A GENOME-WIDE REGULATORY NETWORK OF *INTS12* ASSOCIATED WITH PULMONARY FUNCTION

By

Alexander K. Kheirallah, BSc (Hons)

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

Nottingham, 2017

Abstract

Genome-wide association studies of human lung function and Chronic Obstructive Pulmonary Disease have identified a highly significant and reproducible signal on 4q24. It remains unclear which of the two candidate genes within this locus may regulate lung function: *GSTCD*, a gene with unknown function, and/or *INTS12,* a member of the Integrator Complex which is currently thought to mediate 3'end processing of small nuclear RNAs. An interrogation of bioinformatic datasets showed that in lung tissue, 4q24 polymorphisms associated with lung function correlate with *INTS12* but not neighboring *GSTCD* expression. In contrast to the previous reports in other species, a minor alteration of small nuclear RNA processing was observed following INTS12 depletion. RNA sequencing analysis of knockdown cells instead revealed dysregulation of a core subset of genes relevant to airway biology and a robust downregulation of protein synthesis pathways. Consistent with this, protein translation was decreased in INTS12 knockdown cells. In addition, chromatin immunoprecipitation and sequencing experiments demonstrated INTS12 binding throughout the genome, which was enriched in transcriptionally active regions. Finally, INTS12 regulome was defined which includes genes belonging to the protein synthesis pathways. INTS12 has functions beyond the canonical snRNA processing and evidence is presented showing that it regulates translation by directly controlling the expression of genes belonging to protein synthesis pathways. This thesis provides a detailed analysis of INTS12 activities on a genome-wide scale and contributes to the understanding of biology behind the genetic association for lung function at the 4q24.

Acknowledgments

First, I would like to thank Reham Alhallak, *la mia vera anima gamella*, for her enormous help throughout this journey. I would like to thank my mother Hola Adi for investing in my educational and physical development. I also would like to thank my teachers and tutors back in Poland whose contributions laid the foundations for my intellectual activities. Big appreciations go to my sisters Mira and Maja for the warm times we spent together when visiting me in the UK. My thanks also go to my father Mouetaz Kheirallah.

Secondly, I would like to say big thank you to my supervisors Professor Ian Hall and Dr Ian Sayers. I would like to thank Ian Hall in particular for giving me the freedom during the PhD which allowed me to flourish. Also I would like to thank him for opening the possibility to contribute to scientific publications. And I would like to thank Ian Sayers for offering me the opportunity to write a review paper and showing me the importance of organization and scientific methodology. I also would like to thank my internal assessor Dr Cornelia de Moor for working with me on collaborative experiments and scrutinizing my data. There is no doubt that without Cornelia's suggestions this PhD would not have shaped as it did.

Thirdly, I would like to thank George Blundell-Hunter, Michael Tellier, Annie Quandt, Carlos Carrasco and Phillip Quinlan for working with me on the development of heritability of human voice parameters study which we managed to do outside our core duties.

Finally, my thanks go to my friends at the Division of Respiratory Medicine.

I dedicate this piece of work to the aspiring free Syrian people and innocent victims of war…

Imagination is more important than knowledge. For knowledge is limited to all we now know and understand, while imagination embraces the entire world, and all there ever will be to know and understand.

Albert Einstein

Table of Contents

Abbreviations

ASNS - Asparagine Synthetase

ATF4 – activating transcription factor 4

ChIPseq – chromatin immunoprecipitation and sequencing

cis-eQTL – nearby expression quantitative trait locus

COPD – Chronic Obstructive Pulmonary Disease

CPM – counts per methionine

CTCF - CCCTC-binding factor

D-siRNA – Dicer substrate small interfering RNA

eQTL – expression quantitative trait locus

GARS – glycyl-tRNA synthetase

GSEA – gene set enrichment analysis

GSTCD – Glutathione S-transferase, C-terminal Domain Containing

GWAS – genome-wide association studies

H3K27me3 – histone 3 lysine 27 trimethylation

H3K36me3 – histone 3 lysine 36 trimethylation

H3K4me3 – histone 3 lysine 4 trimethylation

HBEC – human bronchial epithelial cell

IL1R1 - interleukin 1 receptor 1

INTS12 – Integrator Complex subunit 12

INTScom – Integrator Complex

LEP – leptin

lincRNA – long intergenic RNA

MARS – methionyl-tRNA synthetase

PHD – plant homeodomain

POLII – RNA polymerase II

qPCR – quantitative polymerase chain reaction

RNAi – RNA interference

RNAseq – RNA sequencing

SERPINA1 – α1-antitrypsin

snoRNA – small nucleolar RNA

SNP – single nucleotide polymorphism

snRNA – small nuclear RNA

TES – transactional end site

TGFβI - transforming growth factor β 1

trans-eQTL – distant eQTL

TSS – transcriptional start site

1. General introduction

1.1 Overview

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 et al. 2008). Worldwide there are approximately 500 million people suffering from obstructive lung disease. 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 treatments, their quality of life and life expectancy are improved. 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 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 identification of new targets for therapeutic intervention and generation of novel functional hypotheses that can be verified experimentally. Genetic studies of human lung function have identified a highly significant and reproducible signal on 4q24. The

mechanistic basis for this association has not been elucidated. Integrator Complex subunit 12 (*INTS12*[1]) has the potential biological role in normal lung function and development of lung disease as it is located in an replicated locus for genetic variability in lung function and risk of COPD at 4q24.

This chapter describes the 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. As the focus of the thesis is on *INTS12* and its protein product, genetic data which have demonstrated its association with lung function parameters as well as its canonical biological functions are reviewed.

1.2 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 inexpensive platforms for genotyping subjects on a genome wide basis with adequate coverage to permit genetic association signals to be detected. This has also been facilitated by the formation of international consortia (International HapMap, 2005; Genomes Project, 2012) providing a large number of samples and thus adequate statistical power, leading to a number of genome-wide association studies (GWAS) publications on a range of respiratory related phenotypes. GWAS were preceded by linkage analyses that had a limited success, while the very initial studies were concerned with heritability estimations.

1.2.1 Definition of lung function and COPD phenotypes

Before considering genetics, the phenotypic manifestation of COPD and lung function needs to be addressed. 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

[1] Genes are referred in italics (e.g. *INTS12*) while mRNA and protein are not referred in italics (e.g. INTS12)

first second (FEV_1). These are measurements with a general consensus regarding their derivation using spirometry (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 defined as <80% of the predicted value based on age, gender and height (FEV_1 (Percent Predicted)). 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). 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 suggest that $FEV_{25-75\%}$ is a sensitive index of airway obstruction (Lebecque et al. 1993; Simon et al. 2010), while other studies suggest this index is of limited value in this regard (Ciprandi et al. 2012).

COPD is a leading cause of death and chronic morbidity throughout the world. Three in every thousand people are diagnosed with COPD each year and the incidence increases rapidly with age (Afonso, Verhamme, Sturkenboom, and Brusselle, 2011). COPD has previously been defined and graded using the GOLD criteria (Table 1.1; Hurd and Pauwels, 2002; Pauwels et al. 2001) which have been updated to consider symptoms and frequency of exacerbations ((GOLD), 2015). The clinical presentation of COPD is diverse (Hansen et al. 2007; Pellegrino et al. 2005). It is a progressive disabling condition characterised 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. Although cigarette smoking is a major risk

factor for the development of COPD, only 15 to 20% of smokers manifest clinically significant COPD (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. Inflammation is typically long-term and is therefore called chronic bronchitis. Emphysema is a sub-phenotype of COPD and is characterized by enlargement of distal airspaces due to the destruction of the airway walls (i.e. parenchymal destruction) (Hemminki et al. 2008).

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

Table 1.1: GOLD standards of airway limitation severity in COPD.

1.2.2 Familial aggregation of spirometric measures and COPD in families

Familial aggregation studies provide diverse but not completely consistent evidence implicating genetic factors in lung function phenotypes. An early twin study of 127 monozygotic and 141 dizygotic twin pairs by Hubert et al. demonstrated that FEV_1 and FVC 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 et al. 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 (Lewitter et al. 1984; Coultas et al. 1991; McClearn et al. 1994; Wilk et al. 2000; Ober et al. 2001). 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 environment.

Silverman et al. have shown that the risk of COPD is approximately 2-3 higher in smokers who have a first degree relative affected by COPD suggesting the presence of genetic 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 et al. 2008). 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% (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.

1.3 A brief historical overview of molecular genetics and functional studies in pulmonary physiology

1.3.1 The discovery of α-1-antitrypsin (A1AT) deficiency

The first gene linked to emphysema was the *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 et al. 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 (initial discoveries described in 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 and Eriksson, 2013; re-print of original publication). 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). The susceptibility variant was called the Z allele and it was concluded that it 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 (DeMeo and 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 and Aboussouan, 2005). Therefore, despite

the unprecedented genetic, molecular and mechanistic advances in the understanding of A1AT deficiency as related to 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 (e.g. smoking cessation) and, in some countries, A1AT replacement therapy (Petrache, 2009).

1.3.2 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 or dichotomous trait. 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 in advance) 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 and Schork, 1994).

The distance between two genetic markers 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 (Kheirallah et al. 2016).

The statistical significance of the linkage is commonly measured by the LOD score, which is the logarithm to the base of ten of the ratio of the data's likelihood given linkage to the likelihood of no linkage (Morton, 1955). 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 a highly significant linkage (Kheirallah et al. 2016). 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 1.2. These studies were problematic due to lack of sufficient replication leveraging independent clinical cohorts.

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.12 at 36cM	FEV_1/FVC	3.26	DeMeo et al. (2004)

Table 1.2: Summary of linkage studies for lung function and COPD phenotypes.

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 multiple common variants of modest effect sizes important in complex 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.

1.3.3 Translation of genome wide linkage scans to candidate genes

DeMeo et al. performed a follow-up study to identify the most likely causative gene behind the FEV_1/FVC linkage signal on chromosome 2q (Silverman, Palmer, et al. 2002; DeMeo et al. 2006). This was achieved by interrogating the transcriptomic profiling with genetic approaches. 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). Immunohistochemistry (IHC) was used to demonstrate *SERPINE2* expression in both mouse and human lung tissue. Positive staining was demonstrated in healthy, emphysematous, and asthmatic lungs. However, *SERPINE2* expression was only moderately increased in COPD (1.25-fold difference) and the observed effect did not meet the 5% false discovery rate. 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. More recently it was identified that SNPs within *SERPINE2* were associated with airway wall thickness as well as *SERPINE2* levels in the human lung (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 (DeMeo et al. 2009). 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 gene (*IREB2*) met the 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 mechanisms.

In addition to candidate genes from linkage regions there have been a large number of candidate gene studies in COPD (Sandford et al. 2002; Wood and Stockley, 2006; Hersh et al. 2008). 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 (Hardin, 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).

1.3.4 Genome-wide association studies

The advent of GWAS is attributable to advances in genotyping technology (Syvanen, 2005), the Human Genome Project (Lander et al. 2001; Venter et al. 2001) and the completion of the HapMap project (International HapMap, 2005). The basic rational of GWAS is similar to that of linkage scans: 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 and Manolio, 2008). Importantly, these SNPs are selected to capture the maximum information on the human genome by using optimised panels that tag haplotype blocks. This 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% allele frequency) and they are less well designed to evaluate low allele frequency variants (Hirschhorn and Daly, 2005). This is in contrast to family-based linkage approaches which are ideally suited for detecting rare genetic variants of large or moderate phenotypic effects. However, GWAS generally offer greater resolution and more power in association mapping.

GWAS rely on appropriate reconstruction of haplotypes based on a population data however results may be misleading if this reconstruction is erroneous. This is because investigators may use one SNP (also known as tag or sentinel 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). 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 and Daly, 2005). It is critical to remember that mere association does not imply causation and that a significantly associated SNP located in one gene may be tagging the causative variant present elsewhere on a completely different gene. Thus it is advantageous to refer to common haplotypes rather than individual SNPs as associated with any particular phenotype of interest. 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.

1.4 The landscape of GWAS for lung function measure FEV_1 and COPD

Individual GWAS of lung function measures have identified a number of candidate SNPs potentially involved with human lung function measures and risk of COPD. Notably, between 2006 and 2010 there were several small GWAS utilizing high throughput SNP genotyping for association mapping with lung function and COPD. These studies identified several genetic loci underlying these traits including haplotypes containing 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). Importantly, while these studies demonstrated the potential to identify novel lung function and COPD loci using GWAS 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 to increase study power for novel candidate gene discovery. A key component of these meta-analyses is the use of

imputation whereby genetic variation is not directly genotyped on the specific genotyping platform, but can be inferred with a measurable degree of confidence using reference genomes. These are now available from the HapMap project and subsequently the 1,000 and 10,000 genomes initiatives (Marchini and Howie, 2010). This approach makes possible the combining of genotyping data generated on a diverse number of genotyping platforms from individual studies making meta-analysis a feasible approach.

