi et al., 2010).

Genome-wide association studies and meta-analyses in COPD

In addition to associations with FEV_1 and FEV_1/FVC, unsurprisingly these SNPs have also been associated with COPD susceptibility (Table 5) (Brehm et al., 2011; Castaldi et al., 2011; W. Chen et al., 2015; Cho et al., 2010; Cho et al., 2012; Kim et al., 2014; Pillai et al., 2009; Soler Artigas, Wain, et al., 2011; Soler Artigas M, 2015; Van Durme et al., 2010; Wilk et al., 2012). Moreover, recent studies have identified more refined and disease specific SNP associations in COPD subtypes including; emphysema (Kong et al., 2011; Pillai et al., 2010), COPD exacerbations (Pillai et al.,

2010), mild-moderate COPD (Hansel et al., 2013), moderate-severe COPD (Cho et al., 2014) and Chronic Bronchitis (J. H. Lee et al., 2014) (Table 5).

Table 5: Genome-wide association studies and GWAS meta-analyses of COPD

Gene	Locus	SNP	Association/Comparison	Reference
TGFB2	1q41	rs4846480	Severe COPD vs healthy smoker	Cho et al 2014
TNS1	2q35	rs2571445	COPD susceptibility	Soler-Artigas et al 2011
PID1	2q36.3	rs10498230	FEV_1/FVC associated measure - COPD susceptibility	Castaldi et al 2011
PID1	2q36.3	rs1435867	FEV_1/FVC associated measure - COPD susceptibility	Castaldi et al 2011
PID1	2q36.3	rs16825116	Lung function gene associated with COPD susceptibilty	Kim et al 2014
FAM13A	4q22.1	rs2869367	Chronic Bronchitis COPD vs smoker control	Lee et al 2014
FAM13A	4q22.1	rs2045517	Chronic Bronchitis COPD vs smoker control	Lee et al 2014
FAM13A	4q22.1	rs7671167	Chronic Bronchitis COPD vs smoker control	Lee et al 2014
FAM13A	4q22.1	rs7671167	Airway obstruction in COPD	Cho et al 2012
FAM13A	4q22.1	rs7671167	Emphysema in COPD cohort	Pillai et al 2010
FAM13A	4q22.1	rs7671167	COPD susceptibility (also in Hispanics)	Cho et al 2010 and Chen 2015
FAM13A	4q22.1	rs1903003	COPD susceptibility (also in Hispanics)	Cho et al 2010 and Chen 2015
FAM13A	4q22.1	rs2904259	Chronic Bronchitis COPD vs smoker control	Lee et al 2014
FAM13A	4q22.1	rs2609264	Lung function gene associated with COPD susceptibilty	Kim et al 2014
FAM13A	4q22.1	rs2609261	Lung function gene associated with COPD susceptibilty	Kim et al 2014
FAM13A	4q22.1	rs2609260	Lung function gene associated with COPD susceptibilty	Kim et al 2014
FAM13A	4q22.1	rs4416442	Moderate to severe and severe COPD vs healthy smoker	Cho et al 2014
FAM13A	4q22.1	rs1964516	Airway obstruction in COPD	Cho et al 2012
FLJ20184	4q24	rs17035960	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
FLJ20184	4q24	rs17036052	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
INTS12	4q24	rs11727189	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
INTS12	4q24	rs17036090	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
GSTCD	4q24	rs10516526	COPD susceptibility	Soler-Artigas et al 2011
GSTCD	4q24	rs10516526	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
GSTCD	4q24	rs11097901	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
GSTCD	4q24	rs11728716	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
NPNT	4q24	rs17036341	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
NPNT	4q24	rs17331332	FEV_1 associated measure - COPD susceptibility	Castaldi et al 2011
HHIP	4q31.21	rs13141641	Moderate to severe and severe COPD vs healthy smoker	Cho et al 2014
HHIP	4q31.21	rs12504628	COPD susceptibility	Soler-Artigas et al 2011
HHIP	4q31.21	rs13118928	FEV_1/FVC, Emphysema and Exacerbations in COPD cohort	Pillai et al 2010
HHIP	4q31.21	rs13118928	COPD susceptibility	Cho et al 2010
HHIP	4q31.21	rs13118928	COPD susceptibility	van Durme et al 2010
HHIP	4q31.21	rs1828591	COPD susceptibility	van Durme et al 2010
HTR4	5q32	rs7733088	Airway obstruction	Wilk et al 2012
HTR4	5q32	rs3995090	COPD susceptibility	Soler-Artigas et al 2011
ADAM19	5q33	rs2277027	FEV_1/FVC associated measure - COPD susceptibility	Castaldi et al 2011
ADAM19	5q33	rs1422795	FEV_1/FVC associated measure - COPD susceptibility	Castaldi et al 2011
PPT2	6p21	rs10947233	FEV_1/FVC associated measure - COPD susceptibility	Castaldi et al 2011
AGER	6p21	rs2070600	FEV_1/FVC associated measure - COPD susceptibility	Castaldi et al 2011
*KLHL7**	7p15.3	rs858249	COPD susceptibility in Hispanics	Chen et al 2015
ACN9	7q21.3	rs10231916	Lung function gene associated with COPD susceptibilty	Kim et al 2014
ACN9	7q21.3	rs10229181	Lung function gene associated with COPD susceptibilty	Kim et al 2014
RHOBTB1 - TMEM26	10q21.2	rs10761570	Decline in FEV_1 in mild/moderate COPD	Hansel et al 2013
RHOBTB1 - TMEM26	10q21.2	rs7911302	Decline in FEV_1 in mild/moderate COPD	Hansel et al 2013
EFCAB4A	11p15	rs34391416	Chronic Bronchitis COPD vs smoker control	Lee et al 2014
CHID1	11p15	rs147862429	Chronic Bronchitis COPD vs smoker control	Lee et al 2014
ANO3	11p14.2	rs7119465	Lung function gene associated with COPD susceptibilty	Kim et al 2014
*DLG2**	11q14.1	rs286499	COPD susceptibility in Hispanics	Chen et al 2015
MMP12	11q22.2	rs626750	Severe COPD vs healthy smoker	Cho et al 2014
BICD1	12p11.21	rs10844154	Emphysematous COPD	Kong et al 2011
BICD1	12p11.21	rs161976	Emphysematous COPD	Kong et al 2011
LOC100128066 - TTC6	14q21.1	rs177852	Decline in FEV_1 in mild/moderate COPD	Hansel et al 2013
RIN3	14q32.12	rs754388	Moderate to severe and severe COPD vs healthy smoker	Cho et al 2014
IREB2	15q25.1	rs1062980	COPD susceptibility	Brehm et al 2011
IREB2	15q25.1	rs13180	COPD susceptibility	Brehm et al 2011
IREB2	15q25.1	rs8034191	COPD susceptibility	Brehm et al 2011
IREB2	15q25.1	rs265606	COPD susceptibility	Brehm et al 2011
PSMA4	15q25.1	rs2036534	COPD susceptibility	Brehm et al 2011
AGPHD1 - CHRNA3/5	15q25.1	11 SNPs	Airway obstruction in ever smoker	Wilk et al 2012
CHRNA5	15q25.1	rs17486278	Airway obstruction in all ever/never smoker	Wilk et al 2012
CHRNA3/5	15q25.1	rs8034191	FEV_1, FEV_1/FVC and Emphysema in COPD cohort	Pillai et al 2010
CHRNA3	15q25.1	rs12914385	COPD susceptibility	Brehm et al 2011
CHRNA3	15q25.1	rs1051730	COPD susceptibility	Brehm et al 2011
CHRNA3	15q25.1	rs12914385	Moderate to severe and severe COPD	Cho et al 2014
CHRNA3/CHRNA5/IREB2	15q25.1	rs1062980	COPD susceptibility	Cho et al 2010
CHRNA3/CHRNA5/IREB2	15q25.1	rs13180	COPD susceptibility	Cho et al 2010
CHRNA3/5	15q25.1	rs8034191	COPD susceptibility	Pillai et al 2009
CHRNA3/5	15q25.1	rs1051730	COPD susceptibility	Pillai et al 2009
FGF7	15q21.2	rs12591300	COPD susceptibility	Brehm et al 2011
FGF7	15q21.2	rs4480740	COPD susceptibility	Brehm et al 2011
DTWD1	15q21.2	rs17404727	COPD susceptibility	Brehm et al 2011
MCTP2	15q26.2	rs8031759	Lung function gene associated with COPD susceptibilty	Kim et al 2014
AKAP1	17q22	rs886282	Lung function gene associated with COPD susceptibilty	Kim et al 2014
SOX9	17q24.3	rs17178251	Lung function gene associated with COPD susceptibilty	Kim et al 2014
SOX9	17q24.3	rs17765644	Lung function gene associated with COPD susceptibilty	Kim et al 2014
SOX9	17q24.3	rs11870732	Lung function gene associated with COPD susceptibilty	Kim et al 2014
RAB4B - EGLN2	19q13	rs7937	Airway obstruction in COPD	Cho et al 2012
RAB4B - EGLN2	19q13	rs2604894	Airway obstruction in COPD	Cho et al 2012
PDE3A	21q22.3	rs2269145	Lung function gene associated with COPD susceptibilty	Kim et al 2014

Legend: * Trend did not reach significance

Chromosomes 2q, 4q, 5q, 15q and 17q had regions with numerous SNPs associated with COPD diagnosis. SNPs with the lowest *P* values studied ($P \leq 1 \times 10^{-9}$) were in or near to *FAM13A* (Cho et al., 2010; Cho et al., 2012; Cho et al., 2014; J. H. Lee et al., 2014), *HHIP* (Cho et al., 2014; Van Durme et al., 2010), Cholinergic Receptor, Nicotinic, Alpha 3 (Neuronal) (*CHRNA3*) (Cho et al., 2014; Pillai et al., 2009), *ACN9* (renamed *SDHAF3*, Succinate Dehydrogenase Complex Assembly Factor 3) (Kim et al., 2014), RAS oncogene family member, *RAB4B* (Cho et al., 2012) and *AGPHD1* (renamed *HYKK*) a hydroxylysine kinase (Wilk et al., 2012).

Genes with significant SNPs in both Table 3 (FEV_1) and Table 5 (COPD) included *TNS1* SNP rs2571445 (Repapi et al., 2010; Soler Artigas, Wain, et al., 2011), many SNPs at the 4q24 locus (Castaldi et al., 2011; Repapi et al., 2010; Soler Artigas, Wain, et al., 2011), SNP rs12604628 in *HHIP* (Repapi et al., 2010; Soler Artigas, Wain, et al., 2011), rs3995090 in *HTR4* (Repapi et al., 2010; Soler Artigas, Wain, et al., 2011) and SNPs in SRY (Sex Determining Region Y)-Box 9 (*SOX9*) (Hancock et al., 2012; Kim et al., 2014). For SNPs associated with FEV_1/FVC and COPD, different SNPs in Transforming Growth Factor, Beta 2 (*TGFB2*) were identified (Cho et al., 2014; Soler Artigas, Wain, et al., 2011), rs2869967 in *FAM13A* (Hancock et al., 2010; J. H. Lee et al., 2014), rs12504628 in *HHIP* (Repapi et al., 2010; Soler Artigas, Wain, et al., 2011) and different SNPs in *HTR4* (Hancock et al., 2010; Soler Artigas, Wain, et al., 2011; Wilk et al., 2012). Overall these data show genetic variants that were originally identified as genetic determinants of lung function in a general population are also associated with COPD.

Refining COPD SNP associations

The identification of SNPs associated with COPD susceptibility, GOLD stages and COPD sub-types have been aided by utilising data collected in several worldwide initiatives including; COPDGene, ECLIPSE (Evaluation Of COPD Longitudinally to Identify Predictive Surrogate Endpoints), NETT (National Emphysema Treatment Trial), and GenKOLS (Genetics of COPD, Norway) studies, which provide large datasets of clinical, computed tomography (CT) and spirometric information on COPD subjects. These studies are continuing to investigate the underlying genetic and heritable factors of COPD *e.g.* using data collected from over 10,000 individuals in the case of COPDgene. With the use of CT scans, COPDGene® seeks to better

classify COPD based on the pathology observed and understand how the disease may differ from person to person. Furthering our understanding of the genetics underlying clinical features of COPD, Cho et al. recently published findings using these cohorts. By completing a GWAS of CT imaging phenotypes, 5 genetic loci were found to be associated with emphysema-related phenotypes, one locus associated with airway related phenotypes and 2 loci with gas trapping (Table 6) (Cho et al., 2015). The finding that genetic variants associated with both lung function and COPD risk also associate with emphysema related phenotypes *e.g.* AGER SNP rs2070600 is of critical importance as this i) provides greater confidence that this locus is a true association and ii) begins to help dissect the altered biological mechanisms that may underlie the association *i.e.* parenchymal destruction.

Table 6: Phenotypic genetic associations in COPD

Phenotype	Loci	SNP	Candidate Gene	Reference
% Emphysema (african american)	1p35	rs12130495	*MAN1C1*	Manichaikul et al. 2014
CMH in smokers without COPD	1q41	rs3845529	*USH2A*	Dijkstra et al. 2015
Normal LHE pattern	1q41	rs1690789	*TGFB2*	Castaldi et al 2014
Moderate centrilobular LHE pattern	1q41	rs1690789	*TGFB2*	Castaldi et al 2014
% Emphysema (chinese)	4p15.3	rs7698250	*DHX15*	Manichaikul et al. 2014
Airway WAP	4q28.1	rs142200419	*MIR2054*	Cho et al. 2015
CMH in smokers without COPD	4q28.1	rs4863687	*MAML3*	Dijkstra et al. 2015
Emphysema %LAA-950	4q31.21	rs13141641	*HHIP*	Cho et al. 2015
Normal LHE pattern	4q31.21	rs138641402	*HHIP*	Castaldi et al 2014
CMH in smokers with COPD	5p13.1 - p12	rs10461985	*GDNF*	Dijkstra et al. 2015
Emphysema %LAA-950	6p21.3	rs2070600	*AGER*	Cho et al. 2015
Gas Trapping %	6p21.3	rs2070600	*AGER*	Cho et al. 2015
% Emphysema (all races)	6p21.3	rs10947233	*PPT2*	Manichaikul et al. 2014
Emphysema %LAA-950	8p22	rs75200691	*DLC1*	Cho et al. 2015
Normal LHE pattern	11q22	rs17368659	*MMP12*	Castaldi et al 2014
Moderate centrilobular LHE pattern	11q22	rs17368582	*MMP12*	Castaldi et al 2014
CMH in smokers without COPD	12q21.2	rs1690139	*LOC100130336, LOC100131830*	Dijkstra et al. 2015
% Emphysema (all races)	12q23.1	rs7957346	*SNRPF*	Manichaikul et al. 2014
Panlobular LHE pattern	13q14	rs9590614	*VWA8*	Castaldi et al 2014
CMH in smokers without COPD	13q34	rs944899	*SOX1*	Dijkstra et al. 2015
Emphysema %LAA-950	14q32.13	rs45505795	*SERPINA10*	Cho et al. 2015
Emphysema %LAA-950	15q24	rs55676755	*CHRNA3*	Cho et al. 2015
Normal LHE pattern	15q25	rs17486278	*CHRNA5*	Castaldi et al 2014
Moderate centrilobular LHE pattern	15q25	rs114205691	*CHRNA3*	Castaldi et al 2014
Severe centrilobular LHE pattern	15q25	rs9788721	*AGPHD1*	Castaldi et al 2014
Panlobular LHE pattern	15q25	rs11852372	*AGPHD1*	Castaldi et al 2014
Severe centrilobular LHE pattern	17q11	rs379123	*MYO1D*	Castaldi et al 2014
% Emphysema (chinese)	17q25.2	rs7221059	*MGAT5B*	Manichaikul et al. 2014
% Emphysema (hispanic)	19p13.2	rs10411619	*MAN2B1*	Manichaikul et al. 2014
Upper-lower lobe ratio emphysema (hispanic)	19p13.2	rs10411619	*MAN2B1*	Manichaikul et al. 2014
Moderate centrilobular LHE pattern	19q13	rs56113850	*CYP2A6*	Castaldi et al 2014
Gas Trapping %	21q22.11	rs55706246	*LINC00310*	Cho et al. 2015

%LAA2950 = percentage low attenuation area, using a threshold of 2950 HU; Perc15 = HU at the 15th percentile of the density histogram; WAP = percentage of

the wall area compared to the total bronchial area. CMH = Chronic Mucus Hypersecretion. LHE = local histogram-based emphysema.

Also trying to further stratify the complex genetics of COPD, a recent study by Dijkstra and colleagues questioned whether the same SNPs were related to the chronic mucus hypersecretion (CMH) in smokers with and without COPD (Dijkstra, Boezen, et al., 2015). In a meta-analysis of GWAS, the top SNP associated with CMH in smokers with COPD was located in *GDNF*, whilst 4 other CMH associated SNPs in smokers without COPD were also identified (Table 6). Also in 2014, Manichaikul et al. identified associations in percentage emphysema using computed tomography scans (Table 6). Furthermore, on comparison of upper and lower lobes in the CT scans of Hispanic individuals, a difference in emphysema lobe ratio was found to be associated with a SNP in *MAN2B1* (Manichaikul et al., 2014). A separate study by Castaldi et al. used the quantification results of computed tomography emphysema to identify SNPs associated with 4 distinct patterns of emphysema (normal, moderate centrilobular, severe centrilobular and panlobular) (Castaldi et al., 2014).

The missing hereditability

To date, ~49 distinct lung function loci have been identified. However it is estimated that they explain only a modest proportion of the additive polygenic variance (4% for FEV_1, 5.5% for FEV_1/FVC and 3.2% for FVC). There are several reasons for this gap or “missing hereditability”. GWAS focus on common polymorphisms with minor allele frequency greater than 5% that span a small fraction of the human genome, but other forms of genetic variation may be important particularly rare variation and structural variation including copy number variants (S. Lee, Abecasis, Boehnke, & Lin, 2014). Therefore common SNPs either individually or taken together typically may only explain only a small fraction of phenotypic variance. Leveraging larger population sizes may improve the determination of true underlying genetic variance that accounts for phenotypic variance in lung function measures and COPD. Recently in 2013, Klimentidis et al. applied a method developed in the animal breeding field to estimate the heritability of the three main lung function measures FEV_1, FVC, and FEV_1/FVC (Klimentidis et al., 2013). From their all-SNP-inclusive analysis that considered all the genotyped SNPs they found that heritability using

SNP data are nearly identical to estimates based on pedigree information (range from 0.50 for FEV_1 to 0.66 for FEV_1/FVC). However, relying on entire whole-genome SNP data is unlikely to yield accurate approximation of heritability because only single nucleotide or copy-number variants in LD with causative variants should be included in the heritability calculation. Ultimately whole genome sequencing associations with lung function measures are likely to help refine the estimate of genetic variation contributing to the phenotypic variation.

Copy number variation in lung function and COPD

In addition to SNPs potentially contributing to human lung function and COPD susceptibility, copy number variation (duplication or deletion of regions of DNA, CNVs) is also an important area of study, with 4% of the genome harbouring copy number variants (Conrad et al., 2010). For example, previous reports have identified the Beta Defensins to have 2-10 copies in the UK population (Fode et al., 2011; Hardwick et al., 2011). CNVs are associated with a number of diseases including immune-related diseases; psoriasis and Crohn's disease (Bentley et al., 2010; Fellermann et al., 2006; Hollox et al., 2008). In 2011, Lee et al performed a GWAS of copy number variation to test for associations with lung function measures in the Korea Associated Resource (KARE) cohort (B. Y. Lee, Cho, Shin, & Kim, 2011). Interestingly, *TNS1* and *HTR4* showed evidence when leveraging copy number variation, which have previously been identified in SNP association studies (Hancock et al., 2010; Repapi et al., 2010). Recent work in a European study cohort however, did not support previously described associations for lung function measures and COPD susceptibility at the Beta Defensin1 locus (Wain et al., 2014).

In silico approaches in translational studies

In previous sections we have summarized GWAS that have been successful in identifying SNPs associated with lung function and COPD diagnosis and phenotypes with a particular focus on meta-analyses. Despite these successes there is an obvious gap between these genetic findings and their functional and mechanistic translation (Visscher, Brown, McCarthy, & Yang, 2012).

Over 90% of SNPs identified in GWAS of a range of human conditions and traits have been found to localize outside protein-coding regions and this has limited the rate of functional translation (Maurano et al., 2012). This is also true for lung function and COPD associations. This suggests that lung function and COPD associated variants are likely to be involved in normal and aberrant regulation of gene expression. Providing support for this, GWAS SNPs were found to be enriched in chromatin regulatory features (Maurano et al., 2012) and over-represented in gene expression quantitative trait locus (eQTL) studies (Nicolae et al., 2010). Since gene expression signatures are cell type specific and dependent on developmental stage and epigenetic mechanisms, as well as environmental factors, it makes interpretation of putative SNPs identified in GWAS challenging. SNPs located within intergenic regions are particularly difficult to interpret. *In silico* approaches to functionally translate genetic findings can facilitate interpretation and generate testable hypotheses.

The Encyclopaedia of DNA Elements (ENCODE) project

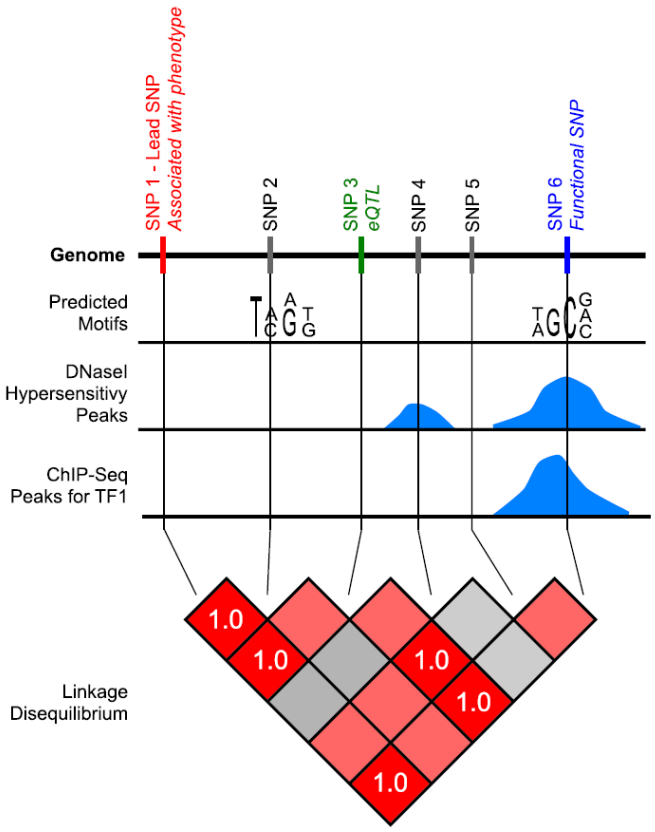
In silico translational approaches have become possible due to the widespread availability of regulatory information on the human genome generated from a diverse set of tissue and cell types. The Encyclopaedia of DNA Elements (ENCODE) project have taken a critical and leading role in this field. This initiative was launched in 2003 by the United States National Human Genome Research Institute (NHGRI) as a follow up to Human Genome Project (Consortium, 2004). This project involves a worldwide consortium and the data generated can be accessed through public databases. The main motivation for ENCODE project was that the mere sequence of a reference haploid genome only provides the physical context of hereditary information and is difficult to interpret without an additional layer of regulation that determines how the cell reads the genetic code. Also, because only 1.5% of the

genome codes for protein, the project aimed at increasing our understanding of the remaining component of the genome which traditionally was inadequately understood (Ohno, 1972). Surprisingly, one of the ENCODE project accomplishments was to demonstrate that 80% of the genome is “associated with at least one biochemical function” (Maher, 2012). The ENCODE project passed through a pilot phase (Consortium et al., 2007), and currently is in the data production phase. Readers are advised to refer to Qu and Fang for a brief review with some more details of the ENCODE project (Qu & Fang, 2013).