Whilst there have now been a number of such studies, the first two of these studies was the SpiroMeta and Charge consortia that investigated FEV_1 as well as FEV_1/FVC and were published in 2010 (Hancock et al. 2010; Repapi et al. 2010). These studies had large discovery and replication samples. In SpiroMeta study 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.

1.4.1 Meta-analyses of Forced Expiratory Volume in the first second

In the SpiroMeta study (Repapi et al. 2010) four loci were reported as reaching genome wide significance for FEV_1 including common variants at both known and novel loci (Figure 1.1):

- 2q35 locus in linkage with Tensin 1 (*TNS1)*
- 4q24 locus near Glutathione S-Transferase, C-Terminal Domain Containing (*GSTCD)*, Intergrator Complex Subunit 12 (*INTS12*), and nephronectin (*NPNT*)
- 5q33 locus in proximity to 5-Hydroxytryptamine (Serotonin) Receptor 4 (*HTR4)*
- 4q31 locus containing *HHIP*

The Charge consortium (Hancock et al. 2010) 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.

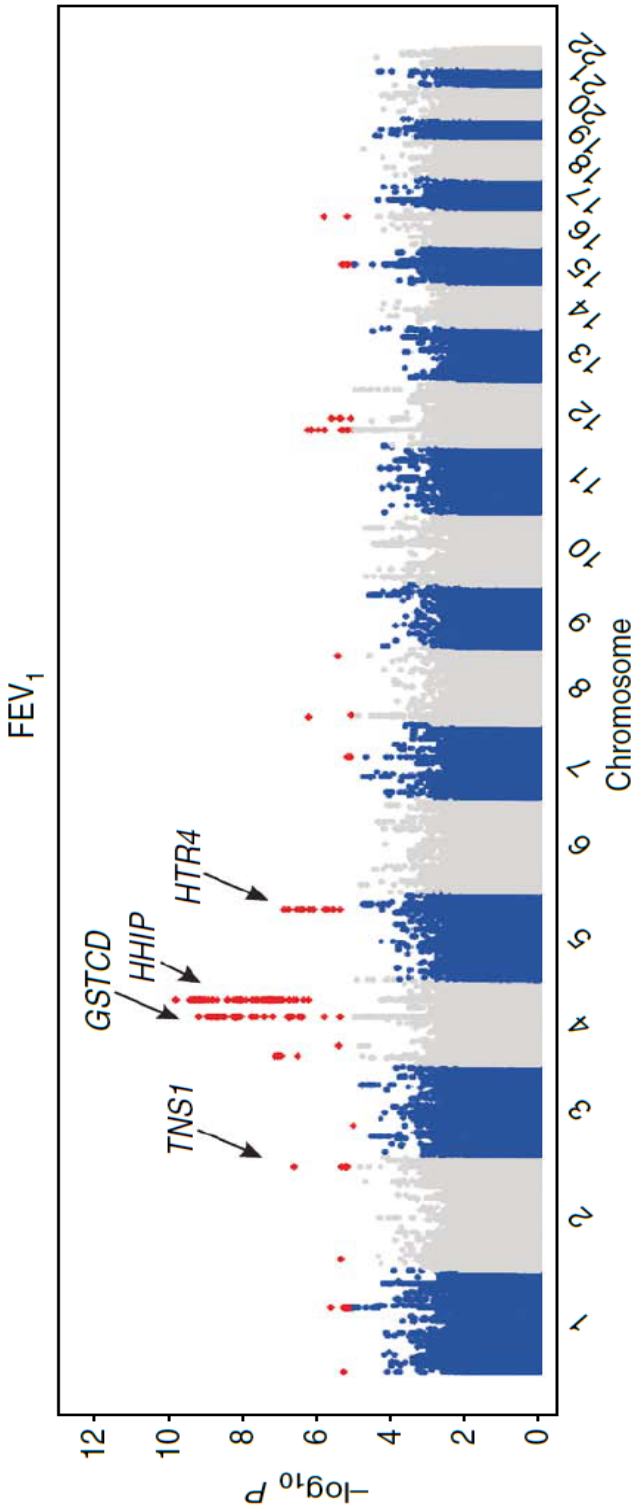


Figure 1.1: Example of manhattan plot of association results for FEV_1. Plot ordered by chromosome position. SNPs with $-\log_{10}P > 5$ are indicated in red. The four loci indicated by arrows showed association with FEV1 ($P < 5 \times 10^{-8}$) in the meta-analysis. Reproduced from Repapi et al. 2010.

1.4.2 Genome-wide association studies and meta-analyses in COPD

In addition to associations with FEV_1, unsurprisingly, these SNPs have also been associated with COPD susceptibility (Brehm et al. 2011; Castaldi et al. 2011; Chen et al. 2015; Cho et al. 2010; Cho et al. 2012; Kim et al. 2014; Pillai et al. 2009; Soler-Artigas, Loth et al. 2011; 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 (Lee et al. 2014).

Numerous COPD susceptibility regions have been identified. SNPs with the lowest P values studied ($P \leq 1 \times 10^{-9}$) were near *FAM13A* (Cho et al. 2010; Cho et al. 2012; Cho et al. 2014), *HHIP* (Cho et al. 2014; Van Durme et al. 2010), *CHRNA3* (Cho et al. 2014; Pillai et al. 2009), Succinate Dehydrogenase Complex Assembly Factor 3 (*SDHAF3*) (Kim et al. 2014), RAS oncogene family member (*RAB4B*) (Cho et al. 2012) and hydroxylysine kinase (*HYKK*) (Wilk et al. 2012). There is a significant overlap between SNPs associated with FEV_1 and COPD. These include *TNS1*'s haplotype SNP rs2571445 (Repapi et al. 2010; Soler-Artigas, Wain, et al. 2011), many SNPs at the 4q24 *INTS12*/*GSTCD*/*NPNT* haplotype locus (Castaldi et al. 2011; Soler-Artigas, Loth et al. 2011), *HHIP*'s haplotype SNP rs12604628 (Soler-Artigas, Loth et al. 2011), *HTR4*'s haplotype SNP rs3995090 (Repapi et al. 2010; Soler-Artigas, Loth et al. 2011). Also, different SNPs within the haplotype of Transforming Growth Factor, Beta 2 (*TGFB2*) were identified (Cho et al. 2014; Soler-Artigas, Wain, et al. 2011), as well as different SNPs in *HTR4*'s haplotype (Hancock et al. 2010; Soler-Artigas, Wain, et al. 2011; Wilk et al. 2012).

1.4.3 *INTS12*'s haplotype associates with pulmonary function and COPD

The effects of genetic variants at 4q24 on lung function have high statistical significance. In the Repapi et al. study the magnitude of effect on FEV_1 of *INTS12*/*GSTCD*/*NPNT* haplotype sentinel SNP was 52mL per alternative allele change which is equivalent to about 3 years of FEV_1 decline in the non-smoking population (Figure 1.2; Repapi et al. 2010). The sentinel SNP for this locus (rs10516526) was also significantly associated with FVC but had no visible effect on FEV_1/FVC, suggesting the correlation to be specifically for expiration parameters rather than for the relationship between them. Interestingly, Repapi et al. found that smoking exposure adjustment does not attenuate the observed association in *INTS12*/*GSTCD*/*NPNT* locus (Repapi et al. 2010), implying a limted role of smoking in driving the association. As mentioned, *INTS12*/*GSTCD*/*NPNT* locus association with FEV_1 was also reported by Hancock et al. (Figure 1.2; Hancock et al. 2010).

Later, Soler-Artigas et al. tested the hypothesis that loci previously associated with lung function are also associated with COPD (Soler-Artigas, Loth et al. 2011). In this study researchers looked at five previously reported sentinel SNPs, including *INTS12*/*GSTCD*/*NPNT* haplotype's rs10516526 variant, between 2890 COPD cases and 13,862 controls. Results showed that the *INTS12*/*GSTCD*/*NPNT* locus represented by rs10516526 is significantly associated with COPD disease status in GOLD stage 3 or 4. Similar study by Castaldi et al. demonstrated similar association between *INTS12*/*GSTCD*/*NPNT* haplotype tagged by rs4235415 with COPD (Castaldi et al. 2011). More recently, two new meta-analyses by Soler-Artigas et al. and Wain et al. have again reported the association of *INTS12*/*GSTCD*/*NPNT* locus with FEV_1 (Soler-Artigas et al. 2015; Wain et al. 2015). The identified association signal represents a linked haplotype that is ~$600x10^3$ – $1000x10^3$ base pairs (bp) long. However, an additional novel signal is thought to have been identified at 4q24 locus and thought to be independent from the *INTS12*/*GSTCD* haplotype and thus it has been argued that there are three independent signals in this region (Wain et al. 2015).

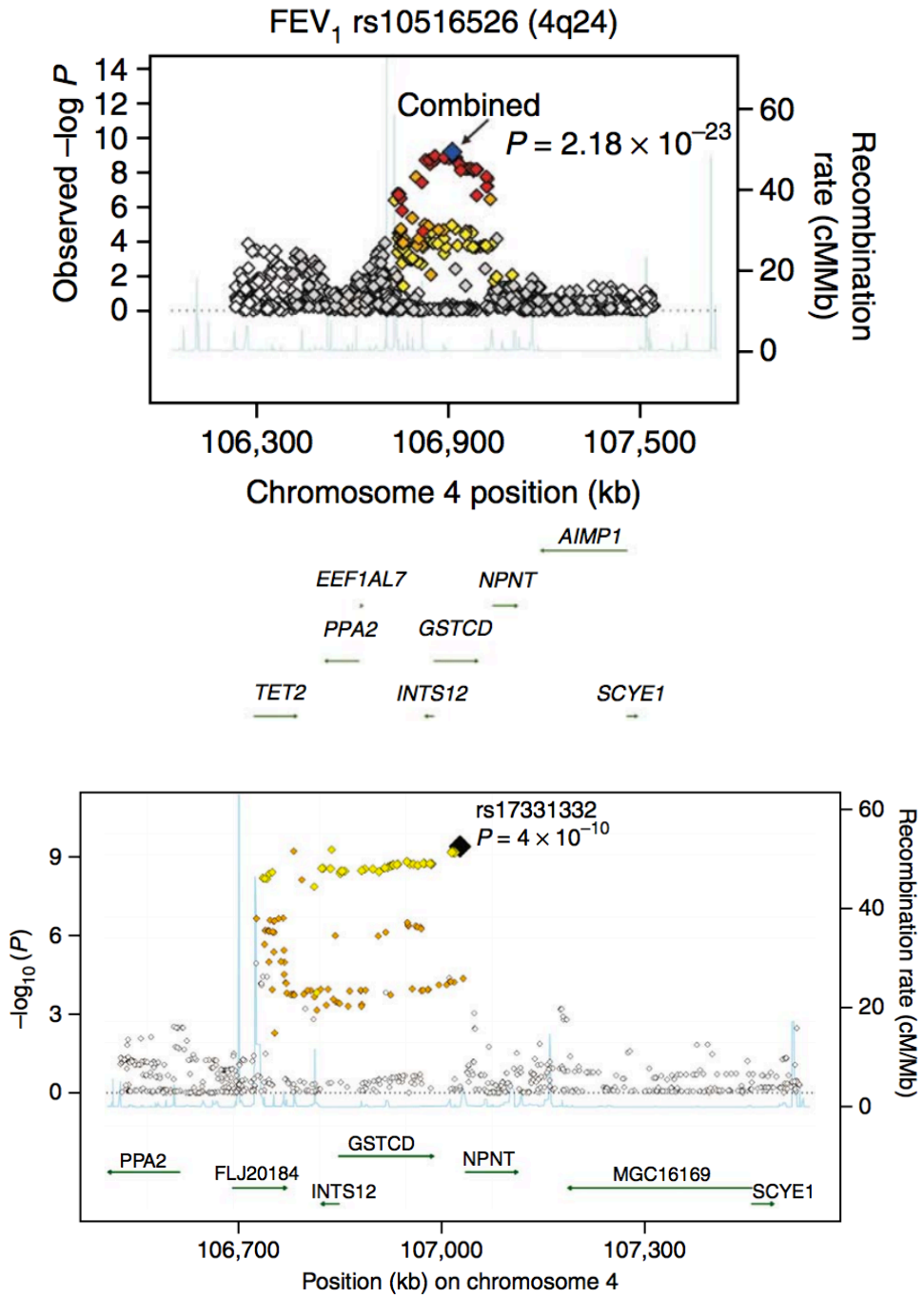


Figure 1.2: Regional association of FEV_1 at 4q24. Statistical significance of each SNP is shown on the $-\log_{10}$ scale as a function of chromosome position (NCBI build GRCh36). The sentinel SNP at each locus is shown in blue. The correlations (r^2) of each of the surrounding SNPs to the sentinel SNP are indicated by colours with red being $r^2>0.8$, orange being $r^2>0.5$, yellow $r^2>0.2$, grey being $r^2<0.2$, and white being unknown. The top plot is of the SpiroMeta consortium, while the bottom plot is of the Charge consortium. Reproduced from Repapi et al. and Hancock et al. (Repapi et al. 2010, Hancock et al. 2010).

Successful replication of 4q24 region association leveraging different population cohorts suggests that at least one gene within this locus is somehow connected with lung function and/or COPD pathogenesis. However genetic epidemiology approaches alone cannot isolate the causative gene. *INTS12* is a candidate lung function gene by virtue of being in strong LD with all sentinels SNPs leveraged by mentioned GWAS meta-analyses investigating the genetic basis of FEV_1. For example, in the case of SpiroMeta consortium, correlation coefficients (r^2) between rs10516526 and large subset of intronic and promoter SNPs for *INTS12* is greater than 0.8 (Figure 1.2).

1.4.4 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 such as COPDGene® (Regan et al. 2010), Evaluation Of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) (Villar Álvarez et al. 2008), and National Emphysema Treatment Trial (NETT) (Criner et al. 2011) 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 accurately 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. published findings using these cohorts (Cho et al. 2015). By completing a GWAS of CT imaging phenotypes, five genetic loci were found to be associated with emphysema-related phenotypes, one locus associated with airway-related phenotypes and two loci with gas trapping (Cho et al. 2015). The finding that genetic variants associated with both lung function and COPD risk also associate with emphysema is of critical importance as this not only provides greater confidence that this locus is a true association but also begins to help

dissect the altered biological mechanisms that may underlie the association.

1.4.5 Copy number variation in lung function and COPD

In addition to SNPs potentially contributing to human lung function and COPD susceptibility, copy number variation (e.g. duplication or deletion of regions of genomic DNA) is also an important area of study, with at least 4% of the genome harbouring copy number variants (Conrad et al. 2010). In 2011, Lee et al. performed a GWAS of copy number variation to test for associations with lung function measures in the Korean cohort (Lee et al. 2011). Interestingly, *TNS1* and *HTR4* showed evidence of correlation with FEV_1 and FVC when leveraging copy number variation. These genes have previously been identified in SNP association studies indicating that different kinds of genetic markers (e.g. point mutations vs duplications) can be in LD with the causative variant thus showing similar associations (Lee et al. 2011; Hancock et al. 2010; Repapi et al. 2010). Recent work in a European cohort however, did not support these copy number variation associations for lung function measures or COPD susceptibility (Wain et al. 2014).

1.4.6 The missing hereditability and the promise of whole genome sequencing associations

Using the approaches outlined above, to date, ~50 distinct lung function loci have been identified (Tang et al. 2014, Soler-Artigas et al. 2015, Repapi et al. 2010, Soler-Artigas, Loth et al. 2011, Hancock et al. 2010, Wain et al. 2015, Hancock et al. 2012). However, it is estimated that they explain only a modest proportion of the additive polygenic variance with 4% for FEV_1, 5.5% for FEV_1/FVC and 3.2% for FVC (Repapi et al. 2010; Hancock et al. 2010). There are several reasons for this gap also known as “missing hereditability”.

As mentioned, GWAS focus on common polymorphisms with minor allele frequency greater than 5% and spanning a small fraction of the human genome. However other forms of genetic variation may be important, particularly rare variation and structural variation (Lee et al.

2014). Therefore, common SNPs either individually or taken together typically may only explain only a small fraction of phenotypic variance. Leveraging larger population sizes could improve the determination of the true underlying genetic variance that accounts for phenotypic variance in lung function measures and COPD by identifying additional loci that contribute to these phenotypes.

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 variants, they found that heritability using SNP data are nearly identical to estimates based on pedigree information ranging from 0.50 for FEV_1 to 0.66 for FEV_1/FVC. However, based on to the formal definition of heritability (Wray and Visscher, 2008), it is possible to say that only variants in strong 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 proportion of to the phenotypic variation that is hereditary. More generally, whole genome sequencing associations are likely to improve our understanding of lung function and COPD genetics by capturing the entire genetic variation including not only point mutations but also copy number variants such as chromosomal translocations, deletions and insertions. Moreover, whole genome sequencing has the potential of being able to determine the actual haplotype block boundaries in the studied population without the need for imputation as it is currently implemented in GWAS. The decrease in the costs of DNA sequencing has made this a viable possibility (Kheirallah et al. 2016).

1.5 *In silico* approaches in translational studies

Despite GWAS successes in mapping lung function and COPD susceptibility loci, there is an obvious gap between these genetic findings and their functional and mechanistic translation (Visscher et al.

2012; Kheirallah et al. 2016). Over 90% of SNPs identified in GWAS of a range of human traits have been found to localize outside protein-coding regions and this has limited the rate of their functional translation (Maurano et al. 2012). This is also true for lung function and COPD associations. Therefore, it has been suggested that lung function and COPD associated variants are likely to be involved in normal and aberrant regulation of gene expression. Providing more general support for this, various GWAS SNPs were found to be enriched in chromatin regulatory features (Maurano et al. 2012) and over-represented in eQTL studies (Nicolae et al. 2010; Luo et al. 2015; Obeidat et al. 2015). Since gene expression signatures are cell type specific and dependent on developmental stage, epigenetic mechanisms, and 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 help in the generation of testable hypotheses.

1.5.1 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, 2007). 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 (Ohno, 1972), the project aimed at increasing our understanding of the remaining component of the genome which

traditionally was inadequately understood. 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, 2007), and currently is in the data production phase.

1.5.2 Integrating human genome regulatory information with candidate loci

The fundamental basis 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 et al. 2012). For instance, in the hypothetical locus displayed in Figure 1.3, 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 ChIPseq identified 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).

There are numerous regulatory features and patterns of gene regulation both of which are cell type specific and may vary at various stages of development. Below in sections 1.5.3 to 1.5.8, elements that can be considered for overlapping with GWAS loci and the possible underlying

biological mechanisms that may be responsible for the genetic association are summarised.

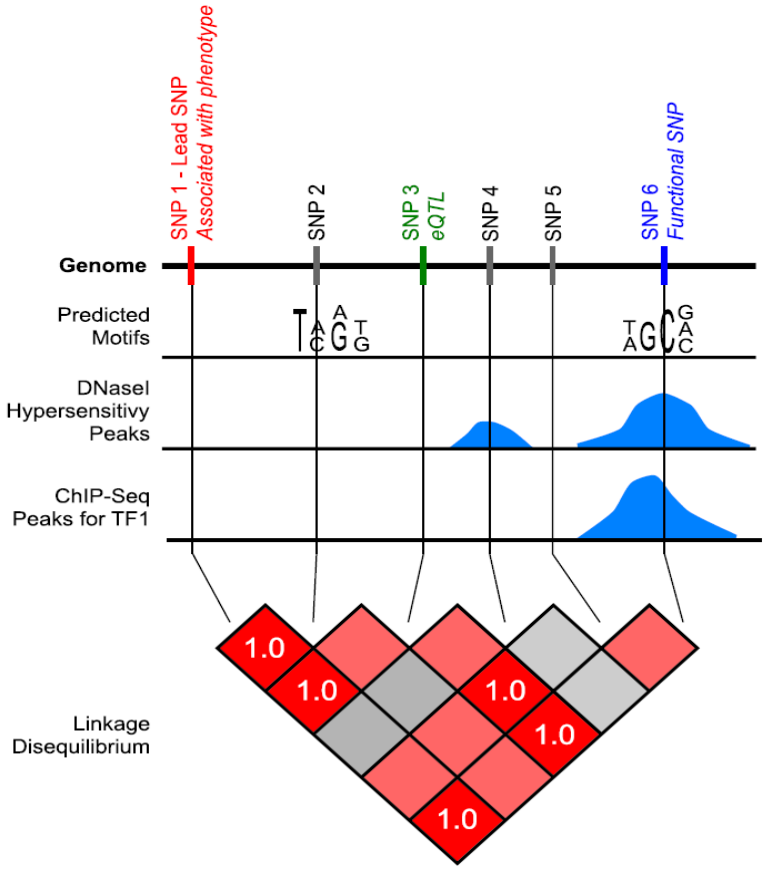


Figure 1.3: Combining GWAS-associated locus with human genome regulatory annotation. Reproduced from Schaub et al. (2012).

1.5.3 Transcription Factor 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 (Lee and 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 and Gann, 1997).

The current method of choice to identify TF binding sites is ChIPseq (Adli and Bernstein, 2011). In ChIPseq proteins are captured while attached to DNA by cross-linking with formaldehyde and the TF of interest is immunoprecipitated using a specific antibody. DNA is purified from the

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 referred to as 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 ChIPseq experiments however the current recommendation of the ENCODE consortium is to use 'input control' instead (Landt et al. 2012). Input control is a sequenced DNA without immunoprecipitation to account for local read distribution biases.

Demonstrating the specificity of antibody is pivotal and can be validated by either Western blotting (WB) or immunofluorescence (IF) combined with protein knockdown or knockout in the 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 et al. 2014) and HOMER (Heinz et al. 2010) toolkits. However, many more programs have been devised for ChIPseq analyses (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 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 1.4). 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.

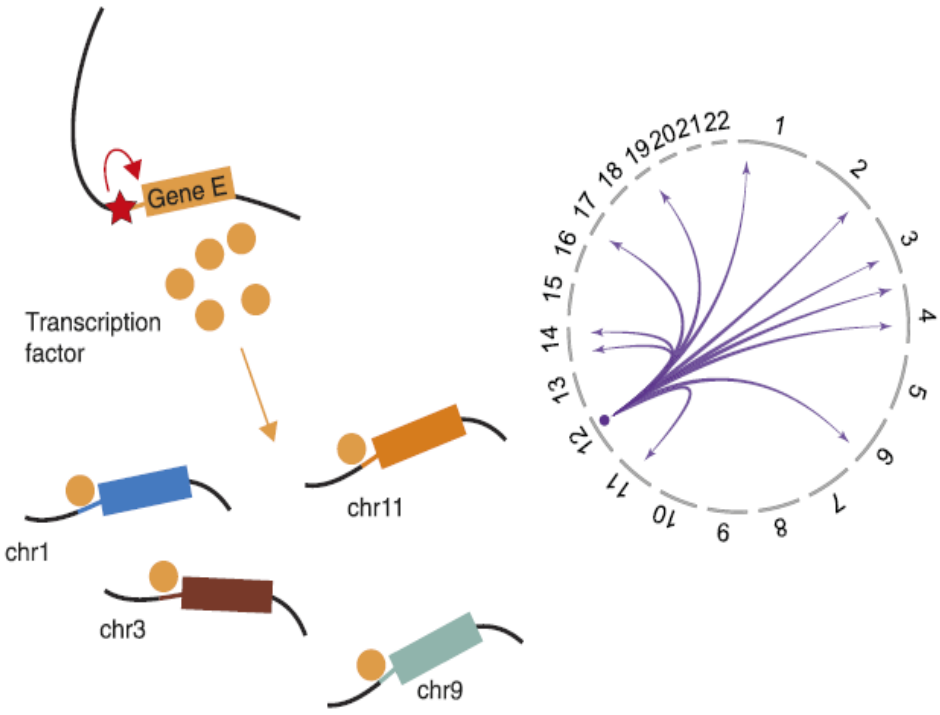


Figure 1.4: One possible mechanism driving a genetic association signal via TF activity. Reproduced from Knight (2014).

1.5.4 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 ChIPseq 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 and Kouzarides, 2011). Thus these modifications act in epigenetic control of gene expression and associate with different gene activities. For example, H3K4me3 mark tends to highlight actively transcribed loci while H3K27me3 associates with the silenced X-chromosome in females (Gibney and Nolan, 2010). Histone modifications are also used to identify the location of other functional elements such as enhancers (Shlyueva et al. 2014). However, the above list is not exhaustive of all possible histone modifications. As it is the case for other regulatory

elements, patterns of histone modifications vary depending on the cell type necessitating the use of datasets from tissues relevant for the phenotype of interest.

It is possible that patients with genetic susceptibility for COPD may have the predisposition for low lung function due to some developmental abnormality. Epigenetic mechanisms were shown to play a central role in embryological development and organogenesis (Kiefer, 2007). Therefore, aberrant resetting of epi-marks could be due to inappropriate levels of effector molecules responsible for their resetting. Because the majority of SNPs in GWAS studies are non-coding and abnormal 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.

1.5.5 Other regulatory elements

DHS 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 regions. DHS 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% is under purifying selection and hence likely to be directly functional (Rands et al. 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 (Li and 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 and 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 e.g. ribosomal RNA depletion protocol.