Integrating human genome regulatory information with candidate loci

The fundamental assumption behind all translational *in silico* approaches is that trait associated SNPs should lie within a functionally annotated region. These functional annotations involve biological or chemical events typically identified via high throughput techniques (Schaub, Boyle, Kundaje, Batzoglou, & Snyder, 2012). For example in the hypothetical locus displayed in Figure 5, six SNPs are in strong LD as demonstrated by an r^2 close to 1. Out of these polymorphisms, SNP 1 was the genotyped sentinel SNP and hence had the most significant *P*-value in the association study. However, SNP 1 does not associate with any of the available regulatory annotations making this SNP unlikely to be driving the observed association signal. On the other hand, SNP 6 associates with a DNaseI hypersensitive site (DHS), a ChIP-seq transcription factor (TF) binding site as well as being at a critical nucleotide of this TF motif signature which makes this SNP much more promising functional candidate. SNP 4 only associates with a DNaseI hypersensitive site while SNP 3 is also in a *cis*-eQTL for a given gene. Thus if we were to follow systematic approach we could prioritize polymorphisms in this region from the ‘most functional’ to ‘least functional’. As in this example, Schaub et al. report that in the majority of associations the SNP most strongly supported by functional annotation is not the sentinel SNP from GWAS but a SNP in LD with the sentinel SNP (Schaub et al., 2012).

Figure 5: Combining GWAS-associated locus with human genome regulatory annotation. Reproduced with permission from Schaub et al. (2012).



There are numerous possible regulatory features and patterns of gene regulation which are both cell type specific and can vary at different stages of development. Here, we briefly explain the features that can be considered for overlapping with GWAS loci and the possible underlying biological mechanisms that may be responsible for some of these. Because patterns of gene expression are cell type specific and highly dependent on the context, annotation of regulatory elements will vary between cell types. Therefore it is important to use those annotations that were generated in the cell types relevant for the phenotype of interest. We discuss the issue of choosing the relevant cell types in the *in vitro* approaches section. In Table 7 we summarize some of the available on-line sources that contain datasets relevant for the respiratory research community.

Table 7: Online human genome regulatory information dataset resources relevant for respiratory research

Source	Cell type	Example Datasets	Website
ENCODE project	Human Airway Epithelial Cells	rRNA-depleted RNAseq	https://www.encodeproject.org/experiments/ENCSR822SUG/
ENCODE project	Human Airway Epithelial Cells	DNase-seq on human NHBE treated with 6uM retinoic acid for 48 hours	https://www.encodeproject.org/experiments/ENCSR000EPN/
ENCODE project	Human Airway Epithelial Cells	H3K27me3, H3K4me3, H3K36me3, CTCF ChIP-seq DNaseI-seq	https://www.encodeproject.org/biosamples/ENCBS417ENC/
NIH Epigenome Project	Adult Lung	Bisulfite-Seq, mRNA-Seq, H3K36me3, H3K4me1, H3K4me3, H3K9me3, H3K27ac ChIP-seq	http://www.roadmapepigenomics.org/data/tables/adult
NIH Epigenome Project	Foetal Lung	H3K27me3, H3K4me1, H3K36me3, H3K4me3, H3K9ac, H3K9me3, ChIP-seq DNaseI-seq	http://www.roadmapepigenomics.org/data/tables/fetal

ENCODE: ENCyclopaedia Of DNA Elements (Consortium, 2004). NIH: National Institutes of Health.

Transcription Factor (TF) binding sites

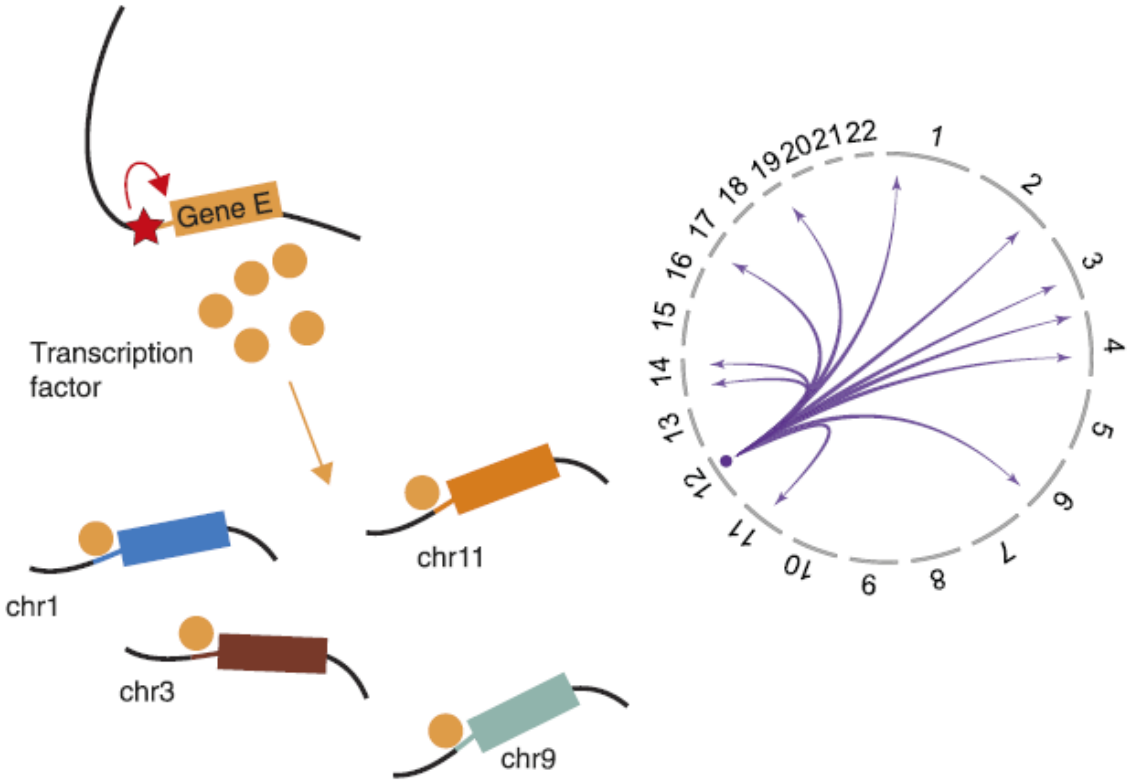
The definition of a TF is a protein that binds to genomic DNA in a sequence-specific manner and controls the rate of gene transcription (Latchman, 1997). TFs can act either individually or as cofactors to promote or repress recruitment of RNA polymerase to specific genes, thus acting as an activator or suppressor of gene expression (T. I. Lee & Young, 2000). A critical characteristic of TFs is that they contain a DNA-binding domain which mediates the binding of TF to its cognate sequences (Ptashne & Gann, 1997).

The current method of choice to identify TF binding sites is ChIP-seq (Adli & Bernstein, 2011). In ChIP-seq proteins are captured while attached to DNA by cross-linking with formaldehyde and the TF is immunoprecipitated using a specific antibody. DNA is purified from precipitated protein and sequenced by the shotgun approach using next-generation sequencing (NGS). Reads are then aligned to the reference genome and from then on sequence reads are called sequence tags. An enrichment of tag density over a particular region suggests that particular site to be the binding site of the TF. Mock immunoprecipitation using non-specific antibody may be used as a control in ChIP-seq experiments however the current recommendation of the ENCODE consortium is to use 'input control' instead. Input control is a sequenced DNA without immunoprecipitation to account for local read distribution biases (Landt et al., 2012). Demonstrating the specificity of antibody is pivotal and can be validated by either western blotting (WB) upon protein lysate or immunofluorescence (IF) combined with RNAi-mediated knockdown in cells (Landt et al., 2012). Several computational approaches have been devised to analyse ChIPseq data, the most popular of which are Model-based Analysis (MACS) (Zhang et al., 2008), Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) (Xu, Grullon, Ge, & Peng, 2014) and HOMER (Heinz et al., 2010) toolkits. However, many more programmes have been devised for ChIPseq analyses, an excellent overview of which has been published by Bailey et al. (Bailey et al., 2013).

For a given GWAS signal locus, a TF binding onto a SNP variant is highly indicative of this variant being functional. Schaub et al. has shown that TF binding (as demonstrated by ChIP-seq) is the most enriched functional element in GWAS loci when compared to the rest of investigated regulatory elements (Schaub et al., 2012).

This finding is highly indicative of the complex nature of phenotypes that were thus far studied by GWAS. One possible scenario for a mechanism behind the genetic association signal is that the causative variant is controlling the expression of the nearby gene which encodes a TF (Figure 6). Different levels of TF in turn affect the expression of TF's regulome (i.e. the set of genes regulated by the TF) which contains genes belonging to molecular pathways important for the investigated phenotype. It is important to bear in mind that these hypotheses can be only considered preliminary and would require experimental validation with respect to genetic loci associated respiratory phenotype.

Figure 6: One possible mechanism driving a genetic association signal via TF activity. Reproduced with permission from Knight (2014).



Post-translational histone tail modifications

Mapping of histone tail modifications and incorporating them onto GWAS loci is another *in silico* approach that can be used to help with the interpretation of non-coding variants. Establishing histone modification sites is similar to establishing TF

binding sites. Antibodies specific for various kinds of histone modifications are used for ChIP-seq analyses. Post-translational histone tail modifications such as histone 3, lysine 4 trimethylation (H3K4me3), H3K27me3, or H3K36me3 act as epigenetic signals regulating gene expression and chromatin modelling (Bannister & Kouzarides, 2011). Thus these modifications act in epigenetic control of gene expression and associate with different gene activities. For example, H3K4me3 tends to highlight actively transcribed loci while H3K27me3 associates with the silenced X-chromosome in females (Gibney & Nolan, 2010). Histone modifications are also used to identify the location of other functional elements such as enhancers (Shlyueva, Stampfel, & Stark, 2014). The above list is not exhaustive of all possible histone modifications and a detailed description of the majority of known histone modifications can be found reviewed in (Bannister & Kouzarides, 2011). Because the majority of SNPs in GWAS studies are non-coding and aberrant gene regulation is thought to play a predominant role in disease pathogenesis, epigenetic control is likely to take a central stage in functional translation of GWAS findings. For example, patients with genetic susceptibility for COPD may have the predisposition for low lung function due to a developmental defect. Epigenetic mechanisms were shown to play a central role in early embryological development and organogenesis (Kiefer, 2007). Therefore aberrant resetting of epigenetic marks could be due to inappropriate expression levels of epigenetic effector molecules. As is the case for other regulatory elements, patterns of histone modifications vary depending on cell type. Hence, again, the investigator is advised to use datasets from tissues relevant for the phenotype of interest.

Other regulatory elements

DHSs are locations of regulatory DNA based on NGS of genomic DNA sensitive to cleavage by DNaseI. These sites mark the accessible chromatin and overlay the majority of known regulatory elements including promoters, enhancers, silencers, insulators and imprint control locus regions. DHSs show evidence of recent functional evolutionary constraint. Interestingly, DHS in pluripotent and immortalised cells show higher mutation rates than that observed in highly differentiated cells (Thurman et al., 2012). Genomic sequences showing conservation of DNA across the species are likely to be functional. Although approximately only 1.5% of the genome is protein coding, about 8.2% is under purifying selection and hence likely to

be directly functional (Rands, Meader, Ponting, & Lunter, 2014). Genome-wide DNA methylation profiling through bisulfite conversion followed by NGS is another high throughput approach to detect a mark important in regulation of gene expression (Y. Li & Tollefsbol, 2011). The effects of DNA methylation are context dependent but they generally associate with silencing of genes in *cis*, especially if it relates to the methylation status at the CpG islands (Deaton & Bird, 2011). Finally, regions associated with short and long non-coding RNA involved in diverse regulatory roles can be identified through RNA sequencing (RNAseq) where a cDNA library is prepared with a ribosomal RNA depletion protocol. Public availability of datasets allows for functional annotation of the human genome enabling the prioritization of genes or SNPs in GWAS loci and testable hypothesis generation.

A very useful on-line tool for rapid preliminary examinations of candidate GWAS loci is the Broad Institute's HaploReg (Ward & Kellis, 2012). This software allows for the exploration of annotations of the genome at particular variants representing haplotype blocks. Information on haplotype blocks is based on the 1000 Genomes Project (Genomes Project et al., 2012). Linked SNPs can be visualized along with sequence conservation, chromatin annotation from the ENCODE and Roadmap Epigenomics projects, the effect of SNPs on gene expression from eQTL studies, as well as the effect of SNPs upon putative regulatory motifs. To illustrate this approach we have investigated sentinel SNP rs10516526 on 4q24 identified as associated with FEV_1 in SpiroMeta (Table 3) using HaploReg (Figure 7). It is apparent using the conservative r2=0.8 that this SNP is in LD with a large number of potentially functional SNPs (Figure 7).

Figure 7: HaploReg output for sentinel SNP rs10516526 identified as associated with FEV_1 in SpiroMeta.

Query SNP: rs10516526 and variants with r^2 >= 0.8

chr	pos (hg38)	LD (r^2)	LD (D')	variant	Ref	Alt	AFR freq	AMR freq	ASN freq	EUR freq	SiPhy cons	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	eQTL results	Motifs changed	GENCODE genes	dbSNP func annot
4	105661987	0.81	0.94	rs72669993	G	A	0.00	0.07	0.00	0.07							4 altered motifs	ARHGEF38	intronic
4	105664099	0.81	0.94	rs72669995	C	A	0.00	0.07	0.00	0.07							CDP	ARHGEF38	intronic
4	105668625	0.86	0.94	rs72669997	A	C	0.00	0.07	0.00	0.07							19 altered motifs	ARHGEF38	intronic
4	105668668	0.81	0.93	rs200077507	A	AGAT	0.00	0.06	0.00	0.06							Pbx3	ARHGEF38	intronic
4	105668671	0.81	0.93	rs201104128	TA	T	0.00	0.06	0.00	0.06							Pbx3	ARHGEF38	intronic
4	105669131	0.86	0.94	rs72669998	A	T	0.02	0.07	0.00	0.07							4 altered motifs	ARHGEF38	intronic
4	105669137	0.86	0.94	rs115599607	T	G	0.02	0.07	0.00	0.07							4 altered motifs	ARHGEF38	intronic
4	105672417	0.86	0.94	rs17036090	T	C	0.02	0.07	0.00	0.07						7 eQTL results	Pou2f2	ARHGEF38	intronic
4	105673359	0.89	0.96	rs72670002	G	A	0.00	0.07	0.00	0.06						Lappalainen2013,Lymphoblastoid,GSTCD	Ik-1,ZBRK1,ZBTB7A	ARHGEF38	intronic
4	105678657	0.96	0.98	rs113767110	C	T	0.02	0.07	0.00	0.07			GI			Lappalainen2013,Lymphoblastoid,GSTCD	YY1	ARHGEF38	3'-UTR
4	105679181	0.96	0.98	rs11731386	A	C	0.02	0.07	0.00	0.07			8 tissues			Lappalainen2013,Lymphoblastoid,GSTCD	4 altered motifs	ARHGEF38	3'-UTR
4	105681857	0.96	0.98	rs72671805	G	A	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD		ARHGEF38	
4	105682115	0.96	0.98	rs17036105	G	A	0.00	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD		ARHGEF38	
4	105683033	0.96	0.98	rs35370743	A	G	0.00	0.07	0.00	0.07			6 tissues	GI,PANC		Lappalainen2013,Lymphoblastoid,GSTCD	9 altered motifs	ARHGEF38	synonymous
4	105683207	0.96	0.98	rs34072732	A	G	0.02	0.07	0.00	0.07			5 tissues	9 tissues	FOXA1,P300	Lappalainen2013,Lymphoblastoid,GSTCD	Rhox11	ARHGEF38	synonymous
4	105683629	0.96	0.98	rs11728044	G	C	0.02	0.07	0.00	0.07			LNG			8 eQTL results	Cdc5,Nkx3	ARHGEF38	intronic
4	105684563	0.96	0.98	rs72671808	A	C	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	ATF3,Homez	ARHGEF38	intronic
4	105685451	0.96	0.98	rs11726569	A	G	0.02	0.07	0.00	0.07						8 eQTL results	8 altered motifs	ARHGEF38	intronic
4	105689774	0.96	0.98	rs112172458	C	T	0.00	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	5 altered motifs	ARHGEF38	intronic
4	105690380	0.96	0.98	rs72671809	A	C	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	5 altered motifs	ARHGEF38	intronic
4	105692596	0.96	0.98	rs72671810	T	C	0.00	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	6 altered motifs	ARHGEF38	intronic
4	105693869	0.96	0.98	rs146581425	CA	C	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	9 altered motifs	ARHGEF38	intronic
4	105694446	0.96	0.98	rs72671811	C	T	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	5 altered motifs	ARHGEF38	intronic
4	105695034	0.96	0.98	rs114171212	C	T	0.02	0.07	0.00	0.07					CTCF	Lappalainen2013,Lymphoblastoid,GSTCD	10 altered motifs	ARHGEF38	intronic
4	105695041	0.96	0.98	rs79571438	C	A	0.00	0.07	0.00	0.07					CTCF	Lappalainen2013,Lymphoblastoid,GSTCD	18 altered motifs	ARHGEF38	intronic
4	105695057	0.96	0.98	rs74497593	C	T	0.02	0.07	0.00	0.07					CTCF	Lappalainen2013,Lymphoblastoid,GSTCD	8 altered motifs	ARHGEF38	intronic
4	105695935	0.96	0.98	rs112119830	C	A	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	ERalpha-a,NF-I,SP1	RP11-311D14.2	intronic
4	105696204	0.94	0.98	rs12374256	G	A	0.02	0.07	0.00	0.06						8 eQTL results	Foxp3,SIX5,ZEB1	RP11-311D14.2	intronic
4	105697062	0.96	0.98	rs72671815	A	G	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	5 altered motifs	RP11-311D14.2	intronic
4	105697983	0.96	0.98	rs11727189	G	T	0.02	0.07	0.00	0.07						8 eQTL results	ERalpha-a,GR,STAT	RP11-311D14.2	intronic
4	105699826	0.96	0.98	rs72671820	C	T	0.00	0.07	0.00	0.07		BRN	BRN			Lappalainen2013,Lymphoblastoid,GSTCD	PRDM1,ZBTB33	RP11-311D14.2	intronic
4	105700351	0.96	0.98	rs17036120	G	C	0.02	0.07	0.00	0.07		BRN	BRN			Lappalainen2013,Lymphoblastoid,GSTCD	Pbx-1	RP11-311D14.2	intronic
4	105701033	0.96	0.98	rs17036123	T	C	0.02	0.07	0.00	0.07			BRN			2 eQTL results	GZF1,PRDM1	RP11-311D14.2	intronic
4	105702066	0.96	0.98	rs72671824	A	G	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	7 altered motifs	RP11-311D14.2	intronic
4	105702285	0.87	0.98	rs189541631	G	A	0.04	0.07	0.00	0.06							Rad21,Znf143,p300	RP11-311D14.2	intronic
4	105702517	0.96	0.98	rs17036125	A	T	0.02	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	5 altered motifs	RP11-311D14.2	intronic
4	105703128	0.96	0.98	rs76419734	C	T	0.00	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	GR	RP11-311D14.2	intronic
4	105703345	0.96	0.98	rs72671826	T	C	0.00	0.07	0.00	0.07						Lappalainen2013,Lymphoblastoid,GSTCD	GATA	RP11-311D14.2	intronic
4	105703372	0.96	0.98	rs17036129	C	T	0.00	0.07	0.00	0.07						8 eQTL results		RP11-311D14.2	intronic
4	105703643	0.96	0.98	rs72671828	C	T	0.02	0.07	0.00	0.07			BLD	BLD		Lappalainen2013,Lymphoblastoid,GSTCD	4 altered motifs	RP11-311D14.2	intronic
4	105707798	0.96	0.98	rs80245547	C	T	0.02	0.07	0.00	0.07		22 tissues	4 tissues	GI	USF1	Lappalainen2013,Lymphoblastoid,GSTCD	Hbp1	RP11-311D14.2	intronic
4	105710199	0.96	0.98	rs72671835	C	T	0.02	0.07	0.00	0.07		13 tissues	9 tissues	BLD		Lappalainen2013,Lymphoblastoid,GSTCD	BATF,GCNF	INTS12	intronic
4	105710234	0.94	0.98	rs201529721	T	TTA	0.00	0.07	0.00	0.06		13 tissues	9 tissues			Lappalainen2013,Lymphoblastoid,GSTCD	10 altered motifs	INTS12	intronic
4	105710235	0.94	0.98	rs200608396	T	TA	0.00	0.07	0.00	0.06		13 tissues	9 tissues			Lappalainen2013,Lymphoblastoid,GSTCD	10 altered motifs	INTS12	intronic