A useful software tool for rapid preliminary examinations of candidate GWAS loci is the Broad Institute's HaploReg (Ward and 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 project, the effect of SNPs on gene expression from eQTL studies, as well as the effect of SNPs upon putative regulatory motifs.

1.5.6 Overview of unbiased analyses of genomic feature overlaps

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 and are outlined below.

In Fisher's exact 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 (Fisher, 1945; Quinlan and Hall, 2010). On the other hand, the Jaccord statistic 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 (Favorov et al. 2012). The final statistic ranges from 0 to 1, where 0 represents no overlap and 1 represents complete overlap. Permutation-based approaches take reference and test regions as input and calculate 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 et al. 2015). Therefore, permutation-based approaches are based on simulation.

1.5.7 Expression quantitative trait loci approaches

The identification of SNPs as eQTL 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 tissues have been characterised for both mRNA expression and have been genotyped on GWAS 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 and Pastinen, 2011). Exon array datasets can be 'noisy' due to short probe design and probe hybridization signal saturation thus 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 and Pastinen, 2011). RNAseq provides a more accurate estimation of known or unknown transcript abundance and in a larger dynamic range (Wang et al. 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 1.5). 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 and can be used to prioritize SNPs or genes in GWAS loci (Rockman and Kruglyak, 2006).

Obeidat et al. 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 (Obeidat et al. 2015; Hao et al. 2012). 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 (Repapi et al. 2010; Hancock et al. 2010). The analyses identified a significant enrichment for both *cis* and *trans*-eQTL variants at those loci. More specifically, for the 6615 SNPs identified as associated with FEV_1, 3413 (i.e. 52%) were also *cis*-eQTL SNPs. Obeidat et al. study is an example of leveraging eQTL approaches to enhance the understanding of biology behind GWAS association signals (Obeidat et al. 2015).

Mapping the RNAseq reads to the reference genome followed by counting the number of SNP-specific reads allows for a detection of allele-specific expression (ASE). It is based on the ability to split the reads depending on the parental chromosome they align to (Figure 1.6). ASE detection is a unique feature of RNAseq and there are major technical challenges in the reliable measurement of ASE (Sun and Hu, 2013).

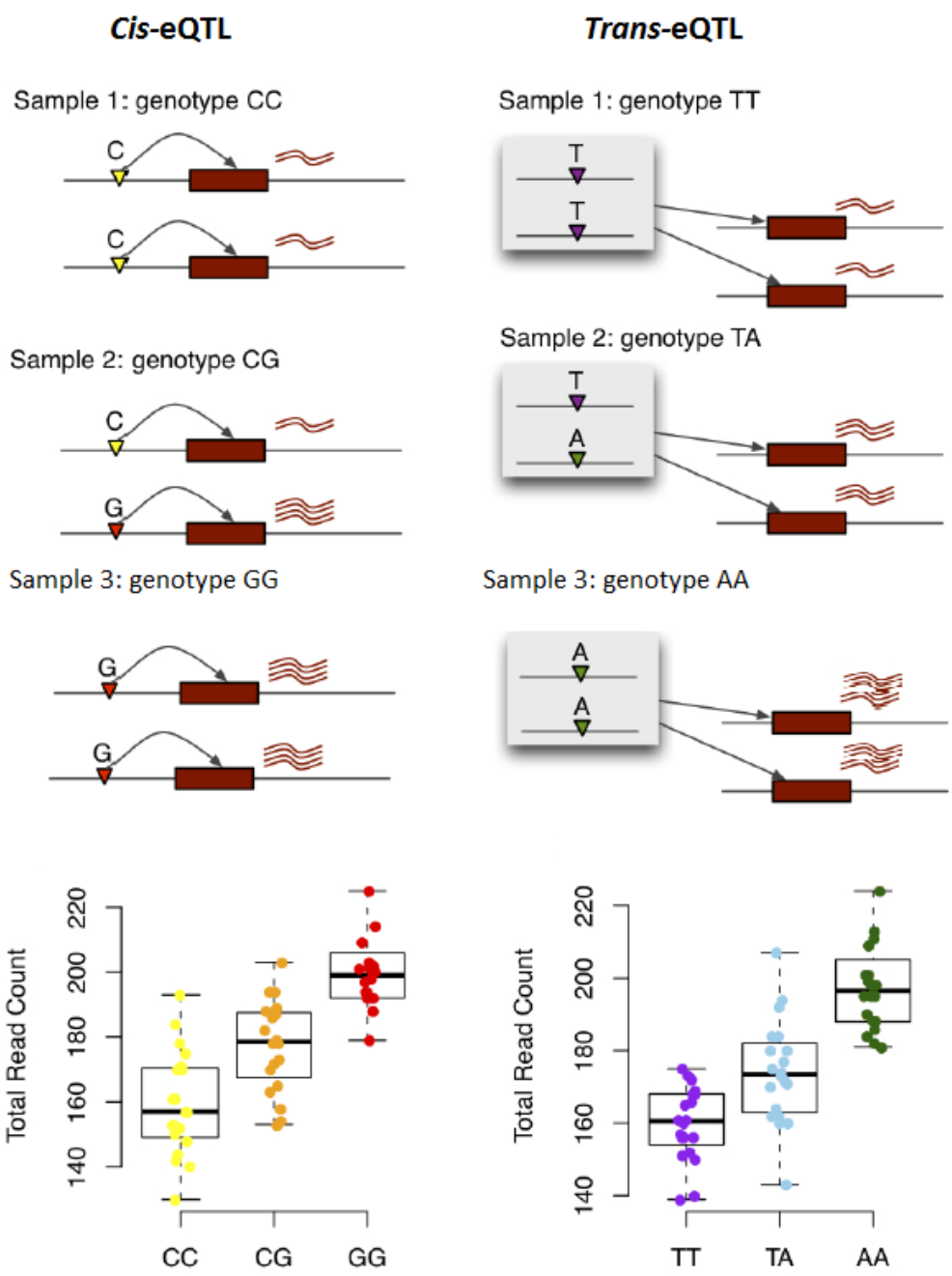


Figure 1.5: An illustrative hypothetical example of the cis-eQTL and trans-eQTL together with their associated per-genotype gene's read counts. On the left hand side we can see the example of cis-eQTL where allele C associate with low gene expression while allele G associates with high gene expression. A heterozygous individual with both alleles is 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. Reproduced from Sun and Hu 2013.

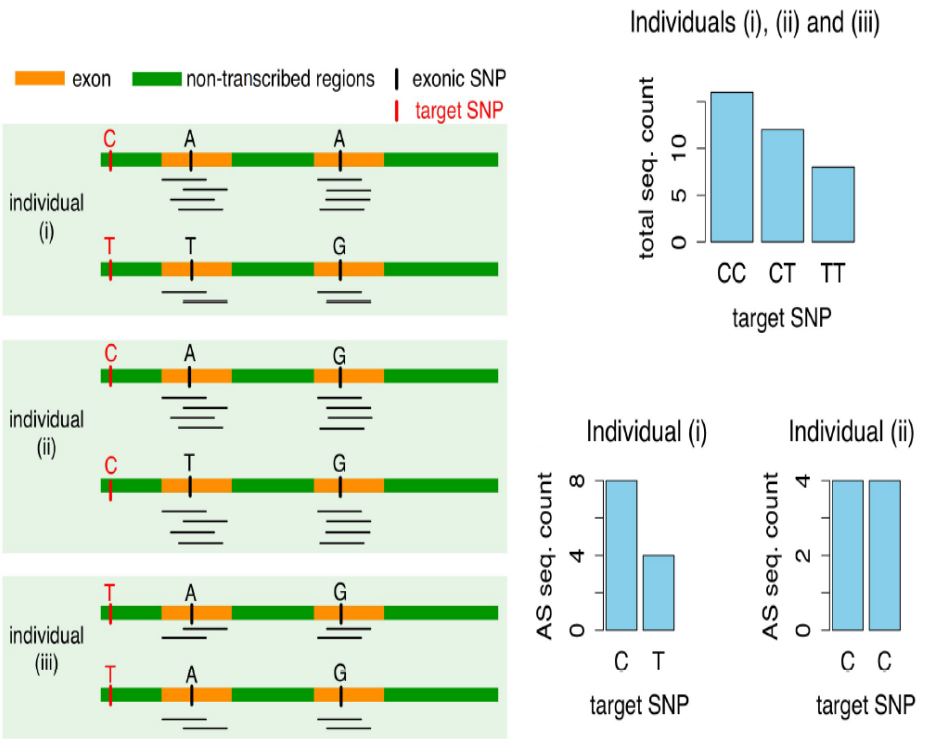


Figure 1.6: A practical example of ASE. Instead of counting total reads per gene, in determination of ASE, exonic (therefore sequenced) SNPs are imputed with not transcribed target (genotyped) SNP and read counts are performed per haplotype. Difference in read counts between the haplotype, as is the case for individual 1, is indicative of ASE. Reproduced from Sun and Hu 2013.

1.5.8 Summary of *in silico* approaches in the translational efforts

In silico approaches to facilitate the translation of genetic association analyses can be useful in providing 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 vitro* or, preferably, in *in vivo* models. Candidate regulatory variants, identified through overlap with publically available functional element annotations require experimental testing using a diverse range of methods in order to have confidence in the observed effects.

1.6 *In vivo* methods to translate GWAS findings

1.6.1 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 support *a priori* evidence about gene function and additional insight can be derived by comparing expression between healthy and disease states and during different developmental stages. IHC is a widely used tool to characterize protein expression in human tissues. In addition to providing information on protein localization within cells, 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 (Uhlen et al. 2005). In addition to identifying protein expression in normal human tissues it is important to consider whether protein expression changes in disease. The key questions are whether the protein expression increases or decreases and whether the change in expression can be used as a biological or prognostic marker.

There are a range of other models to study human lung tissue, the most obvious being primary cell culture which has widely been used to look at responses in airway structural cells. The limitation of 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 the human lung explant model and the human precision cut lung slice (PCLS) model (Wohlsen et al. 2003; Hackett et al. 2008). 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.

1.6.2 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 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.

1.6.2.1 Overview of lung organogenesis

Lung development has five *in utero* stages, with development continuing postnatally. 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 vessels 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 the existing alveoli. Lung volume continues to increase with skeletal growth, and reaches a maximum between 25 and 35 years of age (Figure 1.7; Moore and Persaud, 2003).

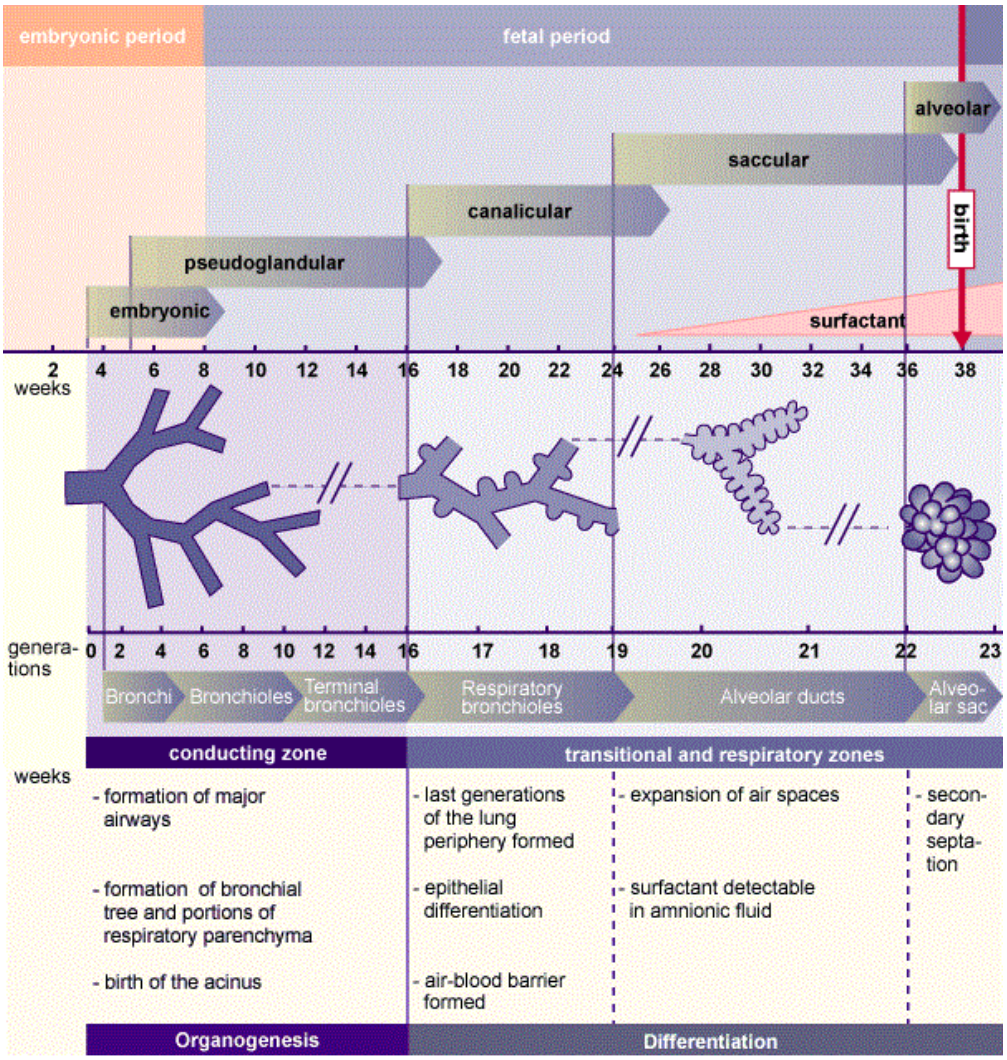


Figure 1.7: A graphical outline of human lung development.

Apart from genes with prior evidence for a role in human development (e.g. *PTCH1* and *HHIP*), little is known about the role and expression of lung function associated genes during lung development (Bellusci et al. 1997; Miller et al. 2004; Pepicelli et al. 1998). The gene expression omnibus is a publically available resource containing large microarray, and more recently RNAseq, datasets which can be used by the scientific community. The Human Developmental Biology Resource (HDBR) is an additional source of human embryonic and foetal tissue samples within the UK with samples ranging from 3 – 20 weeks post-conception

(http://www.hdbr.org/). HDBR can be used to assess lung function associated genes' protein IHC staining in a range of airway relevant tissues. As an example of using these resources, Hodge et al. and Obeidat et al. reported that whilst *INTS12* expression did not change throughout the development of the lungs, *GSTCD* expression significantly decreased and *HTR4* expression increased with rising foetal age throughout the pseudoglandular and canalicular stages (Hodge et al. 2013; Obeidat et al. 2013).

Although these findings are interesting, as it was mentioned before, a gene that is differentially expressed during lung development does not demonstrate this gene to play a direct role in lung development which may be showed by combining mutagenesis approaches altering candidate gene expression and assessment of lung development. Therefore, 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.

1.6.3 Mouse models for respiratory research and the translation of genetic findings

1.6.3.1 Complete gene knockout models

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 (Dawkins and 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 are essential. It is interesting to note that although the chromosomal make up in mice is different to humans with mice having 20 pairs of chromosomes rather than 23, 99% of mouse genes have human orthologues and the order of genes between the two organisms is the same. Many complex human diseases are shared in mice and humans, however drug development

using pre-clinical rodent models has been limited in translation success and this is particularly true in the respiratory field. Review by Edwards et al. focussed on asthma research and highlighted the potential over-reliance on animal models as a contributing factor to the lack of new drugs coming to the clinic (Edwards et al. 2015). Also, respiratory research in the mouse has its own specific considerations. For instance, the basic anatomy of mice's and human's lungs is different, with the make-up of the lung lobes and branching markedly different.

Despite these considerations it is beyond doubt that the use of mice in basic physiology research has provided dramatic advances in the understanding of the role of specific genes in mammalian physiology. In 2011, the approach of gene deletion to understand gene function 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 and Mouse Genetics Project, and has the vision to build a comprehensive catalogue of the functions of every gene in the mammalian genome (Brown and Moore, 2012). This is to be achieved by the generation and extensive phenotyping of ~20,000 knockout mice with removed protein coding genes in a systematic, standardized way. As the IMPC expands, more data will become available for GWAS relevant genes making this a 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 in individual laboratories. Critically, if global gene knockout is lethal, there is a possibility to delete genes in a tissue specific manner. This can be, for instance, achieved in doxycycline or tetracycline regulated systems (Gunther et al. 2002; Shockett et al. 1995).

1.6.3.2 Examples of gene knockout mouse models in respiratory research

Recent work has focused on gaining insight into whether candidate lung function gene *HTR4* plays a functional role in pulmonary physiology (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 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. This was observed with reduced IL1β responses in *HTR4* knockout following LPS instillation in the lungs. 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 (House et al. 2015).

GWAS have identified a number of polymorphisms in the Advanced Glycation End Product-Specific gene (*AGER*) with the pivotal nonsynonymous SNP appearing to be in exon 3 rs2070600 (Gly82Ser, (C/T)) which is associated with emphysema (Cho et al. 2015). *AGER* deficient mice when exposed to cigarette smoke appeared to be modestly protected from the emphysema like phenotype that developed in the lung including airspace enlargement when compared to wild type littermate controls (Sambamurthy et al. 2015). This protection was at least in part thought to be driven by a reduction in the influx of neutrophils into the airways in *AGER* knockout mice. Therefore, Sambamurthy et al. have taken a purely observational association study suggesting *AGER* to harbour a variation contributing to emphysema into a causal relationship demonstrating the power of this approach (Sambamurthy et al. 2015).

1.7 *In vitro* approaches using human cells to translate GWAS findings

This sections focus on the use of human *in vitro* models to further define and translate genetic association signals.

1.7.1 Choosing the cell type to work with

It is important to use cell types relevant to the phenotype of interest in order to avoid misleading biological interpretations. Currently available regulatory genome annotations have been generated in diverse sets of primary cells and immortalized cell lines (Consortium, 2007), 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 a potential, especially as cells can be derived from individuals of a known genotype.

Primary cells are most representative of the human tissues from which they were isolated. However, their phenotype is often context specific and when removed from the body may alter in phenotype. Therefore, 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. 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 flurescence-based preparative procedures.

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, BEAS2B-R1 cells, which are frequently used as a model by those interested in bronchial epithelial cell biology, have 68 chromosomes (unpublished observation; Ian Sayers et al.). These transformations may lead to artificial biochemical activities and misleading 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 approach can be used 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 using this method (Maurano et al. 2012). This method can in principle be applied to any phenotype for which a genetic association study has been undertaken.

1.7.2 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. Luciferase or green fluorescent protein (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 and Fried, 2007) are also used to screen nuclear extract or DNA sequences for specific protein-DNA binding activity.

1.7.3 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 (3C) which is used to analyse the organization of chromosomes. It utilizes oligonucleotide capture technology, 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. 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 by using a combination of chromosome conformation capture, ChIP - quantitative Polymerase Chain Reaction (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 (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.

1.7.4 Studying protein coding candidate genes

With an established candidate protein-coding gene, the traditional assay used for the characterization of the gene function is the gene knockdown using small interfering RNA (siRNA) or short hairpin RNA (shRNA) in a range of relevant human cell lines or primary cells. These methods are particularly useful when little is known about the gene function and may be used for hypothesis generation. Portelli et al. have overexpressed the

asthma associated gene, urokinase plasminogen activator receptor gene (*uPAR*) 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 (*uPAR* encoded protein) in the airway epithelium in asthma patients and association of PLAUR levels with worsening prognosis and increased disease aggressiveness in other diseases such as cancer and COPD (Stewart, Nijmeh et al. 2012; Ivancso et al. 2013; Smith and 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 implicated from GWAS approaches which can be used in cell biology experiments, although appropriate controls are essential as off target effects of transfection are possible (Echeverri et al. 2006).

1.7.4 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 had a relatively limited success in generating hypotheses about the biological basis behind considered phenotypes (Wang et al. 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 a disease state.

With a candidate gene 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 such as greater dynamic range, the possibility of novel splice variant discovery (Wang et al. 2014), identification of differentially expressed genes at individual isoform resolution, identification of differential splicing, identification of coding sequence differential expression and even differential promoter usage (Trapnell et al. 2012).

Having performed a differential gene expression analysis in the presence of approaches to target the gene of interest 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 (FDR) below 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 et al. 2009). The same principles apply for Gene Ontology (GO) analyses but instead of counting the number of genes per pathway, genes are counted per GO term.

It has been argued that the ORA approach is limited in its ability to identify biologically-meaningful pathways that vary between experimental conditions or phenotypes (Huang et al. 2009). 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 (e.g. genes that are differentially expressed at a $P<0.051$). 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 average gene expression in of-interest "reference" condition is subtracted from the average 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 (relative to the "reference" condition). 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 instance, if majority 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 GO 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 agree 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 has its advantages over choosing functional read-out assays on an arbitrary basis. With evidence of pathway or GO 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 1.8).

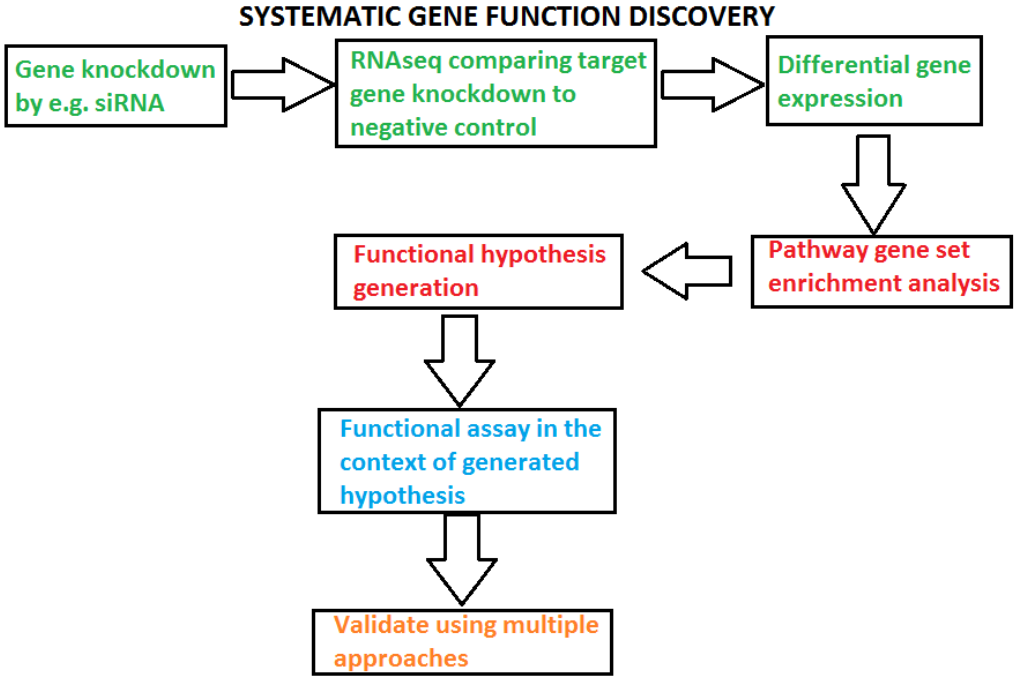


Figure 1.8: A workflow of systematic gene function discovery through combination of transcriptomics, pathway analysis, hypothesis generation and final biological validation.

As an example, 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 (Zhou et al. 2013). In this study, *HHIP* was targeted by shRNA in BEAS-2B airway epithelial cell lines followed by gene 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 was validated using additional qPCR in both BEAS-2B and primary human airway epithelial cells and shown to

be differentially expressed in COPD patients' lung samples versus non-disease controls (Zhou et al. 2013).

1.7.5 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 and 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 (Miller et al. 2011), and clustered regularly interspaced short palindromic repeats (CRISPR) in association with RNA-guided Cas9 nuclease (CRISPR-Cas9 system) (Sander and Joung, 2014).

TALENs are composed of a nuclease domain fused to a protein 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 TF 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 and Sander, 2013). CRISPR-Cas9 systems are alternative to TALENs and have improved cleavage efficiency and easier implementation at a 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 (Konermann et al. 2015; Gilbert et al. 2013; 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 (Ga et al. 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. Authors have combined TALEN domain, light-sensitive CRY2 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 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. This new technique offers great opportunity for biologists studying gene regulation and gene function in their genetic translation efforts, however further experiments are required to determine the specificity of the method and whether it can be used on a routine basis.

1.8 Inferred biology of reproducibly associated lung function genes

As already mentioned, *INTS12* is within a region with a statistically significant evidence of genetic association with lung function and COPD (see section 1.4.2). Crucially, expression of *INTS12* is high in the human bronchial epithelium (Obediat et al. 2013). The known *INTS12* genomic variation can alter the translated INTS12 protein length (Table 1.3). Although it is not entirely clear whether the potential contribution of INTS12 variation to pulmonary function is driven *via* protein coding or non-coding influences, the latter is being suggested to be the predominant case. Therefore, the leading hypothesis is that different *INTS12* expression levels may contribute to differences in lung function or lung pathology.