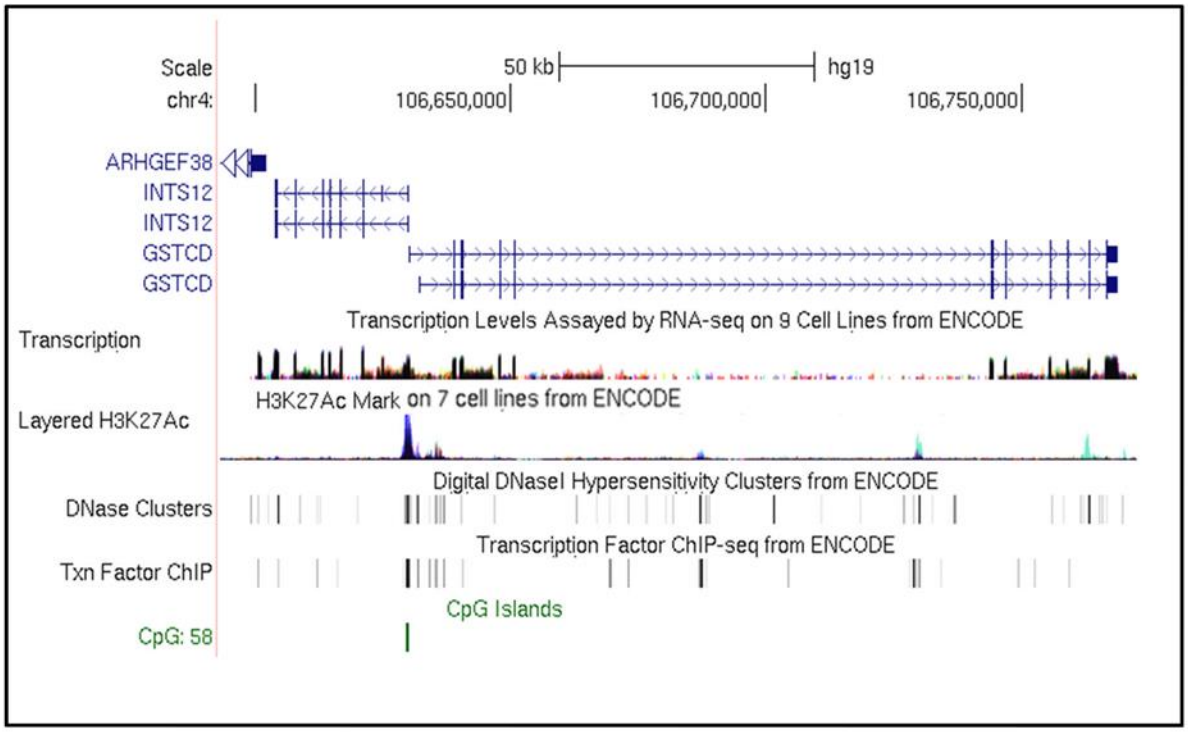
An *in silico* case study: regulatory features at 4q24 locus

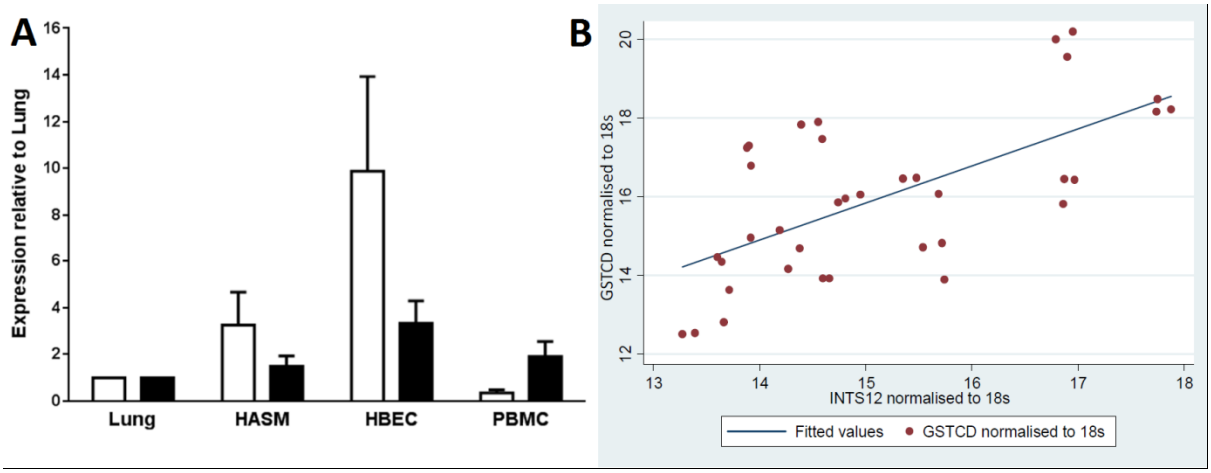
The identification of the regulatory features within the *GSTCD*/*INTS12* locus on 4q24 has been used to provide further insight into the possible regulation of gene expression in this lung function associated locus (Obeidat et al., 2013). This approach identified a possibly shared promoter region for the *INTS12* and *GSTCD* genes with evidence of H3K27 acetylation (Ac) histone mark, DNase hypersensitivity, transcription factor binding and CpG islands detected in-between the oppositely transcribed genes (Figure 8). This finding led to the hypothesis of coordinated expression of these genes which was observed in a range of airway structural cells and lung tissue (Figure 7). However, although SNPs within this putative regulatory region were predictive of *INTS12* expression this was not observed for *GSTCD* suggesting there may be some gene specific effects. Finally, it is important to bear in mind that the regulatory features found enriched at the putative regulatory region were generated in cell lines that were not derived from lung-relevant cells (Obeidat et al., 2013).

Figure 8: Regulatory features within the *GSTCD/INTS12* locus on 4q24. A: The *GSTCD/INTS12* locus is shown, annotated with RNA sequencing, H3K27Ac histone marks, DNase hypersensitivity, transcription factor binding and CpG islands (UCSC Genome Browser (http://genome.ucsc.edu/)) on the Human Feb 2009 (GRCh37/hg19) assembly. For the H3K27Ac histone marks and RNA sequence tracks, peak height is proportional to signal amplitude, with colours representing datasets in different cell backgrounds (pale blue H3K27Ac histone trace = human umbilical vein endothelial cell (HUVEC); blue/grey = K562 erythroleukaemia cells). For the DNase hypersensitivity and transcription factor binding tracks, a grey band indicates the extent of the hypersensitive region and the intensity of the band is proportional to the maximum signal strength observed in any cell line.

B: Expression of GSTCD and INTS12 mRNA in Lung and Airway cells. mRNA expression in human airway smooth muscle (HASM) cells, human bronchial epithelial cells (HBEC) and peripheral blood mononuclear cells (PBMC) is shown relative to mRNA from lung. Open bars depict GSTCD expression whereas black bars show INTS12 expression. Values shown are mean and standard error of the mean (SEM) (n=3). Only the expression of GSTCD in HBEC relative to lung was statistically significant (P<0.05).

C: Correlation between GSTCD and INTS12 ΔCt values in HASM, HBEC, PBMC and lung. The correlation coefficient between these measures was r=0.8, P<0.0001. Adapted with permission from Obeidat et al. 2013.





More advanced approaches: Unbiased analyses of genomic feature overlaps between GWA data and regulatory data

It should be noted that a large degree of non-functional overlap between GWAS loci and functional elements can be anticipated. Therefore, it is important to use an unbiased approach when investigating intersections to determine which overlaps are potentially functional and which overlaps are expected by chance. Several different bioinformatic approaches have been developed to assess the significance of overlaps.

Statistical tests (e.g. Fisher's exact test) have traditionally been used to test the overlap between two regions. In Fisher's test the number of overlaps and number of intervals unique to each feature are calculated and the test of significance is performed given the intervals coverage and the genome size. It is implemented in BEDtools suite for analysis of NGS data (Quinlan & Hall, 2010). On the other hand, the Jaccord statistic is implemented in (Favorov et al., 2012) and it measures the ratio of the number of intersecting base pairs between two regions to the number of base pairs in the union of these regions. Therefore it is a good measure if we expect one set of regions to be within another set of regions. The final statistic ranges from 0 to 1, where 0 represents no overlap and 1 represents complete overlap. Jaccord's statistic is also used in BEDtools (Quinlan & Hall, 2010). Permutation-based approaches take reference and test regions as input and calculates the observed number of overlaps between the reference and test. Test regions are then assigned to random regions with the possibility of masking certain parts, such as non-

mappable repetitive regions of the human genome. The number of overlaps between shuffled test regions and reference are re-calculated multiple times and the distribution of random overlaps are compared to the observed (Diez-Villanueva, Mallona, & Peinado, 2015). The Genomic Association Test (GAT) is another tool and uses a null model that the two sets of intervals are placed independently of one another. Similarly to the permutation-based approaches, the statistical significance is based on simulation (Heger, Webber, Goodson, Ponting, & Lunter, 2013).

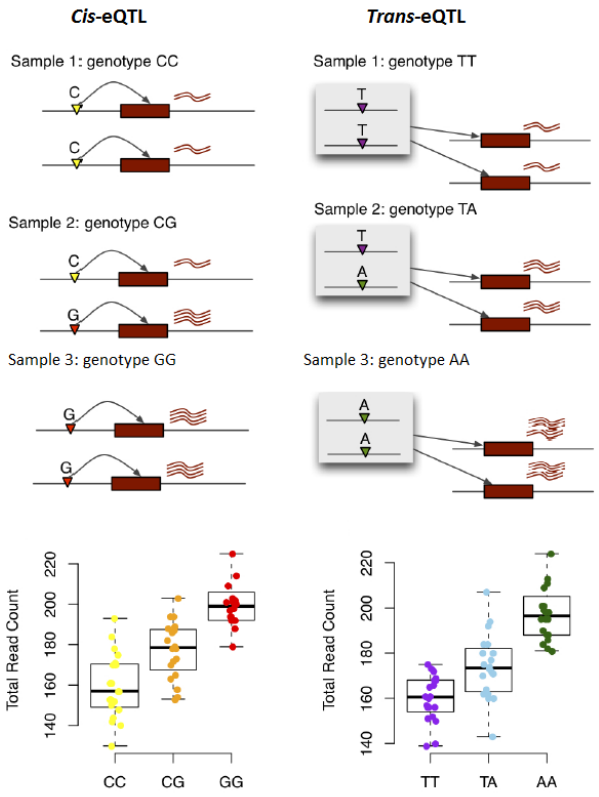
Quantitative Trait Loci (QTL) Approaches: expression QTL, splicing QTL, protein QTL and allele-specific expression

The identification of SNPs as eQTL, splicing QTL (sQTL) (Majewski & Pastinen, 2011), or protein expression QTL (pQTL) raises the possibility of those SNPs being functionally relevant and potentially causative. The most commonly used approach is to study transcript eQTLs where different human primary cells, cell lines or tissue have been characterised for both mRNA transcript expression and have been genotyped on GWA platforms. To date, eQTL studies have relied on microarray based technology with common microarrays utilizing probes located at the 3'UTR regions in order to target areas common to all annotated gene isoforms. On the other hand, exon arrays were designed by implementing probes targeting individual exons (Majewski & Pastinen, 2011). Exon array datasets can be noisy due to short probe design and probe hybridization signal saturation and hence have various analytical challenges (Kwan et al., 2008). Resolution at a splicing level has been achieved by custom arrays targeting splice-junctions (Calarco et al., 2007). Nevertheless, because of limitations of *a priori* gene annotation knowledge as well as complexity of design and analysis, microarrays are gradually being replaced by RNAseq technology (Majewski & Pastinen, 2011). RNAseq provides more accurate estimation of known or unknown transcript abundance and in a larger dynamic range (Z. Wang, Gerstein, & Snyder, 2009).

eQTL analyses particularly reinforce the notion that the observed association signal relates to the expression of either near-by (*cis*-QTLs) or distant (*trans*-QTLs) genes (Figure 9). Although these variants are sometimes said to 'control' the gene expression, the QTL SNP may not be controlling these outcomes but rather be in LD with the truly functional SNP. Nevertheless, mapping gene expression as a QTL trait

is a powerful way to identify markers correlated with differential gene expression at a population level (Rockman & Kruglyak, 2006) and can prioritize SNPs or genes in GWAS loci. Studies have now established that the SNPs identified in GWAS of human traits are enriched for eQTL SNPs (Hao et al., 2012; X. Li et al., 2015; Luo et al., 2015; Schaub et al., 2012).

Figure 9: An illustrative hypothetical example of the cis-eQTL and trans-eQTL together with their associated per-genotype gene's read counts. On the left hand side we can see the example of cis-eQTL where allele C associate with low gene expression while allele G associates with high gene expression. A heterozygous individual with both alleles is showing allele-specific expression (ASE). On the right hand side we can see the example of trans-eQTL where allele T associate with low gene expression while allele A associates with high gene expression. In contrast to cis-eQTL, trans-eQTL is not showing ASE in a heterozygous individual. Note that total per gene read counts cannot distinguish between ASE and non-ASE as reads have to be split depending on what paternal chromosome they align to. Adapted with permission from (Sun & Hu, 2013).



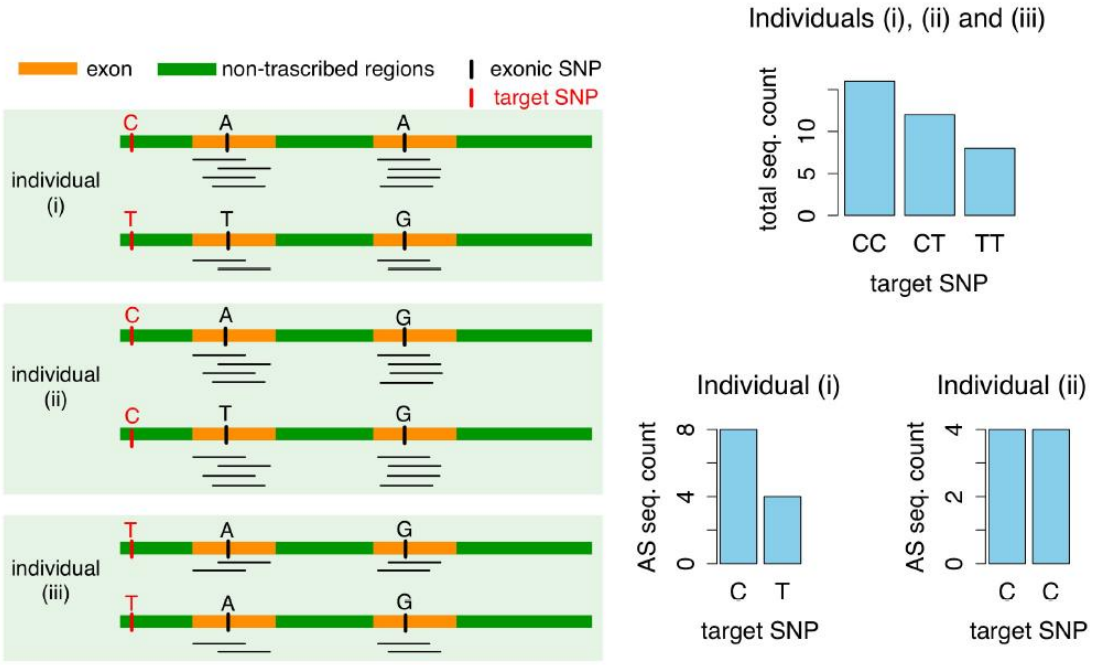
eQTL mapping can be particularly useful for identifying potential candidate genes contained within a broad association region. For example, our research group prioritized genes at locus 4q24 previously associated with the lung function measure FEV_1 and risk of COPD by identifying eQTL SNPs in a range of datasets generated in lymphoblastoid cell line, liver and brain tissue (Obeidat et al., 2013). *INTS12* is

amongst three gene candidates at this locus and was the only gene to have *cis*-eQTL with 4q24 SNPs in different cell types making this candidate gene likely to harbour genetic variation driving the association signal for lung function. These eQTL datasets have now been superseded and from a respiratory perspective excellent additional resources exist including lung tissue (Hao et al., 2012; Obeidat et al., 2015), blood cells (Westra et al., 2013) and airway epithelial cells (X. Li et al., 2015; Luo et al., 2015). The diversity of human tissues and cells and cohort sizes with eQTL data available will continue to expand.

Splicing QTL and Allele Specific Expression

Combining the population genotype or genome sequence data together with RNAseq-derived gene expression can allow for identification of not only eQTLs but also sQTLs. Mapping the RNAseq reads to the reference genome followed by counting the number of allele-specific reads that overlap with a heterozygous SNP allows detection of allele-specific expression (ASE) (Figure 10). The ability to detect ASE is a unique feature of RNA-seq which is not possible to infer using microarrays. There are major technical challenges in the reliable measurement of ASE and developed methods for the determination of ASE in RNAseq data have been reviewed extensively by Sun and Hu (Sun & Hu, 2013).

Figure 10: An example of allele-specific expression (ASE). Instead of counting total reads per gene, in determination of ASE, exonic (therefore sequenced) SNPs are imputed with not transcribed target (genotyped) SNP and read counts are performed per haplotype. Difference in read count between the haplotype, as is the case for individual 1, is indicative of ASE. Reproduced by permission from (Sun & Hu, 2013).



In silico approach case studies: eQTL enrichment at respiratory loci

A very recent study by Obeidat et al. (Obeidat et al., 2015) utilised the lung tissue eQTL dataset (n=1,111) from Hao et al. to investigate the genetic association signals identified in the SpiroMeta-Charge GWAS meta-analyses of both FEV_1 and FEV_1/FVC (Hancock et al., 2010; Hao et al., 2012; Luo et al., 2015; Repapi et al., 2010; Soler ArtigasLoth, et al., 2011). This study compared 468,300 cis and 16,677 trans-eQTL SNPs identified in the lung with the 2,419,122 SNPs interrogated in the SpiroMeta-Charge consortium papers. The analyses identified a significant enrichment for both cis and trans eQTL variants. More specifically, for the 6615 SNPs identified as associated with FEV_1, 3413 were also cis-eQTL SNPs (52%) giving a 2.7 fold enrichment (Obeidat et al., 2015). Using a similar approach for FEV_1, there was also an enrichment for trans-eQTL SNPs of 37.9%. Similarly, for the FEV_1/FVC associated SNPs there was enrichment for both cis-eQTL (2.2 fold enrichment) and trans-eQTL (12.6 fold enrichment) SNPs (Obeidat et al., 2015). More importantly, the identification of gene transcripts that the associated SNPs in

each region regulated was identified providing greater insight into the potential genes underlying the association signal (Obeidat et al., 2015).

Protein eQTL can provide novel insight beyond *cis* and *trans* mRNA eQTL

While eQTL analyses based on mRNA quantification can provide valuable information regarding the functional effect of SNPs and haplotypes, these analyses exclude the role SNPs may play in regulating post transcriptional mechanisms that may also be relevant. A good example is the recent protein eQTL analyses of serum urokinase plasminogen activator receptor (suPAR) levels (Portelli et al., 2014). uPAR has been identified as an asthma susceptibility gene (Barton et al., 2009) and has been shown to be differentially expressed in the airways (Stewart, Nijmeh, Brightling, & Sayers, 2012) and blood (Portelli et al., 2014) in asthma patients. The receptor exists as a glycosyl-phosphatidylinositol (GPI)-linked receptor and is involved in a wide range of functions including migration, proliferation and adhesion and is also involved in fibrinolysis (Portelli et al., 2014). Importantly, the receptor exists as both a soluble form generated by splicing and a soluble form generated by cleavage and these soluble forms are thought to be involved in regulating overall receptor function (Stewart & Sayers, 2009). In the protein eQTL analyses, the cleaved form of suPAR was quantified in the serum of 96 control and 512 asthma subjects and a GWAS completed using this as a quantitative trait. This approach identified a key SNP present in the promoter of human plasma kallikrein gene (*KLKB1*) as a determinant of suPAR protein levels, and subsequent biochemical analyses demonstrated that this SNP was an eQTL for *KLKB1*. We were able to show that KLKB1 subsequently cleaves uPAR from the surface of the cell providing a regulatory genetic post-translational mechanism (Portelli et al., 2014). This study, using a protein eQTL approach, therefore identified a novel way whereby an asthma susceptibility gene can be modified by a genetically driven post-translational mechanism.

In summary*, in silico* approaches to facilitate the translation of genetic association analyses can be invaluable to provide both variant and gene specific information regarding the regulation in these loci. This approach can be effectively used to generate novel hypotheses about the potential genes or variants contributing to the phenotypic variation but alone they do not constitute enough evidence. Ultimately

these hypotheses ought to be validated in *in vivo* and *in vitro* models. Candidate regulatory variants, identified through overlap with publically available functional element annotations as described require experimental testing using a diverse range of methods in order to have confidence in the observed effects.

In vivo methods to translate GWAS findings

Establishing an expression profile of specific genes in the human adult lung

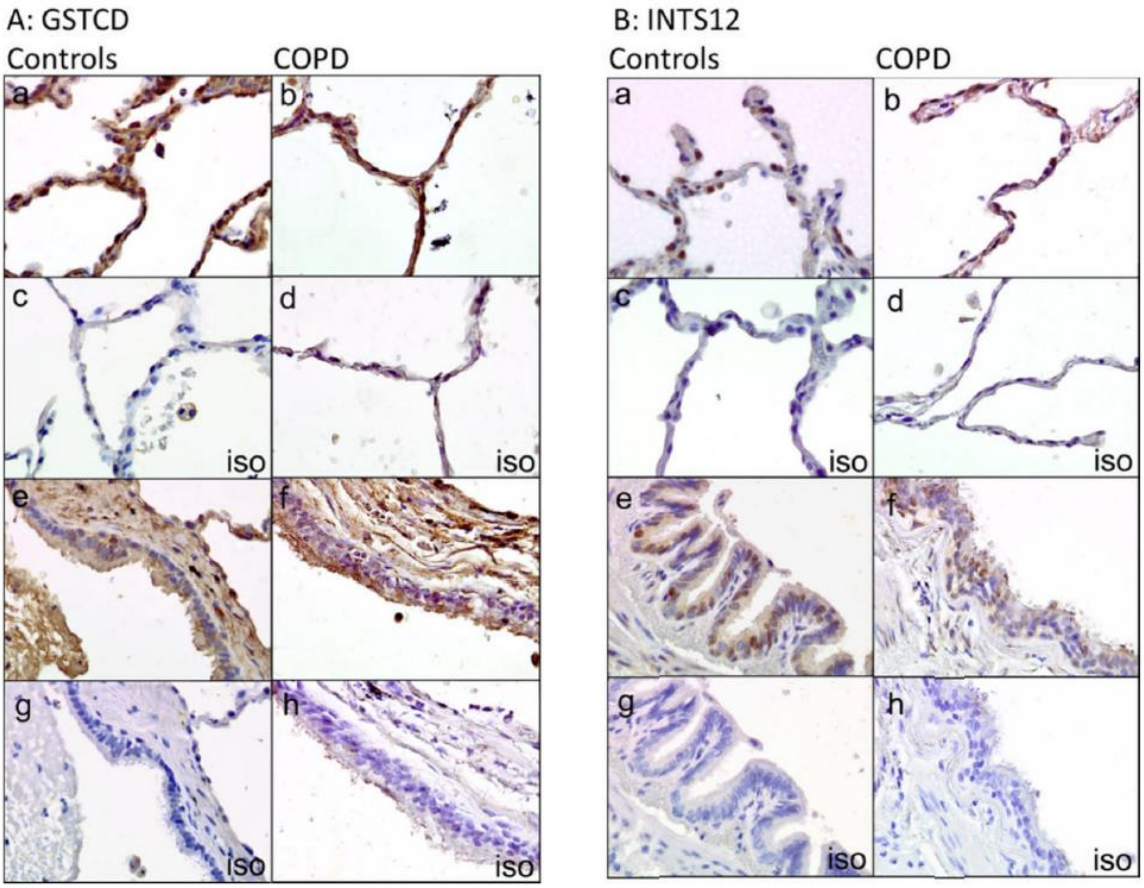
Once a potentially causative gene has been identified, it is essential that the expression profile for this gene, at both mRNA and protein levels, is established in relevant human tissue. Analysing expression at cellular and subcellular levels may provide evidence of gene function and additional insight can be derived by comparing expression between healthy and disease states and during different developmental stages

Immunohistochemisty (IHC) is a widely used tool to characterize protein expression in human tissues. In addition to providing information on protein localization within cells (nuclear, cytoplasm and/or membranous), IHC can be utilized to assess the level of protein expression based on staining intensity. The human protein atlas, a publically available database, encompasses the protein expression of 44 normal human tissues (www.proteinatlas.org, (Uhlen et al., 2005)). To better understand the role of lung function associated genes within the lungs, genes encoding protein can be interrogated for their expression profile in a range of human tissue including the lung shown on the protein atlas. Each antibody has been highly optimized and tested on many of the 44 tissues. For example, lung function associated gene *GSTCD*, was shown to have variable protein expression across multiple human tissues, with little or no staining in the central nervous system and high expression in the male reproductive system, urinary and gastro-intestinal (GI) tracts, liver and pancreas. Specifically in the lung, GSTCD had moderate membranous and cytoplasmic immunopositivity in the bronchial epithelium and lung macrophages, however expression was absent in pneumocytes (Figure 11). Another lung function associated gene at 4q24, *INTS12*, was highly expressed in the nuclei of the majority of tissues. We found moderate protein expression in pneumoctyes and macrophages of the lung whilst the bronchial epithelium was highly immunopositive (Figure 11). This staining pattern in multiple cell types potentially confirms the role as a regulator of RNA processing and/or other functions in the nucleus and implies *INTS12* function may be crucial in many cell types (Obeidat et al., 2013). Similarly, another lung function associated gene *HTR4* (5q33) was found to be highly expressed in the GI tract and neuronal cells of the central nervous system and cerebellum of the brain. In

the lung, bronchial epithelial cells were moderately immunopositive in the cytoplasm and membrane (Hodge et al., 2013). HTR4 protein expression was absent in the pneumocytes and macrophages. Thus, the different patterns of expression of genes identified by GWAS may give clues about function where this not already defined.