Transcript size (bp)	Protein size (aa)	Source	Known IDs
1975	462aa	Ensembl, NCBI	ENST00000451321 Variant 1
1927	462aa	Ensembl, NCBI	ENST00000394735 Variant 2
1710	462aa	Ensembl	ENST00000340139
1502	444aa	Ensembl	ENST00000618810
995	88aa	Ensembl	ENST00000416543
726	132aa	Ensembl	ENST00000420368
653	159aa	Ensembl	ENST00000503746
569	18aa	Ensembl	ENST00000515819
562	88aa	Ensembl	ENST00000433009
542	57aa	Ensembl	ENST00000510876
644	No protein	Ensembl	ENST00000493425
1927 – 1975	237	Obeidat et al. 2013	Variant 3

Table 1.3: Known INTS12 mRNA variants and their corresponding proteins. Information from National Centre for Biotechnology Information (NCBI), Ensembl database and Obeidat et al. 2013. The schematic of INTS12 protein can be seen in Figure 2.7.

This hypothesis is supported by an observation that there is a weak but statistically significant positive correlation between lung *INTS12* expression and FEV_1 (Percent Predicted) (Figure 1.9; Obeidat et al. 2013). Therefore, lower INTS12 levels in the lung associate with poorer lung function and vice versa. Moreover, in diverse tissue types, SNPs associated with lung function are *cis*-eQTLs for *INTS12* expression. The examination of *INTS12*/*GSTCD*/*NPNT* haplotype has detected eQTL effects for *INTS12* but neither for *GSTCD* nor *NPNT* in non-lung tissues such as liver, brain, and blood (Obeidat et al. 2013). In the lung, the eQTL effect had a greater effect size for *INTS12* expression than for *GSTCD* expression albeit not significant when correcting for multiple comparisons (Hao et al. 2012). More recently it was reported that in the lung the significant eQTL effect at *INTS12*/*GSTCD*/*NPNT* haplotype can be detected for *NPNT* expression which was overlooked in previous publication despite using the same publically available microarray dataset (Hao et al. 2012; Obeidat et al. 2015).

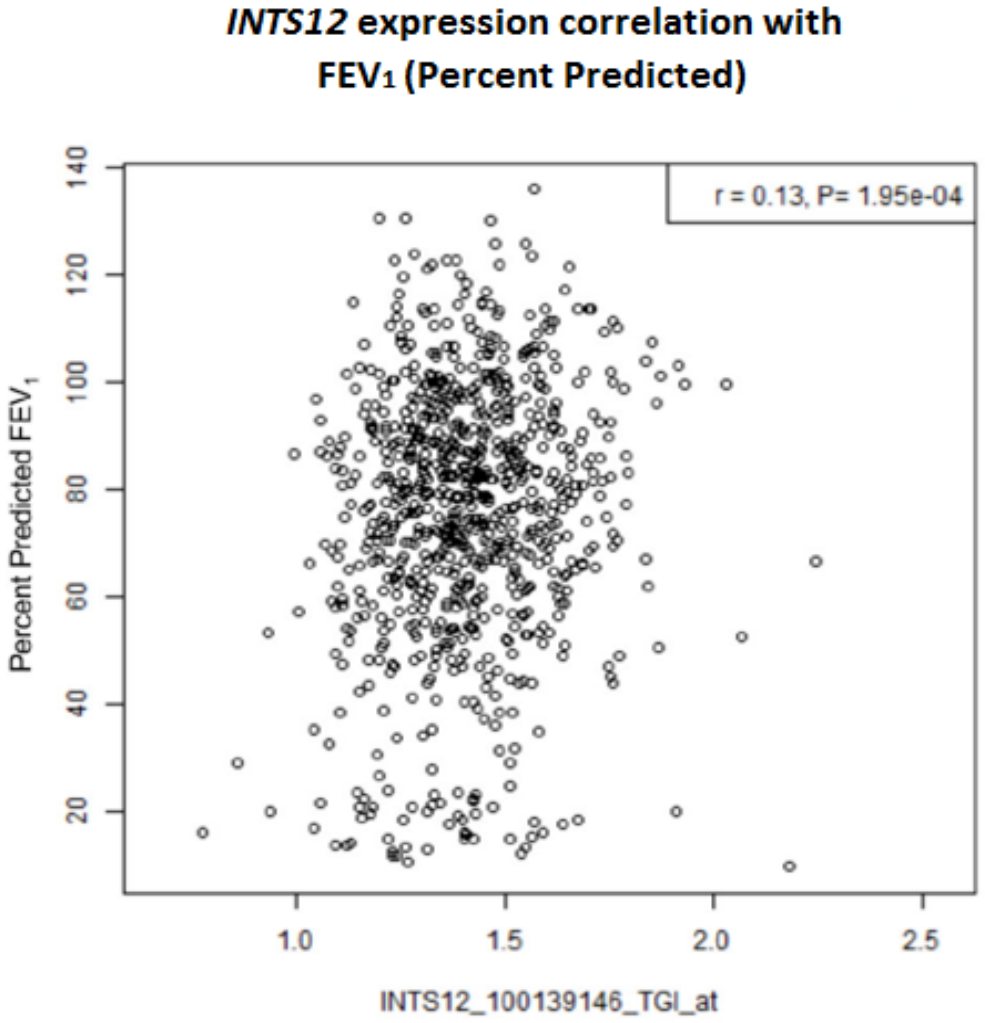


Figure 1.9: INTS12 mRNA levels positively correlate with FEV_1 (Percent Predicted). Reproduced from Obeidat et al. 2013.

However, relying on entire lung tissue to measure gene expression and detect eQTL is challenging due to heterogeneous nature of lung tissue. Therefore, although there may be an eQTL effect detected in a lung relevant cell type (e.g. airway smooth muscle cells) this effect is likely to be masked in a combined gene expression signatures of pooled cell types constituting lung tissue. For example, a study by Li et al. was able to prioritize genes at asthma susceptibility loci *via* eQTL approach by using human bronchial epithelial cells but often was unable to detect eQTL effects in bronchial epithelial lavage consisting of not only epithelial cells but a whole range of other cells (Li et al. 2015). Hence there are good scientific motives that argue for performing eQTL investigations in homogenous cell types.

Not only >90% of GWAS associations are non-coding (Maurano et al. 2012), but recently Obeidat et al. reported that eQTLs are significantly enriched at respiratory loci further giving credit to the idea that associations are largely driven by differences in gene expression (Obeidat et al. 2015). Interestingly, genes that were prioritized at respiratory loci using lung eQTL approach were tested via ORA method (see section 1.7.4) and were found to be enriched for developmental pathways (Obeidat et al. 2015). This goes hand-in-hand with what was said about significant GWAS signals for lung function being detected not only in adults but also in paediatric subjects. Also, the critical role of *INTS12* for mammalian development is demonstrated by the fact that homozygous *INTS12* knockout mouse models show pre-weaning lethality (Obeidat et al. 2013).

1.8.1 Integrator Complex and its subunit 12

INTS12 encodes Integrator Complex subunit 12, a protein that was initially discovered as the smallest member of INTScom (Baillat et al. 2005). Chen et al. identified Asunder and CG4785 as additional core members of the INTScom (Chen et al. 2012). Currently it is believed that INTScom has about ~14 subunits (Stadelmayer et al. 2014). Initially Baillat et al. were investigating the composition of deleted in split hand/split foot 1 (DSS1), the product of candidate gene for split hand/split foot syndrome (Baillat et al. 2005). To determine the identity of DSS1 they developed HeLa cell lines stably expressing a Flag-tagged DSS1. DSS1 was purified through anti-Flag affinity purification and the eluate was separated on a gel and each protein was analysed using mass spectrometry. Investigators found DSS1 to be a component of multiple distinct complexes and identified proteins corresponding to uncharacterized human open reading frames. Almost all DNA fragments predicted from these polypeptides were found to be conserved in metazoans. This approach identified 12 subunits of the INTScom, including INTS12, in physical association with RNA polymerase 2 (RNAPII) (Baillat et al. 2005).

INTS1, INTS3, INTS6, INTS7, INTS8, INTS11 and INTS12 were found to bind to RNAPII C-terminal domain (CTD). More recently, Stadelmayer et al. and Yamamoto et al. found negative elongation factor (NELF) to interact with INTScom proteins and this observation was not abolished after DNase and RNase treatment suggesting direct protein-protein interaction between INTScom and NELF (Stadelmayer et al. 2014, Yamamoto et al. 2014). Sequence investigation revealed INTS11 and INTS9 to have high sequence similarity to cleavage and polyadenylation specificity factor 73 (CPSF-73) and CPSF-100. CPSF-73 and CPSF-100 belong to the superfamily of zinc-dependent β-lactamases. CPSF-73 functions as the pre-mRNA 3'-end-processing endonuclease (Mandel et al. 2006). The same catalytic domain predicted to function as an RNA-specific endonuclease is present in INTS11 molecule (Baillat et al. 2005). The rest of Integrator members displayed little similarity with proteins involved in mRNA processing. The canonical function of INTScom is snRNA 3'end formation, also known as snRNA processing (Baillat et al. 2005; Ezzeddine et al. 2011; Chen et al 2012; Chen et al. 2013; see section 1.8.4).

1.8.2 Small nuclear RNAs

snRNA genes are part of the un-translated fraction of the human genome. Their mature transcripts are highly abundant, non-polyadenylated species and function within the cell nucleus (Matera et al. 2007). They are exported to the cytoplasm where they pair with proteins to form ribonucloproteins (RNP) (Egloff et al. 2008). With the exception of U7 snRNP, which plays a role in histone pre-mRNA processing, the rest of the snRNPs form the core of the spliceosome that removes intronic sequences from the pre-mRNA transcripts. snRNAs that we know about thus far are U1, U2, U4, U4atac, U5, U6, U6atac, U7, U11, and U12 (Matera et al. 2007). There are multiple gene copies of these snRNAs in the human genome presumed to have occurred through ancestral gene duplications. Some of these copies are thought to be transcriptionally silent (i.e. are pseudogenes).

snRNA genes have an snRNA encoding site (i.e. the site that is incorporated into mature snRNA), a TATA-less promoter containing distal sequence element (DSE) and snRNA-specific proximal sequence element (PSE), as well as an snRNA-specific 3'box located 9-19bp downstream of the cleavage site (Figure 1.10). High conservation of elements within snRNA promoters is a characteristic feature of snRNAs throughout the animal kingdom (Ezzaddine et al. 2011). The primary transcripts of snRNA genes extend beyond the 3'box element and sequences after the box display poor conservation in their paralogs even if occurring in the same species (Egloff et al. 2008).

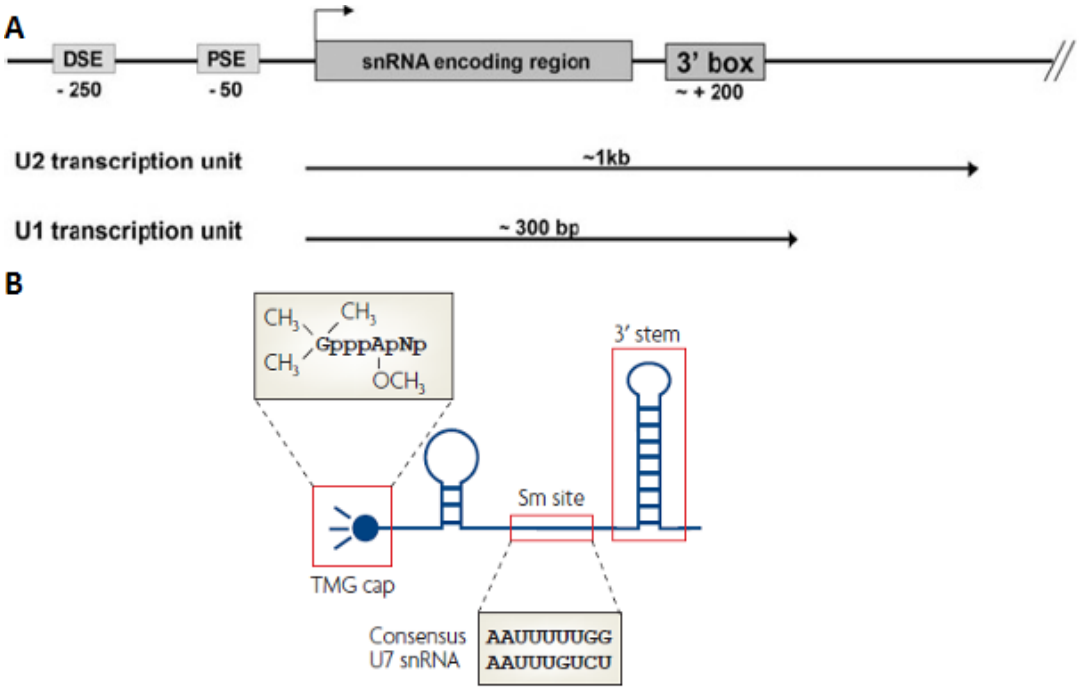


Figure 1.10: The architecture of snRNA genes (A) and the structure of processed U7 snRNA (B). Reproduced from Egloff et al. and Matera et al. (Egloff et al. 2008, Matera et al. 2007).

1.8.3 Functional role of INTScom in RNA polymerase II pause and release

Because INTScom was found stably accompanying RNAPII, Baillat et al. hypothesised that INTScom might mediate an RNAPII-dependent transcription (Baillat et al. 2005). However, their experiments have shown that depletion of INTScom subunits did not alter mRNA levels of protein coding proto-oncogene *FOS*. Furthermore, Baillat et al. were unable to detect INTScom subunits at its promoter using ChIP-PCR

(Baillat et al. 2005). This led researchers to investigate the potential role of INTScom in mediating transcription of non-protein coding genes. As snRNA processing is mediated through RNAPII's CTD (Hernandez 2001; Uguen, 2003), Baillat et al. wanted to test whether INTScom is recruited to snRNA genes (Baillat et al. 2005). Hence promoter and 3'box regions of U1 and U2 snRNAs were examined for the presence of INTScom subunits and RNAPII. ChIP of HeLa nuclear extracts using RNAPII and INTS10 antibodies followed by elution and PCR amplification of fragments corresponding to U1 or U2 resulted in equally strong positive signals around the promoter and 3'box of these snRNA genes. No signal was detected at histone H3 or GAPDH protein-coding genes.

These data have been superseded by more recent studies by Gardini et al. (Gardini et al. 2014). Gardini et al. found that some INTScom proteins are present near TSS of protein coding genes displaying RNAPII pausing. It is thought that ChIPseq data demonstrating higher RNAPII enrichment at the TSS relative to the gene bodies is largely due to pausing phenomenon where RNAPII awaits signals to begin active gene transcription. Prototypical examples of genes with this property are immediate early genes (IEGs) and therefore were tested for INTS11, INTS1, and INTS9 binding. Interestingly, Gardini et al. found these subunits near the TSS of some IEGs in contrast to what was previously reported by Baillat et al. (Gardini et al. 2014, Baillat et al. 2005). This lack of consistency was explained by saying that the antibodies that were used for ChIP in Baillat et al., although efficient in precipitating snRNA genes due to their multiple copies present in the human genome, were incapable of robust precipitation at non-repetitive protein coding genes (Gardini et al. 2014). Another plausible explanation is that single INTScom subunit may not be representative of the binding pattern of other subunits. In Baillat et al. binding near *FOS* was tested by using antibody against INTS10 whereas in Gardini et al. binding to the same gene was tested by using antibodies against INTS11, INTS1, and INTS9 (Gardini et al. 2014, Baillat et al. 2005). Therefore, comparison between these two studies is not straightforward and raises the possibility that

individual INTScom subunits display variable binding patterns arguing against a constant association between them on a genome-wide basis.
Epidermal growth factor (EGF), a potent stimulator of IEG response, was used to test the binding profile of INTS11, INTS1, and INTS9 and there was a robust increase in their occupancy near TSS and gene bodies as a result of EGF treatment of HeLa cells as shown by ChIP-PCR and exemplified by INTS11 ChIPseq. Occupancy decreased 40 to 60 min after EGF induction (Gardini et al. 2014).
Through microarray analysis validated by RNAseq, Gardini et al. have shown that RNAi depletion of INTS1 and INTS11 resulted in diminished EGF mediated response of IEG genes suggesting INTScom to be critical for transcriptional activation of these genes (Gardini et al. 2014). Then INTS11 depletion was combined with RNAPII ChIPseq and it demonstrated INTS11 requirement for polymerase pause release over EGF responsive genes. It was also shown that INTS11 depletion results in diminished 5'-end recruitment of key components of super elongation complex (SEC) onto three EGF responsive genes (Gardini et al. 2014).

Finally, some of these functional requirements were demonstrated in *D. melanogaster* S2 cells. The classical genes displaying RNAPII pause release phenomenon are heat shock response genes. Therefore, the localization of fly's INTS9 and INTS12 was tested over bodies of heat shock response gene *HSP70Aa* and their occupancy was increased near TSS and over the gene body following the heat shock treatment, implying some conservation of this function. However, more experiments are warranted to establish such conservation. Interestingly, despite increased INTS9 and INTS12 occupancy over *HSP70Aa*, the gene upregulation following the heat shock was attenuated (Gardini et al. 2014).
Concurrently to Gardini et al. study, Stadelmayer et al. also reported that INTScom subunits control RNAPII pause and release (Gardini et al. 2014, Stadelmayer et al. 2014). In contrast to Gardini et al. where this phenomenon was investigated in a model set of genes traditionally associated with RNAPII pausing, Stadelmayer et al. created a transgenic

cell line containing HIV-1 long terminal repeat (LTR) in conjunction with luciferase reporter gene. LTR leads to RNAPII pausing and premature termination after synthesis of short RNA, the TransActivation Response element (TAR). Genome occupancy over this inserted element was tested by ChIP-PCR and demonstrated RNAPII, INTS3, INTS11, and INTS3 enrichment over TAR, i.e. where rapid transcriptional termination and pausing typically occurs. An interesting observation is the opposite activity of INTS3 and INTS11 on RNAPII pausing. The presence of INTS3 correlates with low RNAPII density at the TSS and increased RNAPII in gene bodies however its knockdown reduces RNAPII occupancy over gene bodies. In contrast, INTS11 knockdown increases RNAPII occupancy over gene bodies but not over termination sites (Stadelmayer et al. 2014). Therefore, as far as INTS11 is concerned, Stadelmayer et al. observations agree with Gardini et al. observations at the TES but disagree at the gene bodies (Gardini et al. 2014, Stadelmayer et al. 2014).

1.8.4 Functional requirement for INTScom in snRNA biogenesis

As previously mentioned, the canonical function of INTScom is snRNA processing (Baillat et al. 2005; Ezzeddine et al. 2011; Chen et al. 2012; Chen et al. 2013). Baillat et al. suggested a possible mechanistic explanation of snRNA processing, fundamentally involving INTScom in the process. According to them INTScom “is loaded on the RNAPII at the promoter, traveling with it along the gene and cleaving the nascent primary transcript, most likely by recognizing the 3’box” (Figure 1.11; Baillat et al. 2005).

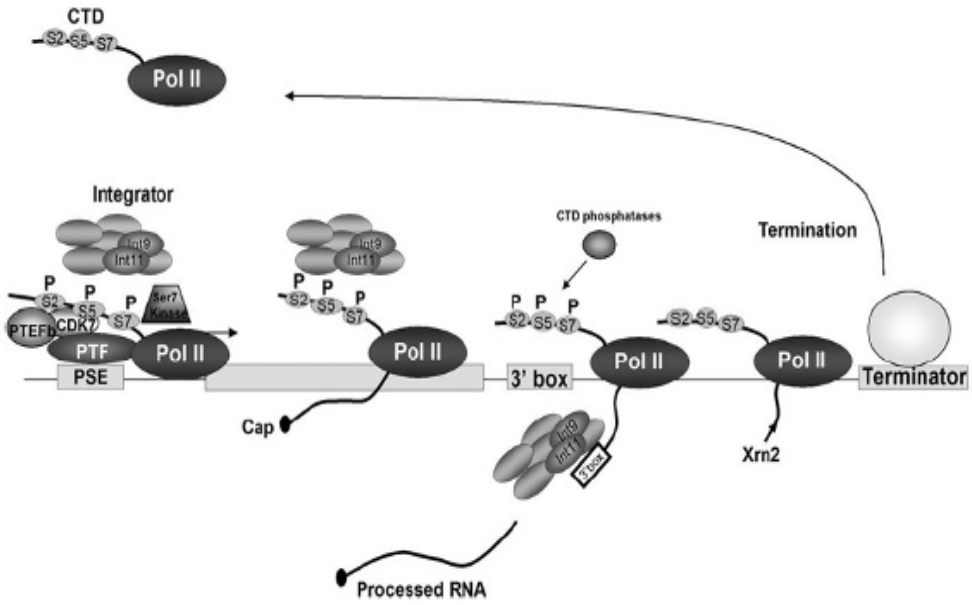


Figure 1.11: Baillat et al. hypothesis on snRNA biogenesis. It states that INTScom associates with CTD of RNAPII following its phosphorylation which initiates this association, recognizes 3'box on emerging pre-snRNA and INTS11 or INTS9 finally cleaves pre-snRNA yielding processed snRNA. Reproduced from Baillat et al. 2005.

To test this hypothesis Baillat et al. depleted INTS1 and INTS11 (Baillat et al. 2005). INTS11 was suspected to be the complex's cleaving engine due to its similarity to CPSF-73. Depletion of either INTS11 or INTS1 resulted in pronounced accumulation of the U1 and U2 misprocessed transcripts but no change in the GAPDH transcripts. To directly examine the catalytic activity of INTS11, researchers developed cell lines expressing a mutant INTS11 predicted to abrogate the catalytic activity of its β-lactamase domain that is thought to cleave primary snRNA transcripts. The mutated tag-INTS11 was validated for association with RNAPII and the rest of the INTScom subunits as well as with the U1, U2 promoters and 3'box. After the positive validation, mutated INTS11 overexpression resulted in a processing defect of U1 and U2 suggesting INTS11 as an snRNA cleaving component of INTScom (Baillat et al. 2005). However, INTS9 is equally likely to play such a role in INTScom activity as it shares sequence similarity to CPSF-73 and CPSF-100. In fact, INTS9 depletion also resulted in pronounced misprocessing of exogenous U7 snRNA (Ezzeddine et al. 2011).

1.8.4.1 Diversification of INTScom dependent functions via snRNA pathway

Given the fact that snRNAs form the core of the spliceosome complex and that INTScom is required for 3'end processing of snRNAs, it was anticipated that the knockdown of INTScom members may result in the disturbance of intron removal. Hence it is not surprising that the loss of INTScom resulted in reporting of diverse yet specific range of phenotypes. For instance, Otani et al. have shown that the expression levels of INTS6 and INTS11 were increased in preadipocytes in the period when the cells were differentiating into adipocytes, while they were reduced to basal levels after complete differentiation (Otani et al. 2013). Subsequently it was demonstrated that the knockdown of INTS6 and INTS11 results in the inhibition of differentiation into mature adipocytes. It was also shown that silencing of INTS4 leads to defects in the formation of Cajal bodies (Takata et al. 2012). It has also been suggested that the induced downregulation of INTS5, INTS9, and INTS11 in zebrafish causes impaired haematopoiesis due to aberrant splicing of smad1 and smad5 via a dominant negative form of these transcripts (Tao et al. 2009). Finally, various INTScom subunits were shown to be required for ciliogenesis (Jodoin, Shboul et al. 2013). It is thought that the primary mechanism behind these observations is the alteration of snRNA 3'-end formation affecting the splicing of mRNAs belonging to genes of particular functional groups explaining the specific phenotypic effects.

It is important to realize that the functional activities of INTScom were often inferred by a targeted depletion of single or a small number of its subunits which were considered to be representative of INTScom as a whole. It remains unclear whether all INTScom subunits are required for some of these processes, especially that there is variability in the relative contributions of various complex members to snRNA processing (Ezzeddine et al. 2011, Chen et al. 2012) and ciliogenesis (Jodoin, Shboul et al. 2013). It is possible that despite physical association of INTScom subunits, individually they may have distinct and different functions. Thus functional inference of the activities of single INTScom

subunits ought to be determined by their respective specific silencing. These functional activities may also vary between different species.

1.8.5 Functional roles for INTS12 in nuclear dynein dynamics

Jodoin, Sitaram et al. showed that not all INTScom subunits are required for perinuclear dynein stability (Jodoin, Sitaram et al. 2013). Researchers sought to determine whether Asunder, a key regulator of cytoplasmic dynein localization and a core member of INTScom identified by Chen et al. derives its separate functions from an independent or a common activity (Jodoin, Sitaram et al. 2013; Chen et al. 2013). By relying on RNAi approach they found that INTS12 knockdown in HeLa cells results in a decrease of nuclear-envelope-to-cytoplasm dynein ratio and an increase of the peak dynein on a nuclear envelope in ~80% of cells transfected with siRNA. The same was true for the majority of other INTScom proteins, except INTS7 and INTS10. Depletion of CPSF30 that is involved in polyA synthesis and 3'end formation of mRNAs, did not alter perinuclear dynein. Researchers thus concluded that the observed effect is not secondary to a general disruption of RNA processing but rather is specific to snRNA processing (Jodoin, Sitaram et al. 2013).