In addition to identifying protein expression in normal human tissues of lung function associated genes, it is important to consider whether protein expression changes in disease. The key questions here are: does protein expression increase or decrease and can the change in expression be used as a potential biological or prognostic marker? Again, as an example, our previous studies have shown no significant change in the protein expression of GSTCD and INTS12 in the small numbers of normal and COPD lung tissues tested to date (Figure 11).

Figure 11: GSTCD and INTS12 protein expression in human tissue. Immunohistochemical studies assessed GSTCD and INTS12 protein expression in tissue sections from controls and individuals with COPD. All images x40 magnification. A: Representative images of GSTCD expression in lung tissue from three healthy donors (images a and e) with matched isotype controls (iso, images c and g) and in lung tissue from three donors with COPD (images b and f) with matched isotype controls (iso, images d and h). B: Representative images of INTS12 expression in lung tissue from three control donors (images a and e) with matched isotype controls (iso, images c and g) and in lung tissue from three donors with COPD (images b and f) with matched isotype controls (iso, images d and h). Reproduced with permission from (Obeidat et al., 2013).



There are a range of other *in vitro* models to study human lung tissue, the most obvious being primary cell culture which has widely been used to look at responses in structural cells: these models are discussed later in the review, The biggest problem with these models is the lack of context, as typically these are cultures of a single cell type. To get around this issue, other approaches have been developed including the use of (i) the human lung explant model and (ii) the human precision cut lung slice (PCLS) model (Hackett, Holloway, Holgate, & Warner, 2008; Wohlsen et al., 2003). The human lung explant model can be used for a wide range of

applications including the identification of regulatory mechanisms defining the expression profile of specific or global gene expression *e.g*. in the presence of environmental triggers such as cigarette smoke, or infection such as respiratory syncytial virus. In the PCLS model, fresh lung tissue is thinly sliced and bronchial contractions can be measured in normal and diseased human lung in the presence and absence of stimuli or drugs which can provide insight into the role of specific genes or sets of genes in airway contraction.

Defining a role for lung function associated genes in human lung development

Data from the lung function GWAS meta-analyses suggest many of the identified genes may be of importance in foetal or early life lung development as the majority of the associations were still present when these analyses were restricted to the paediatric cohorts (Repapi et al., 2010). It is therefore important to question whether spatial or temporal expression of these genes early in human life and/or throughout childhood may be related to or predict lung function and disease later in adult life..

Lung development has five *in utero* stages*,* with development continuing post-natally. Organogenesis occurs during the first two stages of lung development, Embryonic and Pseudoglandular. During the Embryonic stage of lung development (4-8 weeks) formation of the major airways occurs with the lung primordium (~day 30) subdividing into the two main bronchi (~day 33). The trachea and bronchi continue to develop and the pulmonary vein and artery are also formed by this time. Lung buds differentiate from each bronchi into the pseudoglandular stage of development (6-17 weeks). Terminal bronchioles, neural networks and blood vessel continue to develop producing conducting airways. By the end of the Pseudoglandular stage, pneumocyte precursors are present as an epithelium. During the Canalicular, Saccular and Alveolar stages of development, rapid differentiation occurs. At the Canalicular stage of development, respiratory bronchioles are formed and Type II pneumoctyes differentiate into Type I pneumocytes. Surfactant is produced by Type I pneumocytes from the 25th week post conception. The level of surfactant increases until birth. At the Saccular stage, the air spaces expand and alveolar ducts are formed. At the Alveolar stage, alveolar sacs are formed through secondary septation and alveolarization which continues after birth up to around 8 years of age with the generation of new, and growth of

existing alveoli. Lung volume continues to increase with skeletal growth, and reaches a maximum between 25 and 35 years of age.

Apart from genes with prior evidence for a role in human development e.g. *PTCH1* and *HHIP* of the hedgehog signalling pathway, little is known about the role and expression of lung function associated genes during lung development (Bellusci et al., 1997; L. A. Miller et al., 2004; Pepicelli, Lewis, & McMahon, 1998). The gene expression omnibus (GEO) is a publically available resource containing large datasets which can be used by the scientific community. We have previously utilized a gene expression microarray dataset (Kho et al., 2010; Melen et al., 2011) of 38 foetal lung samples analyzed using Affymetrix U133 Plus 2 arrays to assess whether the expression of key genes is altered during normal human lung development (Hodge et al., 2013; Obeidat et al., 2013). The 38 samples analysed contained 27 lungs from the pseudoglandular stage (specifically 7-16 weeks) and 11 from the canalicular (17-26 weeks) stage of lung development. As an example, we were able to show that whilst *INTS12* expression did not change throughout the development of the lungs, *GSTCD* mRNA expression significantly decreased and *HTR4* expression increased with rising foetal age throughout the Pseudoglandular and Canalicular stages (Table 8).

Table 8: Fetal lung gene array data for INTS12, GSTCD and HTR4 expression during Pseudoglandular and Canalicular stages of lung development. Adapted with permission from (Hodge et al., 2013; Obeidat et al., 2013).

Gene name	Probe ID	Average expression	t	*P* value	Adjusted *P* value	β correlation	Significant effect
INTS12	218616_at	8.2248	-1.7318	0.0911	0.2096	-0.0040	n/s
GSTCD	220063_at	5.6144	-2.3059	0.0265	0.0862	-0.0037	n/s
GSTCD	1554518_at	5.8000	-3.2205	0.0026	0.0150	-0.0055	Decreased expression with age
GSTCD	241126_at	3.4619	1.0242	0.3120	0.4878	0.0012	n/s
GSTCD	235387_at	6.3236	-4.4776	0.0001	0.0009	-0.0115	Decreased expression with age
HTR4	216939_s_at	3.2999	0.9672	0.3393	0.5156	0.0009	n/s
HTR4	207577_at	3.6986	3.3242	0.0019	0.0121	0.0024	Increased expression with age
HTR4	207578_s_at	7.0472	0.3855	0.7020	0.8160	0.0008	n/s

Table amalgamated from (Hodge et al., 2013; Obeidat et al., 2013). Probe ID: Affymetrix U133 Plus 2 expression array probe ID; Average expression: across all samples; t: t-statistic describing differential expression; *P* value: unadjusted *P* value; adjusted *P* value: controlling for false discovery rate; β coefficient: log-odds ratio (corresponding to the mean change in gene expression per day during the studied period, 7-22 weeks of gestational age); n/s: no significance observed.

Interestingly, we have recently identified that 29 of the lung function associated genes identified by GWAS meta-analyses (Hancock et al., 2010; Loth et al., 2014; Obeidat et al., 2011; Repapi et al., 2010; Soler ArtigasLoth, et al., 2011) were differentially expressed during lung development at the mRNA level (Miller, S. unpublished data).

The Human Developmental Biology Resource (http://www.hdbr.org/) is an excellent additional source of human embryonic and foetal tissue samples within the UK with samples ranging from 3 – 20 weeks post-conception. In previous studies we have utilized this resource to aid the immunohistochemical assessments of lung function associated genes as an approach to try and reconcile differential expression we have observed at the mRNA level with protein expression data. Again, using

GSTCD, INTS12 and HTR4 as examples, by studying 12 samples between 19 days and 19 weeks post conception (6 Embryonic, 4 Pseudoglandular and 2 Canalicular stage lungs) we were able to see different patterns of protein expression. GSTCD protein expression showed a trend towards a decrease in protein expression from the Pseudoglandular to Canalicular stage (Figure 12a). Overall, INTS12 expression remained consistent throughout the 3 developmental stages studied and was localised to the nucleus in most airway cell types in keeping with the work described above on adult lung (Figure 12b). On the contrary, HTR4 expression increased from low in the early Embryonic lungs to high during the pseudoglandular stage and lower again in the subsequent Canalicular stage samples tested (Figure 12c).

Figure 12: GSTCD, INTS12 and HTR4 expression in the developing lung. (a) GSTCD expression in tissue from human foetuses at a range of developmental stages: embryonic (19 days, a and b; 21 days, c and d; 23 days, e and f); pseudoglandular (10 weeks, g and h; 12 weeks, i and j); canalicular (17 weeks, k; 19 weeks, l). Expression was observed to be increased through the pseudoglandular stage. (b) INTS12 and (c) HTR4 expression in the same panel of tissue samples as described above. Isotype controls were all negative (data not shown). Figures reproduced from (Hodge et al., 2013; Obeidat et al., 2013).

Figure 10a: GSTCD expression in the developing human lung.

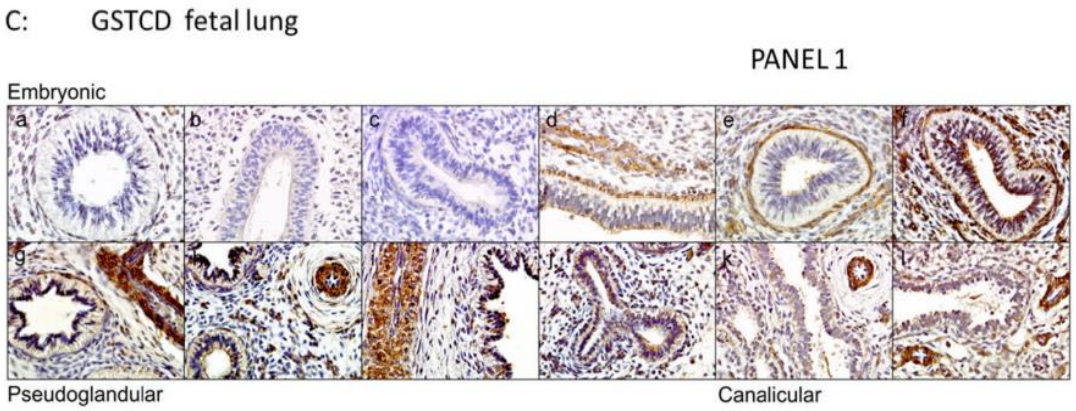


Figure 10b: INTS12 expression in the developing human lung.

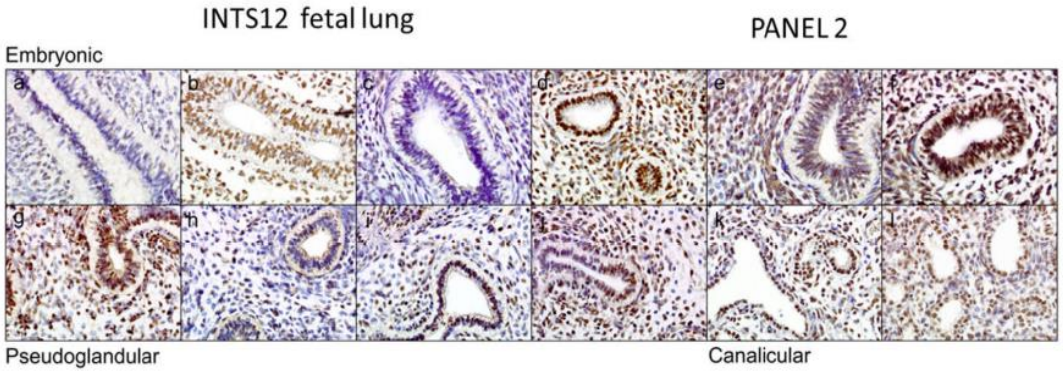
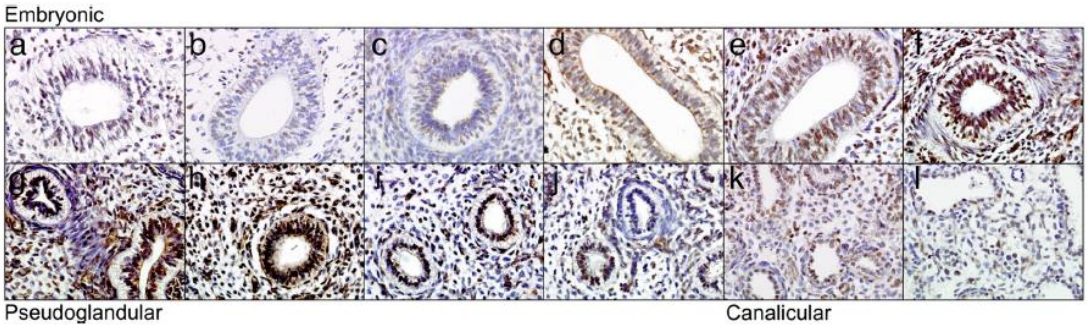


Figure 10c: HTR4 expression in the developing human lung.



Linking the mRNA and protein datasets together, *GSTCD* mRNA expression decreased throughout development which agreed with GSTCD protein expression,. *INTS12* had a high average mRNA expression of >8 and was not differentially expressed during the Pseudoglandular and Canalicular stages of development. This was in agreement with the protein expression data which showed either moderate or strong INTS12 protein expression in the majority of foetal lung samples. *HTR4* mRNA significantly increased in expression during the studied period which was inconsistent with the protein expression data: however. mRNA and protein levels do not always correlate well. For instance, miRNAs can regulate post-transcriptional gene expression through either gene silencing or targeted degradation and thus even though mRNA levels may be high, protein expression could potentially be low or absent. As this and other human tissue resources are expanded, we anticipate that the ability to interrogate the potential differential expression of candidate genes identified in genetic association studies in both foetal, early life and adult human tissue will provide unprecedented insight into the potential role of these genes in pathways involved in developmental and adult lung homeostasis. However, a major limitation of human tissue based approaches is that they are naturally restricted by lack of access to longitudinally obtained tissue samples and are often observational rather than mechanistic in nature. This has led to the extensive use of mice to define genetic mechanisms and interrogate the roles of specific genes *in vivo*, particularly using transgenic knockout mice.

Mouse models for respiratory research and the translation of genetic findings

To functionally characterize genes identified from human lung function GWAS, animal models are a useful tool to better understand the role of a given gene within the whole organism and the lung *in situ* (Dawkins & Stockley, 2001). The use of mice

in research has always been a controversial issue, and a full understanding of both advantages and disadvantages to the study of human health and disease is essential. Table 9 highlights the general advantages and disadvantages of the use of mice in research. It is interesting to note that although 99% of mouse genes have human orthologues, the chromosomal make up in mice is different to humans, with mice having 20 pairs of chromosomes rather than 23. Many complex human diseases are shared in mice and humans however drug development using pre-clinical rodent models has been limited in translation success: this is particularly true in the respiratory field. A recent review focussed on asthma research highlighted the potential over-reliance on animal models as a contributing factor to the lack of new drugs coming to the clinic (Edwards, Belvisi, Dahlen, Holgate, & Holmes, 2015).

Table 9: General advantages and disadvantages of using mice in research.

Advantages	**Disadvantages**
1. Fundamental similarity at the gene level (95%).	1. Mice have 20 pairs of chromosomes whereas humans have 23 pairs.
2. Mouse genome sequenced.	2. Is the use of a mouse to benefit human health ethically correct? Some disagree.
3. Cost effective – breeding, housing and maintenance costs are low.	3. Transgenic mouse rescue can be time-consuming, labour intensive and costly.
4. Quick research, short lifespan so get results soon (One mouse year is 30 human years).	4. Mice live for 2 years whereas humans live ~80 years therefore time disparity.
5. Short generation time.	5. 15% of gene knockouts are developmentally lethal – therefore some studies limited to embryonic development.
6. Many genes involved in complex diseases are shared between mice and humans	6. Mice are not humans and gene function may not be the same in such a different organism.
7. Large litter sizes relative to humans and so lots of samples.	7. Mice have evolved to live in infectious environments and are resistant to many infections and inflammatory stimuli unlike humans.
8. Can directly manipulate the mouse genome.	8. Mice are not a higher organism and cannot express themselves and so some results rely

	on the trained animal technicians' ability.
9. Inbred strains are well characterised.	9. Drug treatment in a mouse does not mean it will be the same in a human; side effects can be different.
10. Gene knockouts very specific and interferes with normal gene expression effectively.	10. Mice are not in natural environment, lab environment is un-natural.
11. Excellent model to test effects of drugs on.	11. Research has found responses in mice to drugs are dissimilar to that seen in humans and also the different organisms have different side effects.
12, Newer approaches to generate genetically modified mice such as CRISPR/Cas9 likely to speed up mouse transgenic research	12. Mouse inflammation is dissimilar to human inflammation and thus there are limitations on the study of inflammatory disease studies (e.g. COPD and asthma).

Respiratory research in the mouse has its own specific considerations. For instance, the basic anatomy of mice and humans lungs is not the same, with the pattern of the lung lobes and lung branching being significantly different (Table 10).

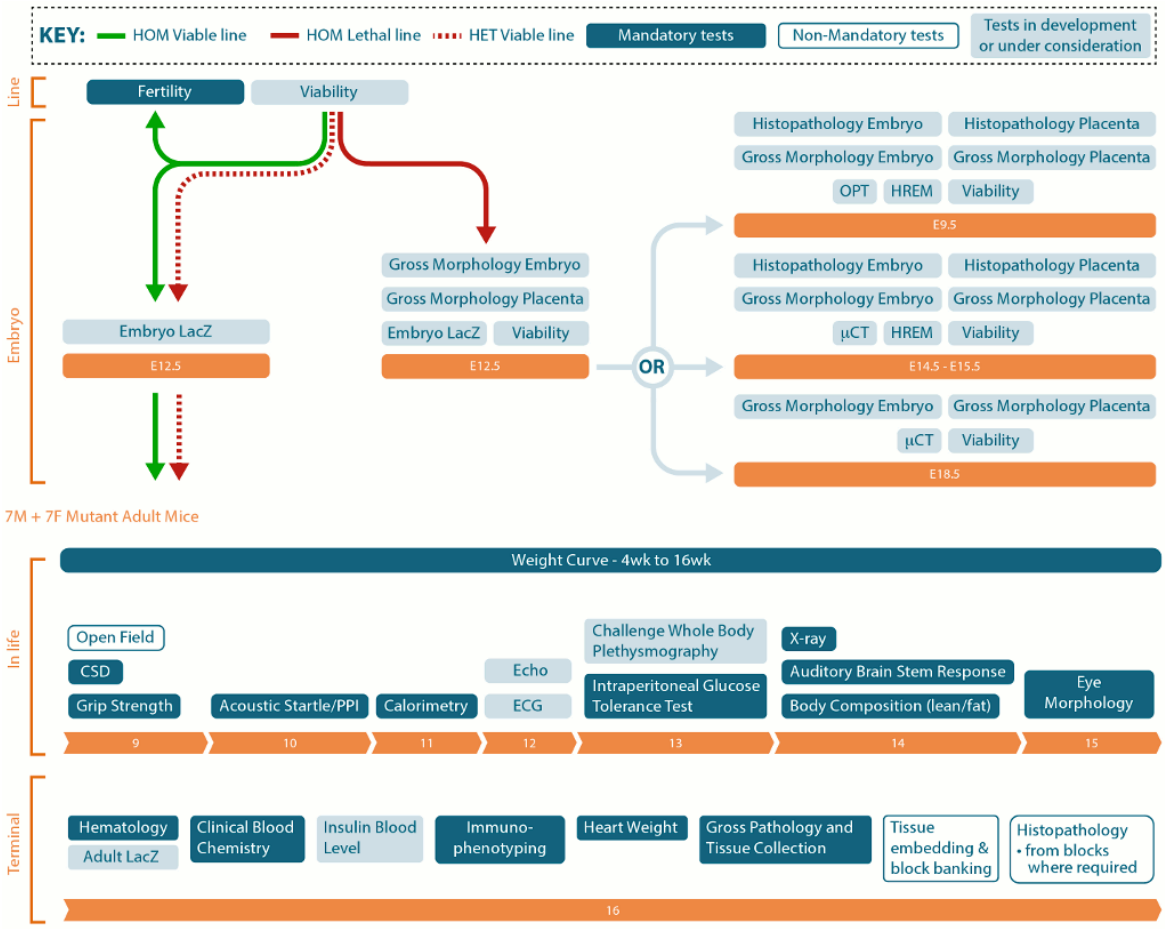
Table 10: Specific considerations for the use of mice in respiratory research.

Advantages	Disadvantages
1. Relatively easy to take non-invasive lung function measurements of mice using Plethysmographic box.	1. Lung anatomy differs between mice and humans: pulmonary lobulation is dissimilar (human right lung has 3 lobes and left has 2 lobes, whereas mice have a single left lung and 4 lobes of right lung).
2. Challenge experiments possible eg for acute lung injury, fibrosis or drug treatment	2. Humans have extensive cilia whereas mice do not.
3..Possible to sample lung tissue whenever required for analysis	3. Mice have few submucosal glands in trachea and no goblet cells which is dissimilar to human trachea.
	4. Mice do not have a cough reflex .
	5. Differences in bronchial branching, mice have 6 airway generations whereas humans have 23.
	6. Mice are obligate nose breathers whereas humans breathe through both mouth and nose.

	7. Mice do not spontaneously acquire Asthma like humans.
	8. The cellular constituents of mouse and human lungs are different.

Despite these considerations it is beyond doubt that the use of mice in basic physiology research has provided major advances in the understanding of the role of specific genes in mammalian physiology, particular when combined with transgenic approaches. In 2011, this area was given a major boost by the formation of The International Mouse Phenotyping Consortium (IMPC) which is a world-wide resource built from previous programmes including The European Mouse Disease Clinic (EUMODIC) and Mouse Genetics Project (MGP), and which has the vision to build a comprehensive catalogue of the functions of every gene in the mammalian genome (www.mousephenotype.org) (Brown & Moore, 2012). This is to be achieved by the generation and extensive phenotyping of ~20,000 knockout mice in a systematic, standardized way. Importantly, the embryonic stem cells and knockout mice are publically available shortly after their generation and data from the standardized phenotype pipeline used to characterise the biological functions of each gene is quickly available online (Figure 13).

Figure 13: IMPReSS (International Mouse Phenotyping Resource of Standardised Screens) Adult and Embryonic Phenotype Pipeline



Baseline data are available through the IPMC on some lung function related genes, incldluing INTS12. Mice with *INTS12* homozygous knock-out have pre-weaning lethality and thus provide evidence that *INTS12* has a fundamental role in mouse development and health. Additionally, *INTS12* heterozygotes were found to have increased circulating levels of magnesium. Importantly, the LacZ gene is introduced into the gene deletion and therefore can be easily visualised and act as a marker of gene expression (Figure 14).

Figure 14: LacZ expression in *INTS12* heterozygote mice characterized as part of the International Mouse Phenotyping Consortium (IMPC).

As the IMPC expands, more data will become available for GWAS relevant genes making this an extremely useful resource. While we have focussed here on IMPC it is of course important to note that many transgenic strains of mice have been generated way in individual laboratories, including mice with gene deletions for lung function associated genes. For example, recent work has focused on gaining insight into whether candidate gene *HTR4* plays a functional role in pulmonary function (House et al., 2015). Knockout of *HTR4* resulted in no difference in the histology of lungs of *HTR4*-null mice and wildtype mice. Furthermore, there was no difference in the lung volume or body weight of these mice. House et al. hypothesized that noncoding variants in *HTR4* may exert trans-regulatory effects. They identified that *HTR4*-deficient mice had a higher baseline lung resistance and increased methacholine-induced airway hyper-responsiveness (AHR) compared to wild type littermates, however these effects were modest. The *HTR4*-deficient mice were also more sensitive to serotonin-induced AHR. Interestingly, challenges with bacterial lipopolysaccharide (LPS), bleomycin, which promotes lung fibrosis, and house dust mite (HDM) to mimic an asthma phenotype were also performed. The pulmonary function and cytokine profiles of HTR4-deficient mice only modestly differed from their wild-type counterparts in these models, for example with a reduced IL1β responses in *HTR4*-/- following LPS instillation in the lungs (House et al., 2015). Thus, the group provided some evidence for a causal relationship between GWAS identified *HTR4* and pulmonary function, with alterations in baseline lung function and increased AHR in *HTR4*-null mice but no differences in lung histology.

Recently, Jin et al. generated a *FAM13A*-mutant mouse and found the mice to be viable and healthy (Jin et al., 2015). As discussed above, SNPS within the *FAM13A* locus have reproducibly been associated with FEV_1/FVC. No morphological differences were identified in the lungs of *FAM13A* homozygous mutant mice versus littermate WT controls (Jin et al., 2015). Additional experiments suggested that FAM13A is involved in Wnt signalling and the authors concluded that under normal physiological conditions *FAM13A* is not an essential gene for developing normal lung function.

One transgenic mouse strain that has dramatically facilitated our translation of human GWAS findings is the receptor for advanced glycation end products (RAGE) gene deletion mouse. Meta-analyses of GWAS have identified a number of polymorphisms in the Advanced Glycation End Product-Specific gene (AGER/RAGE) with the pivotal SNP appearing to be in exon 3 rs2070600 (Gly82Ser, (C/T)) which is associated with FEV_1/FVC (See Table 4). Importantly, this SNP has been shown to be a protein eQTL for the soluble form of the receptor in human serum (Cheng et al., 2013; Gaens et al., 2009). RAGE deficient mice when exposed to cigarette smoke were protected (albeit modestly) from the emphysema like phenotype that developed in the lung including airspace enlargement when compared to WT littermate controls (Sambamurthy, Leme, Oury, & Shapiro, 2015). This protection was at least in part thought to be driven by a reduction in the influx of neutrophils into the airways in RAGE-/- mice (Sambamurthy et al., 2015). This study therefore shows a key link between a gene associated with lung function and a relevant COPD phenotype.

The use of conditional knockout mice may also be helpful, especially where the global knock out is lethal. Currently, the Cre-Lox recombination system is the most commonly used organ specific gene knockout technique. LoxP sites are introduced around the gene of interest and once in the germline, these mice can be crossed with mice containing germline Cre-recombinase. The gene can then be deleted in a tissue specific manner. One great advantage of the use of the Cre-Lox system is that more mice are viable and survive longer than when a gene is knocked out in the whole organism and embryonic or early postnatal lethality can be overcome.

Inducible knock out mice can also be produced whereby gene expression can be turned on or off by doxycycline or tetracycline-regulated systems (Gunther et al., 2002; Shockett, Difilippantonio, Hellman, & Schatz, 1995). A way of producing lung specific inducible knockout mice is by exploiting the cell make-up of the airways, i.e. where Club cells (originally named Clara cells) are exclusively present (Niden, 1967). Bertin et al. (and others) have produced transgenic strains of mice whereby the expression of Cre recombinase was under the expression of Clara Cell 10kD Protein (CC10) (originally (Clara Cell Secretory Protein (CCSP)) and this knockout was thus Clara Cell specific (Bertin, Poujeol, Rubera, Poujeol, & Tauc, 2005; H. Li et al., 2008). The use of other cell specific promoters have also been used; SFTPC (Surfactant Protein C (SP-C)) is used to conditionally express genes in the distal lung structures (Perl et al., 2005). Although the use of conditional cell-specific gene expression systems have many experimental advantages to aid the study of the respiratory system, caution is needed and some researchers in the field have identified that the expression of the tetracycline-transactivator gene causes emphysema-like changes in mice (Sisson et al., 2006), thus identifying a need for the careful consideration and use of experimental controls.

Interestingly, these approaches have been applied to further define the role of RAGE in lung homeostasis (Stogsdill et al., 2013). Conditional overexpression of RAGE was induced under the control of a human SP-C promoter which should lead to enhanced expression in the alveolar epithelial cells. In these mice there was the development of an emphysema-like phenotype demonstrating that alterations in RAGE homeostasis in adult mice following normal lung development is detrimental (Stogsdill et al., 2013).

In vitro approaches using human cells to translate GWAS findings

In the next sections we focus on the use of human model cell systems to further define and translate genetic association signals.

Choosing the cell type to work with

As discussed above it is important to use cell types relevant to the phenotype of interest in order to avoid misleading biological interpretation. Currently available regulatory genome annotations have been generated in diverse sets of primary cells and immortalized cell lines (Consortium, 2004), each with its advantages and disadvantages. In the coming years annotations of un-differentiated and differentiated embryonic cells are likely to rise in prominence due to the likely developmental basis of many of the traits for which genetic association studies have been conducted. Similarly, the use of induced pluripotent stem cells (iPSCs) to re-generate lineages of differentiated human cells has potential, especially as cells can be derived from individuals of known genotype for a region of interest on a known genetic background.

Primary cells are most representative of the human tissues from which they were isolated. However, their phenotype is often context specific and there is extensive literature that primary human cells when removed from the body may alter in phenotype and "dedifferentiate" and so care is needed in interpretation and the use of these cells. Isolation of primary cells or precursor stem cells is inevitably more challenging than the use of immortalised cell lines. For example obtaining primary human bronchial epithelial cells is achieved by bronchoscopy which is invasive for the patient and requires local anaesthesia. It is also difficult to obtain a homogenous population of cells and this may require additional sorting of cells by fluorescently-based preparative procedures. In order to have confidence in the cell type used, characterization of cells is required. This can be achieved by immunofluorescent staining of cell-type specific markers. In general, homogenous primary cell preparations are limited in the numbers obtained and in life span but do provide a useful tool.

There are a number of immortalised cell lines which are often used in respiratory research. Immortalised cell lines have been well characterized by public consortia. Some of these cell lines were shown to retain the properties of the original primary

cells from which they were derived (Bocchini et al., 1992), although it is not advisable to assume that a cell line has the same gene expression signature as the original unmodified cells and this should be examined on a case-by-case basis. Different chromosomal re-arrangements, alteration of chromatin, DNA methylation as well as histone methylation patterns may be observed in cell lines (Masters, 2000). For example we have observed that BEAS2B-R1 cells, which are frequently used as a model by those interested in bronchial epithelial cell biology have 68 chromosomes (unpublished observation). These transformations may lead to artificial biochemical activities and misleading functional genome annotation.. The advantage of cell lines is their ease of propagation allowing access to a large numbers of cells for analyses.

If it is not clear what kind of cell type to utilize, a *de novo* identification of target cells may be applied (Maurano et al., 2012). In this approach annotations from all possibly available cell types are systematically integrated into GWAS loci and cell types showing prominent enrichment of the considered functional element can be deemed relevant for the phenotype of interest. This is a powerful approach if no extensive *a priori* knowledge about the studied phenotype is available. For example, Maurano et al. identified IL-17 producing T helper cells as a target cell type for Crohn's disease (Maurano et al., 2012) using this method. This approach can in principle be applied to any phenotype for which a genetic association study has been undertaken.

Investigating non-coding loci

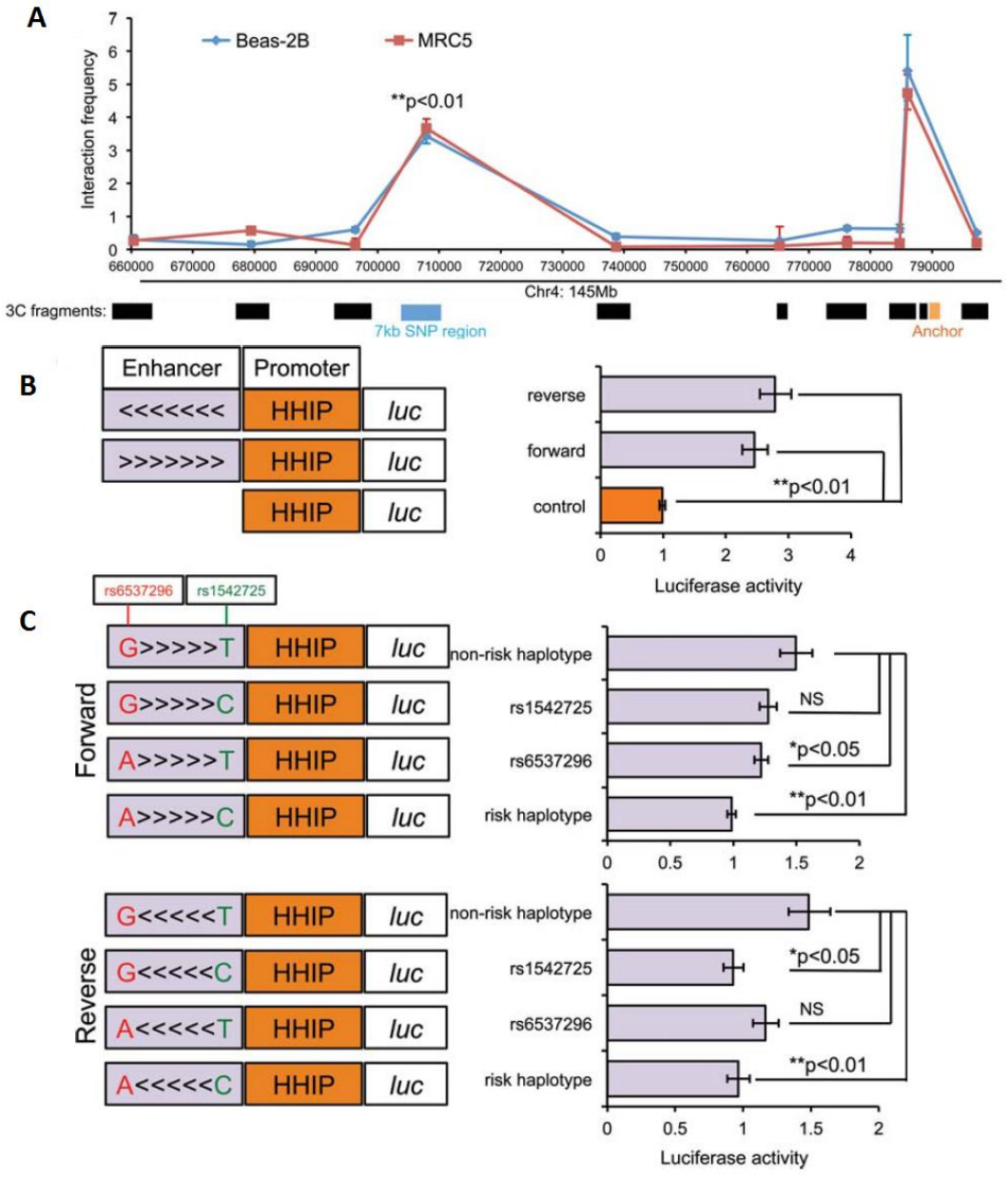
In addition to the approaches discussed above, there are many *in vitro* tools which can be used by researchers wishing to functionally investigate non-coding candidate variants. Of those which have been around for a number of years, luciferase or GFP reporter assays can be used for studying transcriptional and post-transcriptional gene regulation by directly measuring the functional activity of the controlling elements. Causative inferences can then be made by applying mutagenesis to the candidate regulatory regions. Electrophoretic Mobility Shift Assays (Hellman & Fried, 2007). are also often used to screen nuclear extract or DNA sequences for specific protein-DNA binding activity.

The spatial organization of chromosomes: chromosome conformation capture

Studying the spatial organization of chromosomes is crucial if we are to understand the regulation of gene expression. Because of the epistatic effects of genetic variants on the expression of distant genes (Hemani et al., 2014), the emergence of a new tool called Capture-C may prove to be useful in elucidating these relationships (Hughes et al., 2014). Capture-C is a further development of chromosome conformation capture, or 3C, which is used to analyse the organization of chromosomes. It utilizes oligonucleotide capture technology (OCT), 3C and high-throughput sequencing and hence enables researchers to interrogate interactions at hundreds of selected loci at high resolution in a single assay. Therefore this method can provide mechanistic evidence linking genetic variants to genes.

This approach has been used to examine the HHIP locus. The GWAS signal at this locus was primarily located in the 5' region of the *HHIP* gene. This led to the hypothesis that the mechanism underlying the association was at least in part due to alterations in regulatory mechanisms. Zhou et al. formally tested this hypothesis and using a combination of chromosome conformation capture, ChiP-qPCR and reporter based assays they identified a long range enhancer in the *HHIP* gene in the same region as the sentinel SNP associated with lung function (X. Zhou et al., 2012). The authors went on to further demonstrate that the COPD risk haplotype was associated with reduced reporter activity suggesting a causative mechanism leading to reduced *HHIP* expression as observed in lung tissue isolated from COPD patients (Figure 15).

Figure 15: Identification of a long-range enhancer for HHIP in the COPD GWAS locus on chromosome 4q31. (A) Long-range interaction between the COPD susceptibility locus and the HHIP promoter in Beas-2B (bronchial epithelial) and MRC5 (lung fibroblast) cells detected by 3C-PCR. The graph demonstrates 3C interaction frequency of the constant fragment containing the HHIP promoter (orange bar) with other target fragments (black bars). The y-axis refers to 3C-PCR products normalized to the interaction frequency of fragments from the BAC clone. Geometric means and standard errors were from duplicate PCR reactions. (B) The minimal COPD GWAS SNP region (~500 bp around two key SNPs, lightpurple column) cloned at forward orientation and reverse orientation showed enhancer activity for the HHIP promoter ('control', orange) measured by dual-luciferase in Beas-2B cells. (C) The effects of rs1542725 and rs6537296 were evaluated in the reporter assays. Single-nucleotide alterations were introduced individually or combined into minimal enhancer constructs at forward (upper panel) and reverse (lower panel) orientations. Adapted with permission from (Zhou et al. 2012)



Studying protein coding candidate genes

With an established candidate protein-coding gene, the traditional assays used for the characterization of gene function are gene knockdown using small interfering RNA (siRNA) or short hairpin RNA (shRNA) in a range of relevant human cell lines and/or primary cells. These methods are particularly useful when little is known about the gene function and may be used for hypothesis generation. For example, Portelli et al. have overexpressed the asthma associated gene, urokinase plasminogen activator receptor (*uPAR*) gene (encoding PLAUR protein) in human bronchial epithelial cells and observed increased proliferation as a result of this manipulation which was suggested to contribute to airway remodelling. (Portelli et al., 2014). This fits well with the observation of elevated levels of PLAUR in the airway epithelium in asthma patients (Stewart et al., 2012) and association of PLAUR levels with worsening prognosis and increased disease aggressiveness in other diseases such as cancer and COPD (Ivancso et al., 2013; Smith & Marshall, 2010). However it is difficult to infer completely the functional role of the gene with an overexpression approach and gene depletion is potentially more informative from this perspective. There are commercially available siRNAs for many of the genes jmplicated from GWAS approaches which can be used in cell biology experiments, although appropriate controls are essential as off target effects of transfection are frequent.

The in vitro suppression of a specific gene using shRNA followed by global gene analyses in human cell lines has provided a novel insight into the role of lung function and COPD associated gene HHIP in possible pathway-analysis-inferred bronchial epithelial function (J. J. Zhou et al., 2013). In this study, HHIP was targeted by shRNA in BEAS-2B airway epithelial cells followed by expression microarray analyses identifying 296 differentially expressed genes. Subsequent pathway analyses identified a particular enrichment for extracellular matrix proteins and genes associated with cell growth providing a potential insight into how HHIP may be involved in lung homeostasis. Importantly, a subset of genes were validated using additional qPCR in both BEAS-2B and primary human airway epithelial cells and shown to be differentially expressed in COPD patient lung samples versus non-disease controls (J. J. Zhou et al., 2013).

Generating novel functional hypotheses through expression profiling and pathway analyses

Although algorithms taking significant GWAS variants as input and pathways likely to be affected in the phenotype of interest as output have been developed, these algorithms have a relatively limited success in generating hypotheses about the biological basis behind considered phenotypes (K. Wang, Li, & Hakonarson, 2010). This is largely due to the complex nature of the human genome where various epistatic events between alleles are likely to occur. Also, in numerous cases, GWAS signals lie within a large region containing no annotated genes. In these situations it is often the case that a genetic variant is within an enhancer element and has an effect on the expression of a gene distant to its own location. Therefore without the understanding of the inter-genomic interactions it is hard to infer what pathways may be dysregulated in disease state.

With candidate genes prioritized it is possible to generate novel hypotheses by combining the manipulation of the expression of the gene of interest with global transcript expression profiling. For that purpose, RNAseq has advantages that out way microarray based approaches, some of which were outlined in the *in silico* section above including greater dynamic range for quantification and the investigation of splice variation.

Having performed a differential gene expression analysis in the presence and absence of approaches to target the gene of interest, e.g. RNAi, pathway analysis may then be applied. In the classical pathway analysis approach called over-representation analysis (ORA) the first step requires a creation of an input list of genes that are differentially expressed under the considered experimental condition. This list of genes is based on an arbitrary chosen statistic of significance such false discovery rate below (FDR) 5%. Then, input genes that are part of the pathway are counted. This process is repeated using appropriate background of genes (such as all protein-coding genes). Lastly, every pathway is tested for over representation in the list of input genes using hypergeometric, chi-square or binomial distribution (Huang da, Sherman, & Lempicki, 2009). The same principles apply for Gene Ontology analyses but instead of counting the number of genes per pathway, genes

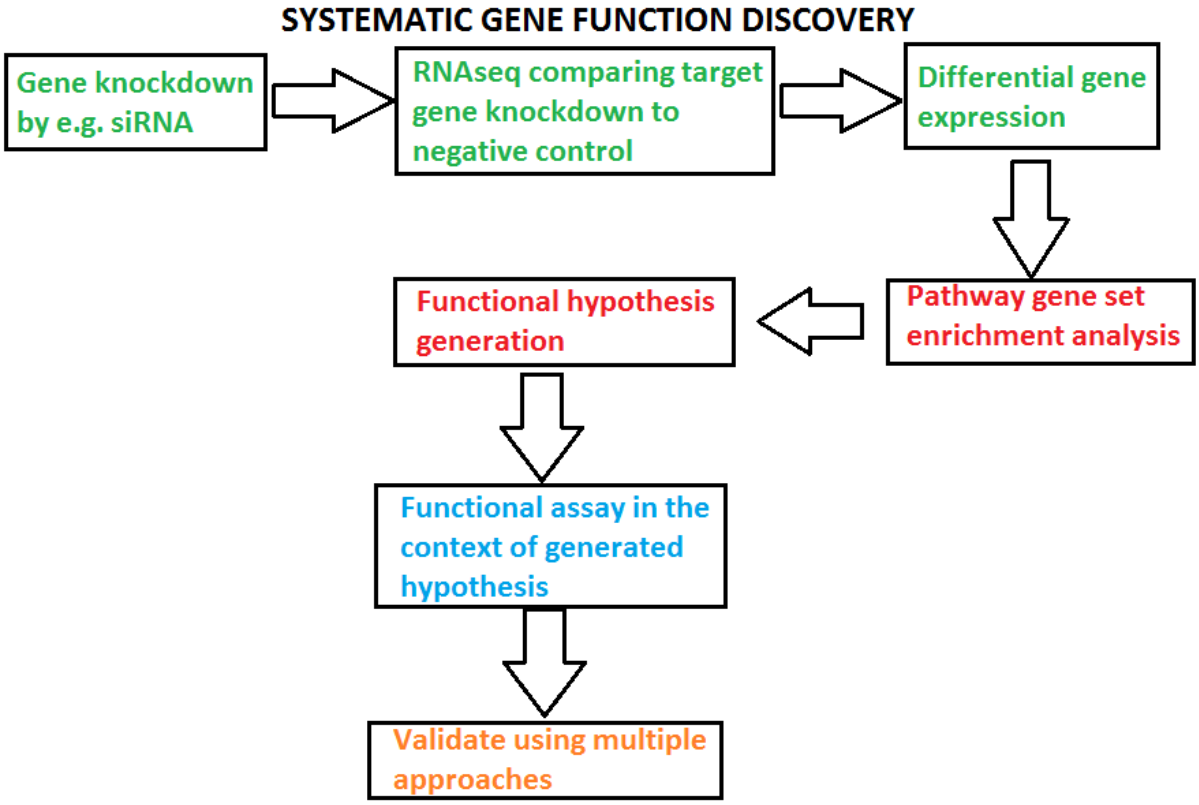
are counted for cellular process terms. However, the ORA approach is limited in its ability to identify biologically-meaningful pathways that vary between experimental conditions or phenotypes. Firstly, in ORA genes that are differentially expressed at FDR above the statistical threshold of significance, are not included in the analysis, and hence this method could miss biologically important genes that do not fulfil the criteria of arbitrary decided statistical significance (for example genes that are differentially expressed at a *P*-value <0.051). Secondly, over-represented pathways are identified based on gene-counts alone and the analysis does not account for quantitative gene expression changes.

These limitations are addressed by gene set enrichment analysis (GSEA). In contrast to ORA, GSEA approach uses all available information regarding gene expression and computes an enrichment score for the gene sets based on effect size or other ranking statistics. The objective of GSEA is to, given *a priori* defined gene set as well as gene-expression-ranked gene list, determine whether members of gene sets are randomly distributed throughout ranked lists, or primarily found at the top or bottom of the ranked list (Subramanian et al., 2005). GSEA calculates the enrichment score by walking down the ranked list of genes, increasing a running-sum statistic when a gene is in the gene set and decreasing it when it is not. A commonly employed ranking algorithm is a signal-to-noise ranking where mean gene expression in one condition is subtracted from the mean expression in the other condition divided by the sum of variances. Therefore genes are ranked from most upregulated with least conditional variability through genes with moderate changes in gene expression at greatest expression variability to most downregulated genes with least variability. Hence, in enrichment score calculation, the magnitude of the increment depends on the strength of differential gene expression and its biological variation. The limitation of GSEA is the assumption that genes and pathways are independent from each other, which is not necessarily true considering the complexity of cellular networks. Also, because both GSEA and ORA are based on *a priori* defined gene sets, both these approaches are limited by the quality of gene set definition. For example, if more than 50% of genes assigned to a particular gene set are erroneous then the identified dysregulated pathway will be flawed as well. Therefore it is advisable to use the most up-to-date pathway or gene ontology definitions that are community curated and adjusted with each new scientific

publication. Finally, it is recommended to validate each result with another analysis, for instance by testing if the identified dysregulated pathways with one gene depletion method agrees with another gene depletion method.

The analyses mentioned above can help in the prioritization of functional *in vitro* assays that may be performed following the experimental gene expression manipulation. This approach is undoubtedly superior to choosing functional read-out assays on an arbitrary basis. With evidence of pathway or gene ontology term dysregulation, a functional assay related to the discovered dysregulation can be performed. For example, with evidence of dysregulation in cellular proliferation genes it is worth testing for cell proliferation with one of the available DNA replication assays. Having determined gene functions in *in vitro* models, researchers can further hypothesise about possible relationships between phenotype and the identified perturbed pathways (Figure 16)

Figure 16: A workflow of systematic gene function discovery through combination of transcriptomics, pathway analysis, hypothesis generation and final biological validation.



The promise of genome editing tools

It has been suggested that the emergence of genome editing is a ‘game changer’ in scientists’ attempt to meaningfully translate genetic association findings (Sander & Joung, 2014). These methods allow editing any genomic sequence by inserting, excluding or modifying sequences in any mammalian cell type or even embryological cells to study the effect on model organisms. Importantly, genome editing allows simultaneous disruption of a multitude of genes or regulatory elements at once, thus allowing the investigation of allele interactions or synergistic effects. This is a huge step forward considering the difficulties of achieving this with traditional RNAi-based approaches, as well as the polygenic character of the majority of phenotypes. Also, it is possible to use these technologies to study the effects of genetic disruptions on lineage-specific cellular differentiation. For that purpose, using totipotent or pluripotent stem cells (or iPSCs generated via epigenetic reprogramming of mature cells) shows great potential promise. It was recently shown that genome editing can be used not only to knockout genes but also to induce their expression from endogenous promoters (Konermann et al., 2015) or for completely other purposes such as modifying epigenetic marks (Gilbert et al., 2013; Maeder et al., 2013; Mali et al., 2013). As mentioned, this can be achieved over multiple loci and has the advantage of recapitulating the transcription at the endogenous genomic template in opposition to recombinant overexpression constructs which may not be representative of the endogenous situation. The two most popular genome editing techniques are Transcription activator-like effector nucleases, abbreviated as TALENs (J. C. Miller et al., 2011), and clustered regularly interspaced short palindromic repeats (CRISPR) in association with RNA-guided Cas9 nuclease (CRISPR-Cas9 system) (Sander & Joung, 2014).

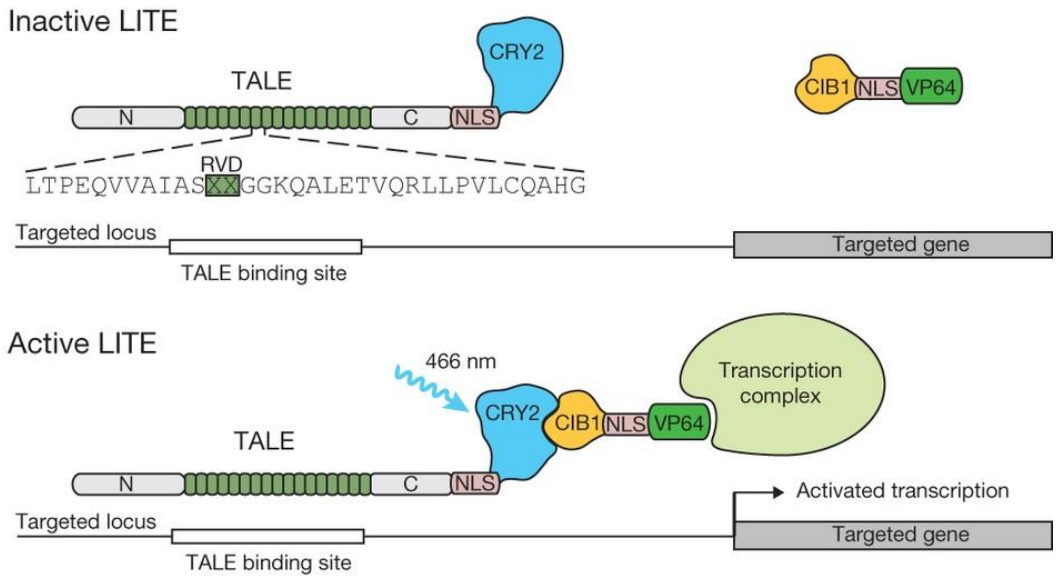
TALENs are composed of a nuclease domain fused to a DNA-binding domain. The nuclease cleaves the genomic DNA in a non-specific manner but the DNA-binding domain confers the needed specificity. This domain is engineer-able to recognize specific DNA sequences and essentially has properties similar to transcription factors capable of activating gene expression (hence the name transcription activator-like effectors molecule). The engineered nuclease binds and causes a double-strand break to DNA. Then non-homologous end-joining or homology-directed repair are activated, thus allowing editing of target sites (Joung & Sander,

2013). CRISPR-Cas9 systems are alternative to TALENs and have improved cleavage efficiency and eased the implementation at reduced cost. In contrast to TALENs, Cas9 nuclease is guided to a target site by a RNA molecule. Therefore, in this case, there is no need to design custom proteins for DNA binding. Konermann et al. leveraged CRISPR-Cas9 system to induce the expression of multitude of genes and this is possible because the entire complex can be provided with distinct effector domains such as activator domains, repressor domains or domains altering the epigenetic mark (Gilbert et al., 2013; Konermann et al., 2015; Maeder et al., 2013; Mali et al., 2013). In those circumstances the Cas9 endonuclease is catalytically inactivated (dCas9). These modified CRISPR-Cas9 constructs can be effectively used to control the activity of candidate regulatory elements or genes that contain significant GWAS signal variants. Introduction into somatic cells can be achieved with typical transfection while introduction into model organisms can be accomplished with injection into the model organism zygote. As with traditional RNAi-based approaches genome editing can occur with off target effects and the current challenges are to minimize these to provide more robust interpretation. For excellent reviews of TALENs and CRISPR/Cas9 approaches see (Gaj, Gersbach, & Barbas, 2013).

A new avenue in genome editing technology has recently emerged with light-inducible transcriptional effectors (LITEs) (Konermann et al., 2013). LITE modules consist of the light-sensitive photoreceptor cryptochrome 2 (CRY2) that is fused to TALEN DNA-binding domain, however, theoretically the concept can be applied to the CRISPR-Cas9 system as well (Figure 16). Authors have combined TALEN domain, light-sensitive cryptochrome 2 protein CIB1 and its co-partners obtained from *Arabidobsis* in order to induce gene expression by exposure to the light at sites determined by specificity of DNA-domain binding. Variable levels of increases in mRNA expression were observed (from little to large) and this was accompanied by an increase in protein level. The construct allows for reversible modulation of gene transcription and epigenetic marks in spatially and temporally sensitive manners via the exposure to light (Konermann et al., 2013). This study is essentially a proof of concept that overexpression using this unique technique may be possible and it was shown to be applicable both *in vitro* and *in vivo*. Undoubtedly this new technique offers great opportunity for biologists studying gene regulation and gene function in

their genetic translation efforts, however further studies are required to determine the specificity of the method and whether it can be used on a routine basis.

Figure 17: The mechanism of action of light-inducible transcriptional effectors: Light-sensitive CRY2 is anchored to DNA-binding TALEN and upon exposure to light recruits CIBI-copartners complex which in turn brings transcriptional machinery resulting in expression of nearby gene. Adapted with permission from (Konermann et al., 2013)



Exemplary case study: The *FTO* locus and obesity

While respiratory genetics translation continues, an exemplary study utilizing many of the approaches outlined in this review is the functional characterization of the fat mass and obesity-associated (*FTO*) locus previously associated with obesity (Claussnitzer et al., 2015). In this study, Claussnitzer and colleagues initially start with a large genomic region on chromosome 16 encompassing the *FTO* gene showing a broad association signal. The initial approach uses a chromatin annotation map of the region (from 127 reference epigenomes) which identified an unusually long enhancer (12.8 kb) in mesenchymal adipocyte progenitors. This was followed by a 10kb-reporter tiling approach encompassing the risk and non-risk haplotypes to identify the key genomic region driving haplotype specific enhancer activity in the relevant cell type, human SGBS adipocytes (i.e., adipocytes derived from a patient with the Simpson–Golabi–Behmel syndrome), which indicated genetic control of enhancer activity (Claussnitzer et al., 2015). Chromosome conformation capture was then used to identify long range chromatin interactions which included 8

genes and importantly developmental regulators *IRX3* and *IRX5* had genotype-associated expression. *IRX3* and *IRX5* were confirmed as being important in obesity associated cellular profiles by examining Human Adipocytes (carriers and non-carriers of the risk haplotype) and by siRNA knockdown. To identify the casual variant underlying the risk haplotype promoter-reporter and EMSA approaches were used which identified the rs1421085 risk allele within an *ARID5B* binding site. Subsequent, *ARID5B* knockdown and over-expression modulated *IRX3* and *IRX5* expression providing further support for the putative mechanism. Finally, using CRISPR-Cas9 genome editing using both risk and non-risk haplotype background to introduce the alternative rs1421085 confirmed the effects on *IRX3* and *IRX5* expression and importantly determined a developmental shift from browning to whitening programs and loss of mitochondrial thermogenesis (Claussnitzer et al., 2015).

>can we put the section below in a Text box figure please?

Inferred biology of selected and most reproducible lung function genes

In the GWAS for lung function and COPD performed to date, a number of genes are reproducibly appearing to be associated with these traits and importantly using several of the approaches outlined above look to be the causative genes. Work is still being undertaken to both understand what the fundamental biological functions of these putative genes are and to elucidate underlying biological mechanisms leading to the observed associations. Nevertheless it is worth considering the current paradigm about the role of some of these genes in altered biology as it may facilitate hypothesis generation.

AGER encodes RAGE which mediates interactions of advanced glycosylation end products and these are glycosylated proteins which accumulate in vascular tissue during aging and at an accelerated rate in several human disorders including diabetes (Kankova et al., 2001). RAGE acts as a receptor for amyloid beta peptide and contributes to the translocation of amyloid-beta peptide across the cell membrane from the extracellular to the intracellular space in cortical neurons (Yan et al., 1996). RAGE signalling plays a critical role in regulating the production and/or expression of TNF-α, oxidative stress, and endothelial dysfunction which is of relevance to damage in the airways. While predominantly from mouse studies (see

in vivo section) the accumulating data suggests that RAGE is critical for both normal lung development and the response of the lung to damage.

HTR4 is a member of the family of serotonin receptors, which are G protein coupled receptors that stimulate cAMP production in response to serotonin. Therefore *HTR4* can play a role in neurotransmission and interestingly this gene was also associated with smoking behaviours. The activity of this receptor is mediated by G proteins that stimulate adenylate cyclase and may have a role in lung development (Hodge et al., 2013). At this time it is likely HTR4 is related to lung function due to the role in neurotransmission.

HHIP is a member of family of hedgehog-interacting proteins. These proteins are conserved morphogens playing a critical role in development. More specifically these proteins take part in formation of anteroposterior patterns of limbs as well as regulation of left-right asymmetry in the embryo (Ingham & McMahon, 2001).

GSTCD and *INTS12* are the two oppositely transcribed genes at 4q24 locus at the centre of association signal for lung function (Figure 2). Nothing is known about the function of the former gene. On the other hand, *INTS12* encodes a protein that was initially discovered as the smallest member of the Integrator Complex that is involved in small nuclear RNA (snRNA) processing (Baillat et al., 2005). The role of Integrator Complex in snRNA processing was demonstrated in human cell lines (Baillat et al., 2005) but the specific requirement of *INTS12* in snRNA processing has only shown in *Drosophila melanogaster* (Ezzeddine et al., 2011) and nothing is currently known about the functional requirement of *INTS12* in snRNA processing in human cells. Interestingly, a recent study by Chen et al. demonstrated that in the fly, evolutionary conserved plant homeodomain (PHD) motif protein domain of INTS12 is dispensable for snRNA processing while N-terminal domain is both necessary and sufficient for this processing to occur (J. Chen, Waltenspiel, Warren, & Wagner, 2013). This strongly suggested the existence of important and unrealized functions for this gene. Although homozygote *INTS12* knockout mouse models are lethal but there is no overt phenotype observed for equivalent manipulation of *GSTCD* (Obeidat et al., 2013). Crucially, Obeidat et al. reported that multiple SNPs associated with lung function at 4q24 are also in eQTL with *INTS12* expression but this was not observed for other genes in the locus (Obeidat et al., 2013). *NPNT* is also present within the

associated 4q24 locus. This gene plays a role in the kidney development by acting as a functional ligand of integrin α-8/β-1. Together with α-8/β-1, *NPNT*'s protein regulates the expression of GDNF which is essential for the kidney development (Linton, Martin, & Reichardt, 2007), however there is a clear role for integrins in the airways including links to airway fibrosis. It has also been suggested that *NPNT* promotes osteoblast differentiation via the epidermal growth factor-like repeats (Kahai, Lee, Seth, & Yang, 2010).

Summary and future perspectives

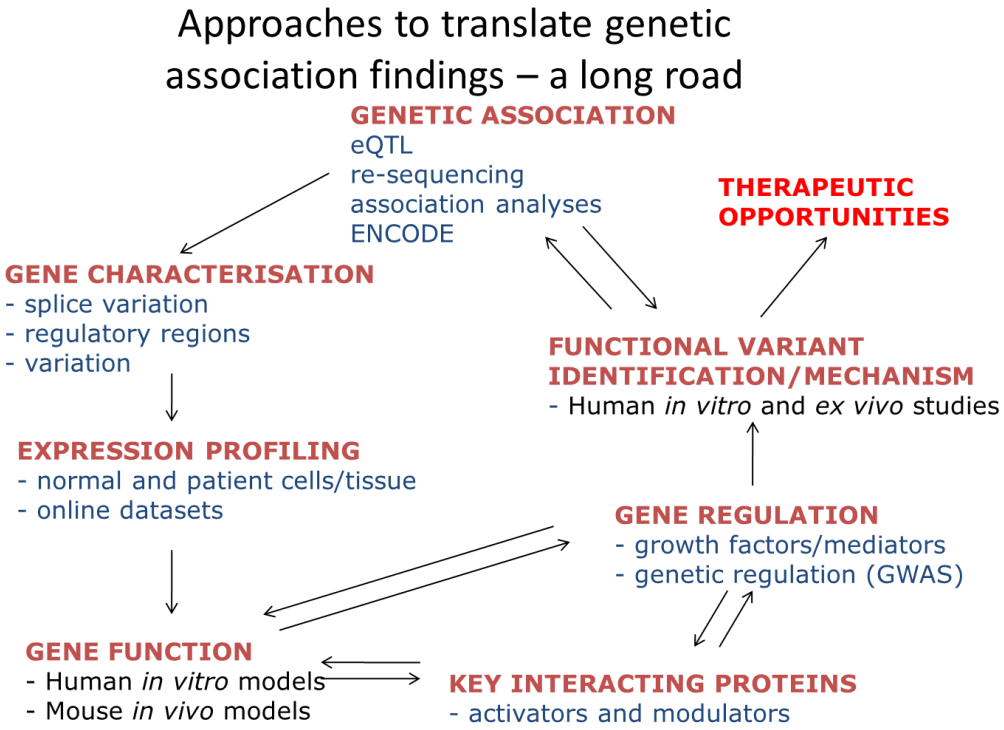
As GWAS and particularly GWAS meta-analyses involve increasingly larger population sizes and improved integration of the genome (including rare sequence variation) continue to identify novel loci for a large number of human traits there is a pressing need to develop technologies to translate these findings. This functional understanding is critical to move from genetics to translational medicine identifying potentially novel targets for therapeutic intervention. This is particularly important in diseases such as COPD where the current medicines available provide relief from symptoms but do not address the underlying progression of the disease.

This translation has been significantly facilitated by recent developments in the areas outlined in this review (Figure 18) but particularly in functional annotation of the genome, mapping chromatin interactions, cell and tissue eQTLs, transgenic mice and more recently genome editing approaches. While all of these approaches have a role to play, it is clear careful experimental design using the most appropriate (ideally human) system that is critical to interpretation and so we envisage a focussing of approaches to include primary adult human cells, or excitingly, the potential for human iPSCs to contribute to these approaches. As genome editing technologies, particularly using CRISPR/Cas9 become routine, efficient and scalable, we envisage this methodology to play a pivotal role in the investigation of both gene and single variant in biological systems both *in vivo* and *in vitro*.

While we have focussed to genetic association in the context of GWAS it is important to note that epigenome-wide association studies (EWAS) with traits and human diseases are increasing and there is an opportunity to combine GWAS and EWAS findings particularly for scalable epigenetic changes such as DNA methylation. This

approach may provide greater insight and initial starting points for functional translation.

Figure 19: An overview of tools that can be used for translation of genetic association findings into biological function allowing for ultimate therapeutic intervention.



Competing interests

Authors declare no competing interests.

Acknowledgements

The authors' laboratory is funded by grants from Medical Research Council (G10000861) and Asthma UK (AUK-PG-2013-188).

References

(GOLD), G. I. f. C. O. L. D. (2015). Global Strategy for the Diagnosis, Management and Prevention of COPD. Retrieved from: http://www.goldcopd.org/

Adli, M., & Bernstein, B. E. (2011). Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. *Nat Protoc, 6*(10), 1656-1668. doi: 10.1038/nprot.2011.402

Afonso, A. S., Verhamme, K. M., Sturkenboom, M. C., & Brusselle, G. G. (2011). COPD in the general population: prevalence, incidence and survival. *Respir Med, 105*(12), 1872-1884. doi: 10.1016/j.rmed.2011.06.012

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., . . . Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput Biol, 9*(11), e1003326. doi: 10.1371/journal.pcbi.1003326

Baillat, D., Hakimi, M. A., Naar, A. M., Shilatifard, A., Cooch, N., & Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell, 123*(2), 265-276. doi: 10.1016/j.cell.2005.08.019

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res, 21*(3), 381-395. doi: 10.1038/cr.2011.22

Barton, S. J., Koppelman, G. H., Vonk, J. M., Browning, C. A., Nolte, I. M., Stewart, C. E., . . . Sayers, I. (2009). PLAUR polymorphisms are associated with asthma, PLAUR levels, and lung function decline. *J Allergy Clin Immunol, 123*(6), 1391-1400 e1317. doi: 10.1016/j.jaci.2009.03.014

Bellusci, S., Furuta, Y., Rush, M. G., Henderson, R., Winnier, G., & Hogan, B. L. (1997). Involvement of Sonic hedgehog (Shh) in mouse embryonic lung growth and morphogenesis. *Development, 124*(1), 53-63.

Bentley, R. W., Pearson, J., Gearry, R. B., Barclay, M. L., McKinney, C., Merriman, T. R., & Roberts, R. L. (2010). Association of higher DEFB4 genomic copy number with Crohn's disease. *Am J Gastroenterol, 105*(2), 354-359. doi: 10.1038/ajg.2009.582

Bertin, G., Poujeol, C., Rubera, I., Poujeol, P., & Tauc, M. (2005). In vivo Cre/loxP mediated recombination in mouse Clara cells. *Transgenic Res, 14*(5), 645-654. doi: 10.1007/s11248-005-7214-0

Bierhaus, A., & Nawroth, P. P. (2009). Multiple levels of regulation determine the role of the receptor for AGE (RAGE) as common soil in inflammation, immune responses and diabetes mellitus and its complications. *Diabetologia, 52*(11), 2251-2263. doi: 10.1007/s00125-009-1458-9

Bocchini, V., Mazzolla, R., Barluzzi, R., Blasi, E., Sick, P., & Kettenmann, H. (1992). An immortalized cell line expresses properties of activated microglial cells. *J Neurosci Res, 31*(4), 616-621. doi: 10.1002/jnr.490310405

Brehm, J. M., Hagiwara, K., Tesfaigzi, Y., Bruse, S., Mariani, T. J., Bhattacharya, S., . . . Celedon, J. C. (2011). Identification of FGF7 as a novel susceptibility locus for chronic obstructive pulmonary disease. *Thorax, 66*(12), 1085-1090. doi: 10.1136/thoraxjnl-2011-200017

Brown, S. D., & Moore, M. W. (2012). Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium. *Dis Model Mech, 5*(3), 289-292. doi: 10.1242/dmm.009878

Calarco, J. A., Xing, Y., Caceres, M., Calarco, J. P., Xiao, X., Pan, Q., . . . Blencowe, B. J. (2007). Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev, 21*(22), 2963-2975. doi: 10.1101/gad.1606907

Castaldi, P. J., Cho, M. H., Litonjua, A. A., Bakke, P., Gulsvik, A., Lomas, D. A., . . . Eclipse, I. (2011). The association of genome-wide significant spirometric loci with chronic obstructive pulmonary disease susceptibility. *Am J Respir Cell Mol Biol, 45*(6), 1147-1153. doi: 10.1165/rcmb.2011-0055OC

Castaldi, P. J., Cho, M. H., San Jose Estepar, R., McDonald, M. L., Laird, N., Beaty, T. H., . . . Investigators, C. O. (2014). Genome-wide association identifies regulatory Loci associated with distinct local histogram emphysema patterns. *Am J Respir Crit Care Med, 190*(4), 399-409. doi: 10.1164/rccm.201403-0569OC

Celedon, J. C., Lange, C., Raby, B. A., Litonjua, A. A., Palmer, L. J., DeMeo, D. L., . . . Silverman, E. K. (2004). The transforming growth factor-beta1 (TGFB1) gene is associated with chronic obstructive pulmonary disease (COPD). *Hum Mol Genet, 13*(15), 1649-1656. doi: 10.1093/hmg/ddh171

Chappell, S., Daly, L., Morgan, K., Baranes, T. G., Roca, J., Rabinovich, R., . . . Kalsheker, N. (2006). The SERPINE2 gene and chronic obstructive pulmonary disease. *Am J Hum Genet, 79*(1), 184-186; author reply 186-187. doi: 10.1086/505268

Chen, J., Waltenspiel, B., Warren, W. D., & Wagner, E. J. (2013). Functional analysis of the integrator subunit 12 identifies a microdomain that mediates activation of the Drosophila integrator complex. *J Biol Chem, 288*(7), 4867-4877. doi: 10.1074/jbc.M112.425892

Chen, W., Brehm, J. M., Manichaikul, A., Cho, M. H., Boutaoui, N., Yan, Q., . . . Celedon, J. C. (2015). A genome-wide association study of chronic obstructive pulmonary disease in Hispanics. *Ann Am Thorac Soc, 12*(3), 340-348. doi: 10.1513/AnnalsATS.201408-380OC

Cheng, D. T., Kim, D. K., Cockayne, D. A., Belousov, A., Bitter, H., Cho, M. H., . . . Investigators, E. (2013). Systemic soluble receptor for advanced glycation endproducts is a biomarker of emphysema and associated with AGER genetic variants in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med, 188*(8), 948-957. doi: 10.1164/rccm.201302-0247OC

Cho, M. H., Boutaoui, N., Klanderman, B. J., Sylvia, J. S., Ziniti, J. P., Hersh, C. P., . . . Silverman, E. K. (2010). Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet, 42*(3), 200-202. doi: 10.1038/ng.535

Cho, M. H., Castaldi, P. J., Hersh, C. P., Hobbs, B. D., Barr, R. G., Tal-Singer, R., . . . Investigators, C. O. (2015). A Genome-wide Association Study of Emphysema and Airway Quantitative Imaging Phenotypes. *Am J Respir Crit Care Med*. doi: 10.1164/rccm.201501-0148OC

Cho, M. H., Castaldi, P. J., Wan, E. S., Siedlinski, M., Hersh, C. P., Demeo, D. L., . . . Investigators, C. O. (2012). A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet, 21*(4), 947-957. doi: 10.1093/hmg/ddr524

Cho, M. H., McDonald, M. L., Zhou, X., Mattheisen, M., Castaldi, P. J., Hersh, C. P., . . . Investigators, C. O. (2014). Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med, 2*(3), 214-225. doi: 10.1016/S2213-2600(14)70002-5

Ciprandi, G., Capasso, M., Tosca, M., Salpietro, C., Salpietro, A., Marseglia, G., & La Rosa, M. (2012). A forced expiratory flow at 25-75% value <65% of predicted should be considered abnormal: a real-world, cross-sectional study. *Allergy Asthma Proc, 33*(1), e5-8. doi: 10.2500/aap.2012.33.3524

Claussnitzer, M., Dankel, S. N., Kim, K. H., Quon, G., Meuleman, W., Haugen, C., . . . Kellis, M. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med, 373*(10), 895-907. doi: 10.1056/NEJMoa1502214

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., . . . Hurles, M. E. (2010). Origins and functional impact of copy number variation in the human genome. *Nature, 464*(7289), 704-712. doi: 10.1038/nature08516

Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science, 306*(5696), 636-640. doi: 10.1126/science.1105136

Consortium, E. P., Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., . . . de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature, 447*(7146), 799-816. doi: 10.1038/nature05874

Coultas, D. B., Hanis, C. L., Howard, C. A., Skipper, B. J., & Samet, J. M. (1991). Heritability of ventilatory function in smoking and nonsmoking New Mexico Hispanics. *Am Rev Respir Dis, 144*(4), 770-775. doi: 10.1164/ajrccm/144.4.770

Dawkins, P. A., & Stockley, R. A. (2001). Animal models of chronic obstructive pulmonary disease. *Thorax, 56*(12), 972-977.

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev, 25*(10), 1010-1022. doi: 10.1101/gad.2037511

Demeo, D. L., Mariani, T. J., Lange, C., Srisuma, S., Litonjua, A. A., Celedon, J. C., . . . Silverman, E. K. (2006). The SERPINE2 gene is associated with chronic obstructive pulmonary disease. *Am J Hum Genet, 78*(2), 253-264. doi: 10.1086/499828

DeMeo, D. L., & Silverman, E. K. (2003). Genetics of chronic obstructive pulmonary disease. *Semin Respir Crit Care Med, 24*(2), 151-160. doi: 10.1055/s-2003-39014

DeMeo, D. L., & Silverman, E. K. (2004). Alpha1-antitrypsin deficiency. 2: genetic aspects of alpha(1)-antitrypsin deficiency: phenotypes and genetic modifiers of emphysema risk. *Thorax, 59*(3), 259-264.

Diez-Villanueva, A., Mallona, I., & Peinado, M. A. (2015). Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics Chromatin, 8*, 22. doi: 10.1186/s13072-015-0014-8

Dijkstra, A. E., Boezen, H. M., van den Berge, M., Vonk, J. M., Hiemstra, P. S., Barr, R. G., . . . LifeLines Cohort Study, g. (2015). Dissecting the genetics of chronic mucus hypersecretion in smokers with and without COPD. *Eur Respir J, 45*(1), 60-75. doi: 10.1183/09031936.00093314

Dijkstra, A. E., Postma, D. S., van Ginneken, B., Wielputz, M. O., Schmidt, M., Becker, N., . . . Groen, H. J. (2015). Novel genes for airway wall thickness identified with combined genome-wide association and expression analyses. *Am J Respir Crit Care Med, 191*(5), 547-556. doi: 10.1164/rccm.201405-0840OC

Edwards, J., Belvisi, M., Dahlen, S. E., Holgate, S., & Holmes, A. (2015). Human tissue models for a human disease: what are the barriers? *Thorax, 70*(7), 695-697. doi: 10.1136/thoraxjnl-2014-206648

Eriksson, S. (1965). Studies in alpha 1-antitrypsin deficiency. *Acta Med Scand Suppl, 432*, 1-85.

Ezzeddine, N., Chen, J., Waltenspiel, B., Burch, B., Albrecht, T., Zhuo, M., . . . Wagner, E. J. (2011). A subset of Drosophila integrator proteins is essential for efficient U7 snRNA and spliceosomal snRNA 3'-end formation. *Mol Cell Biol, 31*(2), 328-341. doi: 10.1128/MCB.00943-10

Favorov, A., Mularoni, L., Cope, L. M., Medvedeva, Y., Mironov, A. A., Makeev, V. J., & Wheelan, S. J. (2012). Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput Biol, 8*(5), e1002529. doi: 10.1371/journal.pcbi.1002529

Fellermann, K., Stange, D. E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C. L., . . . Stange, E. F. (2006). A chromosome 8 gene-cluster polymorphism

with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet, 79*(3), 439-448. doi: 10.1086/505915

Ferhani, N., Letuve, S., Kozhich, A., Thibaudeau, O., Grandsaigne, M., Maret, M., . . . Pretolani, M. (2010). Expression of high-mobility group box 1 and of receptor for advanced glycation end products in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med, 181*(9), 917-927. doi: 10.1164/rccm.200903-0340OC

Fode, P., Jespersgaard, C., Hardwick, R. J., Bogle, H., Theisen, M., Dodoo, D., . . . Hollox, E. J. (2011). Determination of beta-defensin genomic copy number in different populations: a comparison of three methods. *PLoS One, 6*(2), e16768. doi: 10.1371/journal.pone.0016768

Gaens, K. H., Ferreira, I., van der Kallen, C. J., van Greevenbroek, M. M., Blaak, E. E., Feskens, E. J., . . . Schalkwijk, C. G. (2009). Association of polymorphism in the receptor for advanced glycation end products (RAGE) gene with circulating RAGE levels. *J Clin Endocrinol Metab, 94*(12), 5174-5180. doi: 10.1210/jc.2009-1067

Gaj, T., Gersbach, C. A., & Barbas, C. F., 3rd. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol, 31*(7), 397-405. doi: 10.1016/j.tibtech.2013.04.004

Genomes Project, C., Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., . . . McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi: 10.1038/nature11632

Gibney, E. R., & Nolan, C. M. (2010). Epigenetics and gene expression. *Heredity (Edinb), 105*(1), 4-13. doi: 10.1038/hdy.2010.54

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., . . . Qi, L. S. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell, 154*(2), 442-451. doi: 10.1016/j.cell.2013.06.044

Gunther, E. J., Belka, G. K., Wertheim, G. B., Wang, J., Hartman, J. L., Boxer, R. B., & Chodosh, L. A. (2002). A novel doxycycline-inducible system for the transgenic analysis of mammary gland biology. *FASEB J, 16*(3), 283-292. doi: 10.1096/fj.01-0551com

Hackett, T. L., Holloway, R., Holgate, S. T., & Warner, J. A. (2008). Dynamics of pro-inflammatory and anti-inflammatory cytokine release during acute inflammation in chronic obstructive pulmonary disease: an ex vivo study. *Respir Res, 9*, 47. doi: 10.1186/1465-9921-9-47

Hallberg, J., Dominicus, A., Eriksson, U. K., Gerhardsson de Verdier, M., Pedersen, N. L., Dahlback, M., . . . Svartengren, M. (2008). Interaction between smoking and genetic factors in the development of chronic bronchitis. *Am J Respir Crit Care Med, 177*(5), 486-490. doi: 10.1164/rccm.200704-565OC

Hancock, D. B., Artigas, M. S., Gharib, S. A., Henry, A., Manichaikul, A., Ramasamy, A., . . . London, S. J. (2012). Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS Genet, 8*(12), e1003098. doi: 10.1371/journal.pgen.1003098

Hancock, D. B., Eijgelsheim, M., Wilk, J. B., Gharib, S. A., Loehr, L. R., Marciante, K. D., . . . London, S. J. (2010). Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet, 42*(1), 45-52. doi: 10.1038/ng.500

Hankinson, J. L., Odencrantz, J. R., & Fedan, K. B. (1999). Spirometric reference values from a sample of the general U.S. population. *Am J Respir Crit Care Med, 159*(1), 179-187. doi: 10.1164/ajrccm.159.1.9712108

Hansel, N. N., Ruczinski, I., Rafaels, N., Sin, D. D., Daley, D., Malinina, A., . . . Barnes, K. C. (2013). Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. *Hum Genet, 132*(1), 79-90. doi: 10.1007/s00439-012-1219-6

Hansen, J. E., Sun, X. G., & Wasserman, K. (2007). Spirometric criteria for airway obstruction: Use percentage of FEV1/FVC ratio below the fifth percentile, not < 70%. *Chest, 131*(2), 349-355. doi: 10.1378/chest.06-1349

Hao, K., Bosse, Y., Nickle, D. C., Pare, P. D., Postma, D. S., Laviolette, M., . . . Sin, D. D. (2012). Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet, 8*(11), e1003029. doi: 10.1371/journal.pgen.1003029

Hardin M, S. E. (2014). Chronic obstructive pulmonary disease genetics: a review of the past and a look into the future. *J COPD F, 1*(1), 33-46

Hardwick, R. J., Machado, L. R., Zuccherato, L. W., Antolinos, S., Xue, Y., Shawa, N., . . . Hollox, E. J. (2011). A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia. *Hum Mutat, 32*(7), 743-750. doi: 10.1002/humu.21491

Heger, A., Webber, C., Goodson, M., Ponting, C. P., & Lunter, G. (2013). GAT: a simulation framework for testing the association of genomic intervals. *Bioinformatics, 29*(16), 2046-2048. doi: 10.1093/bioinformatics/btt343

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., . . . Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell, 38*(4), 576-589. doi: 10.1016/j.molcel.2010.05.004

Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc, 2*(8), 1849-1861. doi: 10.1038/nprot.2007.249

Hemani, G., Shakhbazov, K., Westra, H. J., Esko, T., Henders, A. K., McRae, A. F., . . . Powell, J. E. (2014). Detection and replication of epistasis influencing transcription in humans. *Nature, 508*(7495), 249-253. doi: 10.1038/nature13005

Hemminki, K., Li, X., Sundquist, K., & Sundquist, J. (2008a). Familial risks for chronic obstructive pulmonary disease among siblings based on hospitalisations in Sweden. *J Epidemiol Community Health, 62*(5), 398-401. doi: 10.1136/jech.2007.063156

Hemminki, K., Li, X., Sundquist, K., & Sundquist, J. (2008b). Familial risks for common diseases: etiologic clues and guidance to gene identification. *Mutat Res, 658*(3), 247-258. doi: 10.1016/j.mrrev.2008.01.002

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet, 6*(2), 95-108. doi: 10.1038/nrg1521

Hodge, E., Nelson, C. P., Miller, S., Billington, C. K., Stewart, C. E., Swan, C., . . . Sayers, I. (2013). HTR4 gene structure and altered expression in the developing lung. *Respir Res, 14*, 77. doi: 10.1186/1465-9921-14-77

Hofmann, M. A., Drury, S., Hudson, B. I., Gleason, M. R., Qu, W., Lu, Y., . . . Schmidt, A. M. (2002). RAGE and arthritis: the G82S polymorphism amplifies

the inflammatory response. *Genes Immun, 3*(3), 123-135. doi: 10.1038/sj.gene.6363861

Hollox, E. J., Huffmeier, U., Zeeuwen, P. L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., . . . Schalkwijk, J. (2008). Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet, 40*(1), 23-25. doi: 10.1038/ng.2007.48

Hori, O., Brett, J., Slattery, T., Cao, R., Zhang, J., Chen, J. X., . . . et al. (1995). The receptor for advanced glycation end products (RAGE) is a cellular binding site for amphoterin. Mediation of neurite outgrowth and co-expression of rage and amphoterin in the developing nervous system. *J Biol Chem, 270*(43), 25752-25761.

House, J. S., Li, H., DeGraff, L. M., Flake, G., Zeldin, D. C., & London, S. J. (2015). Genetic variation in HTR4 and lung function: GWAS follow-up in mouse. *FASEB J, 29*(1), 323-335. doi: 10.1096/fj.14-253898

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res, 37*(1), 1-13. doi: 10.1093/nar/gkn923

Hubert, H. B., Fabsitz, R. R., Feinleib, M., & Gwinn, C. (1982). Genetic and environmental influences on pulmonary function in adult twins. *Am Rev Respir Dis, 125*(4), 409-415.

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., . . . Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat Genet, 46*(2), 205-212. doi: 10.1038/ng.2871

Hukkinen, M., Kaprio, J., Broms, U., Viljanen, A., Kotz, D., Rantanen, T., & Korhonen, T. (2011). Heritability of lung function: a twin study among never-smoking elderly women. *Twin Res Hum Genet, 14*(5), 401-407. doi: 10.1375/twin.14.5.401

Hunninghake, G. M., Cho, M. H., Tesfaigzi, Y., Soto-Quiros, M. E., Avila, L., Lasky-Su, J., . . . Celedon, J. C. (2009). MMP12, lung function, and COPD in high-risk populations. *N Engl J Med, 361*(27), 2599-2608. doi: 10.1056/NEJMoa0904006

Hurd, S., & Pauwels, R. (2002). Global Initiative for Chronic Obstructive Lung Diseases (GOLD). *Pulm Pharmacol Ther, 15*(4), 353-355.

Ingham, P. W., & McMahon, A. P. (2001). Hedgehog signaling in animal development: paradigms and principles. *Genes Dev, 15*(23), 3059-3087. doi: 10.1101/gad.938601

International HapMap, C. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi: 10.1038/nature04226

Irving, J. A., Pike, R. N., Lesk, A. M., & Whisstock, J. C. (2000). Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Res, 10*(12), 1845-1864.

Ivancso, I., Toldi, G., Bohacs, A., Eszes, N., Muller, V., Rigo, J., Jr., . . . Tamasi, L. (2013). Relationship of circulating soluble urokinase plasminogen activator receptor (suPAR) levels to disease control in asthma and asthmatic pregnancy. *PLoS One, 8*(4), e60697. doi: 10.1371/journal.pone.0060697

Janciauskiene, S. M., Bals, R., Koczulla, R., Vogelmeier, C., Kohnlein, T., & Welte, T. (2011). The discovery of alpha1-antitrypsin and its role in health and disease. *Respir Med, 105*(8), 1129-1139. doi: 10.1016/j.rmed.2011.02.002

Jin, Z., Chung, J. W., Mei, W., Strack, S., He, C., Lau, G. W., & Yang, J. (2015). Regulation of nuclear-cytoplasmic shuttling and function of Family with

sequence similarity 13, member A (Fam13a), by B56-containing PP2As and Akt. *Mol Biol Cell, 26*(6), 1160-1173. doi: 10.1091/mbc.E14-08-1276

Joung, J. K., & Sander, J. D. (2013). TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol, 14*(1), 49-55. doi: 10.1038/nrm3486

Kahai, S., Lee, S. C., Seth, A., & Yang, B. B. (2010). Nephronectin promotes osteoblast differentiation via the epidermal growth factor-like repeats. *FEBS Lett, 584*(1), 233-238. doi: 10.1016/j.febslet.2009.11.077

Kankova, K., Zahejsky, J., Marova, I., Muzik, J., Kuhrova, V., Blazkova, M., . . . Vacha, J. (2001). Polymorphisms in the RAGE gene influence susceptibility to diabetes-associated microvascular dermatoses in NIDDM. *J Diabetes Complications, 15*(4), 185-192.

Kiefer, J. C. (2007). Epigenetics in development. *Dev Dyn, 236*(4), 1144-1156. doi: 10.1002/dvdy.21094

Kim, W. J., Lim, M. N., Hong, Y., Silverman, E. K., Lee, J. H., Jung, B. H., . . . Lee, S. D. (2014). Association of lung function genes with chronic obstructive pulmonary disease. *Lung, 192*(4), 473-480. doi: 10.1007/s00408-014-9579-4

Klimentidis, Y. C., Vazquez, A. I., de Los Campos, G., Allison, D. B., Dransfield, M. T., & Thannickal, V. J. (2013). Heritability of pulmonary function estimated from pedigree and whole-genome markers. *Front Genet, 4*, 174. doi: 10.3389/fgene.2013.00174

Konermann, S., Brigham, M. D., Trevino, A. E., Hsu, P. D., Heidenreich, M., Cong, L., . . . Zhang, F. (2013). Optical control of mammalian endogenous transcription and epigenetic states. *Nature, 500*(7463), 472-476. doi: 10.1038/nature12466

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., . . . Zhang, F. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature, 517*(7536), 583-588. doi: 10.1038/nature14136

Kong, X., Cho, M. H., Anderson, W., Coxson, H. O., Muller, N., Washko, G., . . . Investigators, E. S. N. (2011). Genome-wide association study identifies BICD1 as a susceptibility gene for emphysema. *Am J Respir Crit Care Med, 183*(1), 43-49. doi: 10.1164/rccm.201004-0541OC

Kwan, T., Benovoy, D., Dias, C., Gurd, S., Provencher, C., Beaulieu, P., . . . Majewski, J. (2008). Genome-wide analysis of transcript isoform variation in humans. *Nat Genet, 40*(2), 225-231. doi: 10.1038/ng.2007.57

Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science, 265*(5181), 2037-2048.

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., . . . Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res, 22*(9), 1813-1831. doi: 10.1101/gr.136184.111

Latchman, D. S. (1997). Transcription factors: an overview. *Int J Biochem Cell Biol, 29*(12), 1305-1312.

Laurell, C. B., & Eriksson, S. (2013). The electrophoretic alpha1-globulin pattern of serum in alpha1-antitrypsin deficiency. 1963. *COPD, 10 Suppl 1*, 3-8. doi: 10.3109/15412555.2013.771956

Lebecque, P., Kiakulanda, P., & Coates, A. L. (1993). Spirometry in the asthmatic child: is FEF25-75 a more sensitive test than FEV1/FVC? *Pediatr Pulmonol, 16*(1), 19-22.

Lee, B. Y., Cho, S., Shin, D. H., & Kim, H. (2011). Genome-wide association study of copy number variations associated with pulmonary function measures in Korea Associated Resource (KARE) cohorts. *Genomics, 97*(2), 101-105. doi: 10.1016/j.ygeno.2010.11.001

Lee, J. H., Cho, M. H., Hersh, C. P., McDonald, M. L., Crapo, J. D., Bakke, P. S., . . . Investigators, E. (2014). Genetic susceptibility for chronic bronchitis in chronic obstructive pulmonary disease. *Respir Res, 15*, 113. doi: 10.1186/s12931-014-0113-2

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet, 95*(1), 5-23. doi: 10.1016/j.ajhg.2014.06.009

Lee, T. I., & Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annu Rev Genet, 34*, 77-137. doi: 10.1146/annurev.genet.34.1.77

Lewitter, F. I., Tager, I. B., McGue, M., Tishler, P. V., & Speizer, F. E. (1984). Genetic and environmental determinants of level of pulmonary function. *Am J Epidemiol, 120*(4), 518-530.

Li, H., Cho, S. N., Evans, C. M., Dickey, B. F., Jeong, J. W., & DeMayo, F. J. (2008). Cre-mediated recombination in mouse Clara cells. *Genesis, 46*(6), 300-307. doi: 10.1002/dvg.20396

Li, X., Hastie, A. T., Hawkins, G. A., Moore, W. C., Ampleford, E. J., Milosevic, J., . . . Bleecker, E. R. (2015). eQTL of bronchial epithelial cells and bronchial alveolar lavage deciphers GWAS-identified asthma genes. *Allergy*. doi: 10.1111/all.12683

Li, Y., & Tollefsbol, T. O. (2011). DNA methylation detection: bisulfite genomic sequencing analysis. *Methods Mol Biol, 791*, 11-21. doi: 10.1007/978-1-61779-316-5_2

Lieberman, J. (1969). Heterozygous and homozygous alpha-antitrypsin deficiency in patients with pulmonary emphysema. *N Engl J Med, 281*(6), 279-284. doi: 10.1056/NEJM196908072810601

Linton, J. M., Martin, G. R., & Reichardt, L. F. (2007). The ECM protein nephronectin promotes kidney development via integrin alpha8beta1-mediated stimulation of Gdnf expression. *Development, 134*(13), 2501-2509. doi: 10.1242/dev.005033

Loth, D. W., Artigas, M. S., Gharib, S. A., Wain, L. V., Franceschini, N., Koch, B., . . . London, S. J. (2014). Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat Genet, 46*(7), 669-677. doi: 10.1038/ng.3011

Lozano, R., Naghavi, M., Foreman, K., Lim, S., Shibuya, K., Aboyans, V., . . . Memish, Z. A. (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), 2095-2128. doi: 10.1016/S0140-6736(12)61728-0

Luo, W., Obeidat, M., Di Narzo, A. F., Chen, R., Sin, D. D., Pare, P. D., & Hao, K. (2015). Airway Epithelial Expression Quantitative Trait Loci Reveal Genes Underlying Asthma and Other Airway Diseases. *Am J Respir Cell Mol Biol*. doi: 10.1165/rcmb.2014-0381OC

Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H., & Joung, J. K. (2013). CRISPR RNA-guided activation of endogenous human genes. *Nat Methods, 10*(10), 977-979. doi: 10.1038/nmeth.2598

Maher, B. (2012). ENCODE: The human encyclopaedia. *Nature, 489*(7414), 46-48.

Majewski, J., & Pastinen, T. (2011). The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet, 27*(2), 72-79. doi: 10.1016/j.tig.2010.10.006

Mali, P., Aach, J., Stranges, P. B., Esvelt, K. M., Moosburner, M., Kosuri, S., . . . Church, G. M. (2013). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol, 31*(9), 833-838. doi: 10.1038/nbt.2675

Manichaikul, A., Hoffman, E. A., Smolonska, J., Gao, W., Cho, M. H., Baumhauer, H., . . . Barr, R. G. (2014). Genome-wide study of percent emphysema on computed tomography in the general population. The Multi-Ethnic Study of Atherosclerosis Lung/SNP Health Association Resource Study. *Am J Respir Crit Care Med, 189*(4), 408-418. doi: 10.1164/rccm.201306-1061OC

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511. doi: 10.1038/nrg2796

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature, 470*(7333), 198-203. doi: 10.1038/nature09796

Masters, J. R. (2000). Human cancer cell lines: fact and fantasy. *Nat Rev Mol Cell Biol, 1*(3), 233-236. doi: 10.1038/35043102

Mathers, C., Boerma, T. and Ma Fat, D. (2008). The global burden of disease: 2004 update: World Health Organisation.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., . . . Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science, 337*(6099), 1190-1195. doi: 10.1126/science.1222794

McClearn, G. E., Svartengren, M., Pedersen, N. L., Heller, D. A., & Plomin, R. (1994). Genetic and environmental influences on pulmonary function in aging Swedish twins. *J Gerontol, 49*(6), 264-268.

McCloskey, S. C., Patel, B. D., Hinchliffe, S. J., Reid, E. D., Wareham, N. J., & Lomas, D. A. (2001). Siblings of patients with severe chronic obstructive pulmonary disease have a significant risk of airflow obstruction. *Am J Respir Crit Care Med, 164*(8 Pt 1), 1419-1424. doi: 10.1164/ajrccm.164.8.2105002

Miller, J. C., Tan, S., Qiao, G., Barlow, K. A., Wang, J., Xia, D. F., . . . Rebar, E. J. (2011). A TALE nuclease architecture for efficient genome editing. *Nat Biotechnol, 29*(2), 143-148. doi: 10.1038/nbt.1755

Miller, L. A., Wert, S. E., Clark, J. C., Xu, Y., Perl, A. K., & Whitsett, J. A. (2004). Role of Sonic hedgehog in patterning of tracheal-bronchial cartilage and the peripheral lung. *Dev Dyn, 231*(1), 57-71. doi: 10.1002/dvdy.20105

Miller, M. R., Hankinson, J., Brusasco, V., Burgos, F., Casaburi, R., Coates, A., . . . Force, A. E. T. (2005). Standardisation of spirometry. *Eur Respir J, 26*(2), 319-338. doi: 10.1183/09031936.05.00034805

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet, 6*(4), e1000888. doi: 10.1371/journal.pgen.1000888

Niden, A. H. (1967). Bronchiolar and large alveolar cell in pulmonary phospholipid metabolism. *Science, 158*(3806), 1323-1324.

Obeidat, M., Hao, K., Bosse, Y., Nickle, D. C., Nie, Y., Postma, D. S., . . . Pare, P. D. (2015). Molecular mechanisms underlying variations in lung function: a systems genetics analysis. *Lancet Respir Med*. doi: 10.1016/S2213-2600(15)00380-X

Obeidat, M., Miller, S., Probert, K., Billington, C. K., Henry, A. P., Hodge, E., . . . Hall, I. P. (2013). GSTCD and INTS12 regulation and expression in the human lung. *PLoS One, 8*(9), e74630. doi: 10.1371/journal.pone.0074630

Obeidat, M., Wain, L. V., Shrine, N., Kalsheker, N., Soler Artigas, M., Repapi, E., . . . SpiroMeta, C. (2011). A comprehensive evaluation of potential lung function associated genes in the SpiroMeta general population sample. *PLoS One, 6*(5), e19382. doi: 10.1371/journal.pone.0019382

Ober, C., Abney, M., & McPeek, M. S. (2001). The genetic dissection of complex traits in a founder population. *Am J Hum Genet, 69*(5), 1068-1079. doi: 10.1086/324025

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven Symp Biol, 23*, 366-370.

Palmer, L. J., Celedon, J. C., Chapman, H. A., Speizer, F. E., Weiss, S. T., & Silverman, E. K. (2003). Genome-wide linkage analysis of bronchodilator responsiveness and post-bronchodilator spirometric phenotypes in chronic obstructive pulmonary disease. *Hum Mol Genet, 12*(10), 1199-1210.

Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA, 299*(11), 1335-1344. doi: 10.1001/jama.299.11.1335

Pellegrino, R., Viegi, G., Brusasco, V., Crapo, R. O., Burgos, F., Casaburi, R., . . . Wanger, J. (2005). Interpretative strategies for lung function tests. *Eur Respir J, 26*(5), 948-968. doi: 10.1183/09031936.05.00035205

Pepicelli, C. V., Lewis, P. M., & McMahon, A. P. (1998). Sonic hedgehog regulates branching morphogenesis in the mammalian lung. *Curr Biol, 8*(19), 1083-1086.

Perl, A. K., Wert, S. E., Loudy, D. E., Shan, Z., Blair, P. A., & Whitsett, J. A. (2005). Conditional recombination reveals distinct subsets of epithelial cells in trachea, bronchi, and alveoli. *Am J Respir Cell Mol Biol, 33*(5), 455-462. doi: 10.1165/rcmb.2005-0180OC

Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., . . . Investigators, I. (2009). A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet, 5*(3), e1000421. doi: 10.1371/journal.pgen.1000421

Pillai, S. G., Kong, X., Edwards, L. D., Cho, M. H., Anderson, W. H., Coxson, H. O., . . . Investigators, I. (2010). Loci identified by genome-wide association studies influence different disease-related phenotypes in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med, 182*(12), 1498-1505. doi: 10.1164/rccm.201002-0151OC

Portelli, M. A., Siedlinski, M., Stewart, C. E., Postma, D. S., Nieuwenhuis, M. A., Vonk, J. M., . . . Sayers, I. (2014). Genome-wide protein QTL mapping identifies human plasma kallikrein as a post-translational regulator of serum uPAR levels. *FASEB J, 28*(2), 923-934. doi: 10.1096/fj.13-240879

Ptashne, M., & Gann, A. (1997). Transcriptional activation by recruitment. *Nature, 386*(6625), 569-577. doi: 10.1038/386569a0

Qu, H., & Fang, X. (2013). A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics Proteomics Bioinformatics, 11*(3), 135-141. doi: 10.1016/j.gpb.2013.05.001

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi: 10.1093/bioinformatics/btq033

Rabe, K. F., Hurd, S., Anzueto, A., Barnes, P. J., Buist, S. A., Calverley, P., . . . Global Initiative for Chronic Obstructive Lung, D. (2007). Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am J Respir Crit Care Med, 176*(6), 532-555. doi: 10.1164/rccm.200703-456SO

Rands, C. M., Meader, S., Ponting, C. P., & Lunter, G. (2014). 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet, 10*(7), e1004525. doi: 10.1371/journal.pgen.1004525

Redline, S., Tishler, P. V., Lewitter, F. I., Tager, I. B., Munoz, A., & Speizer, F. E. (1987). Assessment of genetic and nongenetic influences on pulmonary function. A twin study. *Am Rev Respir Dis, 135*(1), 217-222.

Repapi, E., Sayers, I., Wain, L. V., Burton, P. R., Johnson, T., Obeidat, M., . . . Tobin, M. D. (2010). Genome-wide association study identifies five loci associated with lung function. *Nat Genet, 42*(1), 36-44. doi: 10.1038/ng.501

Rockman, M. V., & Kruglyak, L. (2006). Genetics of global gene expression. *Nat Rev Genet, 7*(11), 862-872. doi: 10.1038/nrg1964

Sambamurthy, N., Leme, A. S., Oury, T. D., & Shapiro, S. D. (2015). The receptor for advanced glycation end products (RAGE) contributes to the progression of emphysema in mice. *PLoS One, 10*(3), e0118979. doi: 10.1371/journal.pone.0118979

Sander, J. D., & Joung, J. K. (2014). CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol, 32*(4), 347-355. doi: 10.1038/nbt.2842

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res, 22*(9), 1748-1759. doi: 10.1101/gr.136127.111

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet, 15*(4), 272-286. doi: 10.1038/nrg3682

Shockett, P., Difilippantonio, M., Hellman, N., & Schatz, D. G. (1995). A modified tetracycline-regulated system provides autoregulatory, inducible gene expression in cultured cells and transgenic mice. *Proc Natl Acad Sci U S A, 92*(14), 6522-6526.

Silverman, E. K., Chapman, H. A., Drazen, J. M., Weiss, S. T., Rosner, B., Campbell, E. J., . . . Speizer, F. E. (1998). Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. *Am J Respir Crit Care Med, 157*(6 Pt 1), 1770-1778. doi: 10.1164/ajrccm.157.6.9706014

Silverman, E. K., Mosley, J. D., Palmer, L. J., Barth, M., Senter, J. M., Brown, A., . . . Weiss, S. T. (2002). Genome-wide linkage analysis of severe, early-onset chronic obstructive pulmonary disease: airflow obstruction and chronic bronchitis phenotypes. *Hum Mol Genet, 11*(6), 623-632.

Silverman, E. K., Palmer, L. J., Mosley, J. D., Barth, M., Senter, J. M., Brown, A., . . . Weiss, S. T. (2002). Genomewide linkage analysis of quantitative spirometric phenotypes in severe early-onset chronic obstructive pulmonary disease. *Am J Hum Genet, 70*(5), 1229-1239. doi: 10.1086/340316

Simon, M. R., Chinchilli, V. M., Phillips, B. R., Sorkness, C. A., Lemanske, R. F., Jr., Szefler, S. J., . . . Blood, I. (2010). Forced expiratory flow between 25% and 75% of vital capacity and FEV1/forced vital capacity ratio in relation to clinical

and physiological parameters in asthmatic children with normal FEV1 values. *J Allergy Clin Immunol, 126*(3), 527-534 e521-528. doi: 10.1016/j.jaci.2010.05.016

Sisson, T. H., Hansen, J. M., Shah, M., Hanson, K. E., Du, M., Ling, T., . . . Christensen, P. J. (2006). Expression of the reverse tetracycline-transactivator gene causes emphysema-like changes in mice. *Am J Respir Cell Mol Biol, 34*(5), 552-560. doi: 10.1165/rcmb.2005-0378OC

Smith, H. W., & Marshall, C. J. (2010). Regulation of cell signalling by uPAR. *Nat Rev Mol Cell Biol, 11*(1), 23-36. doi: 10.1038/nrm2821

Soler Artigas, M., Loth, D. W., Wain, L. V., Gharib, S. A., Obeidat, M., Tang, W., . . . Tobin, M. D. (2011). Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet, 43*(11), 1082-1090. doi: 10.1038/ng.941

Soler Artigas, M., Wain, L. V., Repapi, E., Obeidat, M., Sayers, I., Burton, P. R., . . . SpiroMeta, C. (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. *Am J Respir Crit Care Med, 184*(7), 786-795. doi: 10.1164/rccm.201102-0192OC

Soler Artigas M, W. L., Miller S, Kheirallah AK, Huffman J, Ntalla I, Shrine N, Obeidat M, Trochet H, McArdle W, Couto Alves A, Hui J, Zhao JH, Joshi P, Teumer A, Albrecht E, Imboden M, Rawal R, Lopez L, Marten J, Enroth S, Surakka I, Polasek O, Lyytikäinen LP, Granell R, Hysi P, Flexeder C, Mahajan A, Beilby J, Bossé Y, Brandsma CA, Campbell H, Gieger C, Gläser S, Gonzalez J, Grallert H, Hammond C, Harris S, Hartikainen AL, Hayward C, Heliövaara M, Henderson J, Hocking L, Horikoshi M, Hutri-Kähönen N, Ingelsson E, Johansson A, Kemp J, Kolcic I, Kumar A, Lind L, Melén E, Musk A, Navarro P, Nickle D, Padmanabhan S, Raitakari O, Ried J, Ripatti S, Schulz H, Scott R, Sin D, Starr J, Viñuela A, Völzke H, Wild S, Wright A, Zemunik T, Jarvis D, Spector T, Evans D, Lehtimäki T, Vitart V, Kähönen M, Gyllensten U, Rudan I, Deary I, Karrasch S, Probst-Hensch N, Heinrich J, Koch B, Wilson J, Wareham N, James A, Morris A, Jarvelin MR, Sayers I, Strachan D, Hall IP, and Tobin M. (2015). Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nature Communications*.

Sparvero, L. J., Asafu-Adjei, D., Kang, R., Tang, D., Amin, N., Im, J., . . . Lotze, M. T. (2009). RAGE (Receptor for Advanced Glycation Endproducts), RAGE ligands, and their role in cancer and inflammation. *J Transl Med, 7*, 17. doi: 10.1186/1479-5876-7-17

Stewart, C. E., Nijmeh, H. S., Brightling, C. E., & Sayers, I. (2012). uPAR regulates bronchial epithelial repair in vitro and is elevated in asthmatic epithelium. *Thorax, 67*(6), 477-487. doi: 10.1136/thoraxjnl-2011-200508

Stewart, C. E., & Sayers, I. (2009). Characterisation of urokinase plasminogen activator receptor variants in human airway and peripheral cells. *BMC Mol Biol, 10*, 75. doi: 10.1186/1471-2199-10-75

Stogsdill, M. P., Stogsdill, J. A., Bodine, B. G., Fredrickson, A. C., Sefcik, T. L., Wood, T. T., . . . Reynolds, P. R. (2013). Conditional overexpression of receptors for advanced glycation end-products in the adult murine lung causes airspace enlargement and induces inflammation. *Am J Respir Cell Mol Biol, 49*(1), 128-134. doi: 10.1165/rcmb.2013-0013OC

Stoller, J. K., & Aboussouan, L. S. (2005). Alpha1-antitrypsin deficiency. *Lancet, 365*(9478), 2225-2236. doi: 10.1016/S0140-6736(05)66781-5

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A, 102*(43), 15545-15550. doi: 10.1073/pnas.0506580102

Sun, W., & Hu, Y. (2013). eQTL Mapping Using RNA-seq Data. *Stat Biosci, 5*(1), 198-219. doi: 10.1007/s12561-012-9068-3

Syvanen, A. C. (2005). Toward genome-wide SNP genotyping. *Nat Genet, 37 Suppl*, S5-10. doi: 10.1038/ng1558

Tang, W., Kowgier, M., Loth, D. W., Soler Artigas, M., Joubert, B. R., Hodge, E., . . . Cassano, P. A. (2014). Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. *PLoS One, 9*(7), e100776. doi: 10.1371/journal.pone.0100776

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., . . . Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature, 489*(7414), 75-82. doi: 10.1038/nature11232

Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C. A., Amini, B., Andersen, E., . . . Ponten, F. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics, 4*(12), 1920-1932. doi: 10.1074/mcp.M500279-MCP200

Van Durme, Y. M., Eijgelsheim, M., Joos, G. F., Hofman, A., Uitterlinden, A. G., Brusselle, G. G., & Stricker, B. H. (2010). Hedgehog-interacting protein is a COPD susceptibility gene: the Rotterdam Study. *Eur Respir J, 36*(1), 89-95. doi: 10.1183/09031936.00129509

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351. doi: 10.1126/science.1058040

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *Am J Hum Genet, 90*(1), 7-24. doi: 10.1016/j.ajhg.2011.11.029

Wain L, S. N., Miller S, Jackson V, Ntalla I, Soler Artigas M, Billington CK, Kheirallah AK, Allen R, Cook JP, Probert K, Obeidat M, Bossé Y, Hao K, Postma DS, Paré PD, Ramasamy A, UK Brain Expression Consortium, Mägi R, Mihailov E, Reinmaa E, Melén E, O'Connell J, Frangou E, Delaneau O, OxGSK Consortium, Freeman C, Petkova D, McCarthy M, Sayers I, Deloukas P, Hubbard R, Pavord I, Hansell A, Thomson NC, Zeggini E, Morris AP, Marchini J, Strachan DP, Tobin MD, Hall IP (2015). Novel insights into the genetics of smoking behaviour, lung function and chronic obstructive pulmonary disease in UK Biobank. *The Lancet Respiratory Medicine*.

Wain, L. V., Odenthal-Hesse, L., Abujaber, R., Sayers, I., Beardsmore, C., Gaillard, E. A., . . . Hollox, E. J. (2014). Copy number variation of the beta-defensin genes in europeans: no supporting evidence for association with lung function, chronic obstructive pulmonary disease or asthma. *PLoS One, 9*(1), e84192. doi: 10.1371/journal.pone.0084192

Wang, K., Li, M., & Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nat Rev Genet, 11*(12), 843-854. doi: 10.1038/nrg2884

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63. doi: 10.1038/nrg2484

Ward, L. D., & Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically

linked variants. *Nucleic Acids Res, 40*(Database issue), D930-934. doi: 10.1093/nar/gkr917

Westra, H. J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., . . . Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet, 45*(10), 1238-1243. doi: 10.1038/ng.2756

Wilk, J. B., Chen, T. H., Gottlieb, D. J., Walter, R. E., Nagle, M. W., Brandler, B. J., . . . O'Connor, G. T. (2009). A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet, 5*(3), e1000429. doi: 10.1371/journal.pgen.1000429

Wilk, J. B., DeStefano, A. L., Arnett, D. K., Rich, S. S., Djousse, L., Crapo, R. O., . . . Myers, R. H. (2003). A genome-wide scan of pulmonary function measures in the National Heart, Lung, and Blood Institute Family Heart Study. *Am J Respir Crit Care Med, 167*(11), 1528-1533. doi: 10.1164/rccm.200207-755OC

Wilk, J. B., Djousse, L., Arnett, D. K., Rich, S. S., Province, M. A., Hunt, S. C., . . . Myers, R. H. (2000). Evidence for major genes influencing pulmonary function in the NHLBI family heart study. *Genet Epidemiol, 19*(1), 81-94. doi: 10.1002/1098-2272(200007)19:1<81::AID-GEPI6>3.0.CO;2-8

Wilk, J. B., Shrine, N. R., Loehr, L. R., Zhao, J. H., Manichaikul, A., Lopez, L. M., . . . Stricker, B. H. (2012). Genome-wide association studies identify CHRNA5/3 and HTR4 in the development of airflow obstruction. *Am J Respir Crit Care Med, 186*(7), 622-632. doi: 10.1164/rccm.201202-0366OC

Wilk, J. B., Walter, R. E., Laramie, J. M., Gottlieb, D. J., & O'Connor, G. T. (2007). Framingham Heart Study genome-wide association: results for pulmonary function measures. *BMC Med Genet, 8 Suppl 1*, S8. doi: 10.1186/1471-2350-8-S1-S8

Wohlsen, A., Martin, C., Vollmer, E., Branscheid, D., Magnussen, H., Becker, W. M., . . . Uhlig, S. (2003). The early allergic response in small airways of human precision-cut lung slices. *Eur Respir J, 21*(6), 1024-1032.

Wu, L., Ma, L., Nicholson, L. F., & Black, P. N. (2011). Advanced glycation end products and its receptor (RAGE) are increased in patients with COPD. *Respir Med, 105*(3), 329-336. doi: 10.1016/j.rmed.2010.11.001

Xu, S., Grullon, S., Ge, K., & Peng, W. (2014). Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. *Methods Mol Biol, 1150*, 97-111. doi: 10.1007/978-1-4939-0512-6_5

Yan, S. D., Chen, X., Fu, J., Chen, M., Zhu, H., Roher, A., . . . Schmidt, A. M. (1996). RAGE and amyloid-beta peptide neurotoxicity in Alzheimer's disease. *Nature, 382*(6593), 685-691. doi: 10.1038/382685a0

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol, 9*(9), R137. doi: 10.1186/gb-2008-9-9-r137

Zhou, J. J., Cho, M. H., Castaldi, P. J., Hersh, C. P., Silverman, E. K., & Laird, N. M. (2013). Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am J Respir Crit Care Med, 188*(8), 941-947. doi: 10.1164/rccm.201302-0263OC

Zhou, X., Baron, R. M., Hardin, M., Cho, M. H., Zielinski, J., Hawrylkiewicz, I., . . . Silverman, E. K. (2012). Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Hum Mol Genet, 21*(6), 1325-1335. doi: 10.1093/hmg/ddr569

Zhu, G., Warren, L., Aponte, J., Gulsvik, A., Bakke, P., Anderson, W. H., . . . International, C. G. N. I. (2007). The SERPINE2 gene is associated with chronic obstructive pulmonary disease in two large populations. *Am J Respir Crit Care Med, 176*(2), 167-173. doi: 10.1164/rccm.200611-1723OC