1.8.5.1 Subcellular localization and expression of INTS12

Another interesting aspect of work by Jodoin, Sitaram et al. is the investigation of INTScom members' subcellular localizations (Jodoin, Sitaram et al. 2013). Researchers fused GFP onto individual INTScom proteins to visualize their location. It turns out that based on their location INTScom subunits can be divided into three categories: predominantly cytoplasmic, predominantly nuclear and evenly distributed between nucleus and cytoplasm. INTS12 falls into the nuclear category. Interestingly, INTS12 was the only INTScom member that had an exclusively nuclear localization, i.e. 100% of cells had INTS12 present in the nucleus (Jodoin, Sitaram et al. 2013). The predominantly nuclear localization of INTS12 was also reported by Obeidat et al. (Obeidat et al. 2013). In this study immunohistochemical staining of normal and COPD adult lung tissue revealed INTS12 expression to be confined to the

nucleus of epithelial cells and pneumocytes. Furthermore, mRNA analyses found INTS12 to be expressed in a range of airway cell types, with the highest expression in human bronchial epithelial cells.

1.8.6 The known biology of Glutathione S-Transferase C-Terminal Domain Containing

Just like *INTS12*, *GSTCD* is in strong linkage with SNPs associated with lung function and COPD. Little is known about the molecular or cellular function of *GSTCD* gene. As the name implies, it has a GST motif which is present in other proteins as well. A homology search revealed that Eukaryotic Translation Elongation Factor 1, Titin and Chloride intracellular channel protein are the closest matching proteins. Interestingly, Titin proteins are important in striated muscle contraction, while Chloride channels have a role in number of respiratory conditions including Cystic Fibrosis (Obeidat et al. 2013). Importantly, in contrast to INTS12, although lung function SNPs are not eQTLs for GSTCD, this gene is differentially expressed between the pseudoglandular and canalicular stages of lung development. Considering the possible developmental basis of respiratory traits this makes GSTCD an interesting candidate to follow. Nevertheless, this thesis concentrated on INTS12 studies.

1.9 Introduction summary

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 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

Introduction but particularly in the 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 the careful experimental design using the most appropriate systems that is critical to interpretation. As genome editing technologies become routine, efficient and scalable these methodologies are going to play a pivotal role in the investigation of gene and single variants both *in vivo* and *in vitro*.

INTS12 is within 4q24 locus at the centre of reproducible association signal for lung function and encodes a protein that is a member of INTScom consisting of ~14 subunits. This complex was shown to stably accompany RNA polymerase II (POLII) and at a molecular level has been implicated in small nuclear RNA (snRNA) and Cajal body biogenesis, perinuclear dynein dynamics and recently with POLII pause and release. At the functional level, targeted knockdown and mutagenesis experiments demonstrated INTScom to be necessary for mouse adipogenesis, zebrafish haemopoiesis as well as human primary ciliogenesis. The relative contribution and requirement for each INTScom subunit in these processes is unclear at this time.

What is known directly about the function of INTS12 is that in *D.melanogaster* S2 cells it is necessary for snRNA processing and POLII pause release. In human cells, INTS12 was shown to be required for the maintenance of perinuclear dynein and primary ciliogenesis (Jodoin, Sitaram et al. 2013; Jodoin, Shboul et al. 2013), however these observations could not have been replicated by independent silencing experiments. Interestingly, Chen et al. demonstrated that in the fly, the evolutionary conserved plant homeodomain (PHD) motif of INTS12 is dispensable for snRNA processing while N-terminal subdomain is both necessary and sufficient for this processing to occur (Chen et al. 2013). This suggests the existence of important and unrealized functions for this gene which require further elucidation. More importantly, even though few functions have been identified for INTS12, no studies have addressed by which molecular mechanisms these functions are implemented.

1.10 Aims

The overarching aim of this thesis is to provide a greater insight on the function and genome-wide regulatory properties of INTS12 in primary human bronchial epithelial cells using some of methods outlined in this chapter. The key objectives of this thesis can be stated as follows:

1. To further test the hypothesis that the variable expression of INTS12 is a plausible driver of association signal for lung function at 4q24 locus using a lung specific eQTL dataset.
2. To develop the tools to silence INTS12 expression in human primary bronchial epithelial cells.
3. To design the necessary qPCR assays and evaluate the contribution of INTS12 to human snRNA processing.
4. To predict molecular and/or cellular functions of INTS12 through combination of gene knockdown and genome-wide RNAseq profiling (hypothesis-free study).
5. To test the predicted functions by using relevant assays.
6. To provide mechanistic insight into the observed differential gene expression changes by combining RNAseq and INTS12 chromatin immunoprecipitation followed by sequencing (ChIPseq) data.

2. Materials and methods

A range of high throughput and functional readout techniques were used as part of this thesis and they are described below. Issues in relation to data analyses and theoretical background of the methods are also considered.

2.1 Cell culture methods

Basal undifferentiated human bronchial epithelial cells (HBECs) obtained from three donors were purchased from Lonza© (Berkshire, UK) and used throughout this project (see Appendix for specification of donor demographics and experiments in which their respective cells were used). Cells were cultured in Bronchial Epithelial Cell Growth Medium (BEGM™) prepared by addition of bovine pituitary extract (0.4 % v/v), hydrocortisone (0.1 % v/v), hEGF (0.1 % v/v), epinephrine (0.1 % v/v), transferrin (0.1 % v/v), insulin (0.1 % v/v), retinoic acid (0.1 % v/v), triiodothyronine (0.1 % v/v), and GA-1000 (0.1 % v/v) to BEBM Basal Medium (500ml, Lonza©).

To ensure availability of the same passage of cells for further experiments, cells were batch frozen down. Thus passage 2 (P2) HBECs were grown adhered to 75cm^2 culture flask (T75) until ~95% confluent. Cells were then incubated in trypsin/EDTA at 37^{o}C for ~5min and re-suspended in a culture media (i.e. passaged). Finally, freezing mixture composed of 10% dimethyl sulfoxide (DMSO) BEGM was used to quickly transfer 250,000 of cells into a sterile cryovial which was stored at -80^{o}C for ~24 hours (h) followed by long term storage in liquid nitrogen. All subsequent experiments were performed by using these stocks of passage 3 (P3) HBECs. Cells were brought up from liquid nitrogen by thawing a cryovial at room temperature (RT) and pouring the entire vial content into 20ml of BEGM and growing cells on T75 for ~24h after which the growth medium was replaced. Prior to specific experiments cells were grown at 37^{o}C with 5% CO_2 until ~95% confluent with BEGM media change every 48h. All HBECs were characterized for expression of epithelial markers by either Lonza (Berkshire, UK) or internally (Stewart, Torr et al. 2012).

2.1.1 Haemocytometer counts

Haemocytometer-based cell counts were used to determine cell concentrations prior to cell seeding or to assess cell number in different experimental conditions. Concentration was defined as the number of cells per ml of cell suspension. Cells were diluted to a suitable suspension so that once in the chamber, the cells were uniformly distributed and not overlapping each other. A cover slip was cleaned and placed over the haemocytometer and 10µl of cell suspension was added to fill the chamber. The suspension fills the chamber by capillary action. The grid is divided into 9 squares, each with a surface area of $1mm^2$. The depth of the chamber is 0.1mm giving a total volume of $0.1mm^3$ for each square. Cells were counted in the two squares and averaged, giving a number of cells in $0.1mm^3$ volume. This number was then multiplied by 10^4 to give a final cell count per ml.

2.2 RNA interference

2.2.1 RNA interference and off target effects

RNA interference (RNAi) is extensively used in functional gene studies. RNAi is a gene silencing system where small interfering RNAs (siRNA) and microRNAs (miRNAs) play a central stage (Jackson and Linsley, 2010). miRNAs are endogenous non-coding RNA used by eukaryotic cells for post-transcriptional regulation of gene expression. These RNAs are transcribed and processed in the nucleus by enzymes Drosha and Pasha. After the export to the cytoplasm pro-miRNAs are processed by enzyme Dicer yielding double stranded ~21-22 base pairs (bp) long mature miRNA. One of the strands is called the guide strand and this strand is incorporated onto RNA-induced silencing complex (RISC) to guide silencing complex to target mRNA to promote translation arrest or mRNA cleavage (Bartel et al. 2009). siRNAs on the other hand are exogenous and in nature they are introduced into eukaryotic cells by an invading virus as Dicer substrate RNA (D-siRNA). This D-siRNA is enzymatically cleaved by RNase-III class endoribonuclease Dicer, generating ~21-22bp long siRNA (Jackson and Linsley, 2010). As it is

the case for miRNAs, siRNA's guide strand binds to RISC and ensures that activated catalytic component of RISC, Argonate, cleaves the mRNA complementary to the guide strand. The passenger strand is degraded upon incorporation of guide strand onto RISC.

Therefore, siRNAs and miRNAs share similar machinery downstream of their initial processing. In addition to their specific silencing property conferred via sequence complementarity, siRNAs may generate off-target effects (Jackson et al. 2003). miRNA-like off-targeting is driven by partial complementarity to other mRNA sequences and this limitation is inherent to any siRNA used in knockdown experiments. Other off-target effects include oligonucleotide or delivery vehicles disruptions resulting in innate immune responses. These limitations of the RNAi mediated knockdown approach makes it imperative to carefully design functional experiments looking at the effect of gene silencing on the phenotype.

RNAi techniques are used to selectively knockdown specific mRNA transcripts within a cell in order to investigate the function of the protein or, in case of non-coding genes, the function of the transcript. Commonly employed techniques are transfections of either siRNAs or Dicer substrate siRNA (D-siRNAs; Amarzguioui et al. 2006). siRNAs are chemically synthesised 21bp long RNAs and therefore bypass the need for enzymatic processing by mimicking Dicer products. However, in this project, 25bp long D-siRNA molecules were used for gene knockdown in which 25bp substrates are processed by Dicer into 21bp long siRNA followed by their incorporation into RISC complex. Dicer-Substrate duplexes provide two critical improvements over the use of traditional 21bp siRNA designs. D-siRNA takes advantage of the natural processing by Dicer producing 10-fold higher potency and specificity than the shorter 21bp siRNA forms. D-siRNA duplexes also evade the mammalian interferon response when expressed in mammalian cells (Amarzguioui et al. 2006).

2.2.2 D-siRNA transfections experimental optimizations

2.2.2.1 Determination of FuGENE6® and INTERFERin® transfection efficiencies

The first step in any RNAi gene silencing experiment is the optimization of siRNA or D-siRNA delivery into the cells. Therefore fluorescent Cy3 labelled scrambled siRNA (excitation max 556nm, emission max 570nm; Ambion Life Technologies, cat. num. AM4621) as well as FuGENE6® (Promega cat. num. E2691) and INTERFERin® (Polyplus cat. nu. 406-10) transfection reagents were purchased and used in order to compare their transfection efficiencies. Cells were imaged using epifluorescent microscopy (see section 2.4.1). INTERFERin® is a cationic polymer that binds to anionic D-siRNAs creating a complex that interacts with negatively charged heparan sulfate proteoglycans on the outer cell membrane and is introduced to cytoplasm by endocytosis with the actin involvement (Jelena Vjetrovic, personal communication, 17th April 2014). FuGENE6® transfects DNA or RNA in a process called lipofection whereby cationic lipids bind to the negative molecules and fuse with the cell membrane or undergo endocytosis (Jacobsen et al. 2004). Concentration of the transfection reagents was 2µl/ml while Cy3-D-siRNA concentrations where 10nM, 50nM, and 100nM. These concentrations were chosen because the cell florescence is hardly visible below 10nM even with excellent transfection efficiency and at this stage the objective was to compare FuGENE6® and INTERFERin® transfection efficiencies. The experiment was performed in three independent biological replicates (i.e. using different cell vials of the same donor and at different times).

2.2.2.2 Validation of RNAi functionality and prioritization of D-siRNAs targeting INTS12

Having determined the desired transfection reagent and conditions, the next step was to demonstrate RNAi functionality using a known effective D-siRNA. Hypoxanthine Phosphoribosyltransferase 1 (HPRT1) positive control duplex transfected as per manufacturer's recommendation was used for this purpose (OriGene cat. num. SR302223A). Following

successful suppression of HPRT1, three different INTS12 targeting D-siRNAs (OriGene cat. num. SR311359; Table 1 of Appendix) were tested for specific gene knockdown. D-siRNAs A, B and C were found to be complementary to exons 6, 7-8, and 8 respectively when aligned against canonical mRNA variant 1 (NM_020395.3; Figure 2.1). INTS12 D-siRNA were transfected at 10nM and INTS12 expression was tested by PCR (see section 2.3.5) 48h after the initiation of knockdown. On the basis of obtained results, D-siRNAs A and C were taken forward for further optimizations. Because off-target effects were shown to be increased at the higher D-siRNA doses (Jackson et al. 2003), knockdown efficiency was tested at 0.1nM, 1nM and 10nM D-siRNA concentrations in order to determine the lowest concentration yielding the desired level of INTS12 knockdown (again 48h after the initial transfection). On the basis of obtained results, 1nM D-siRNA concentration using 1μl/ml of transfection reagent were chosen for subsequent functional experiments to reduce the severity of off target effects.

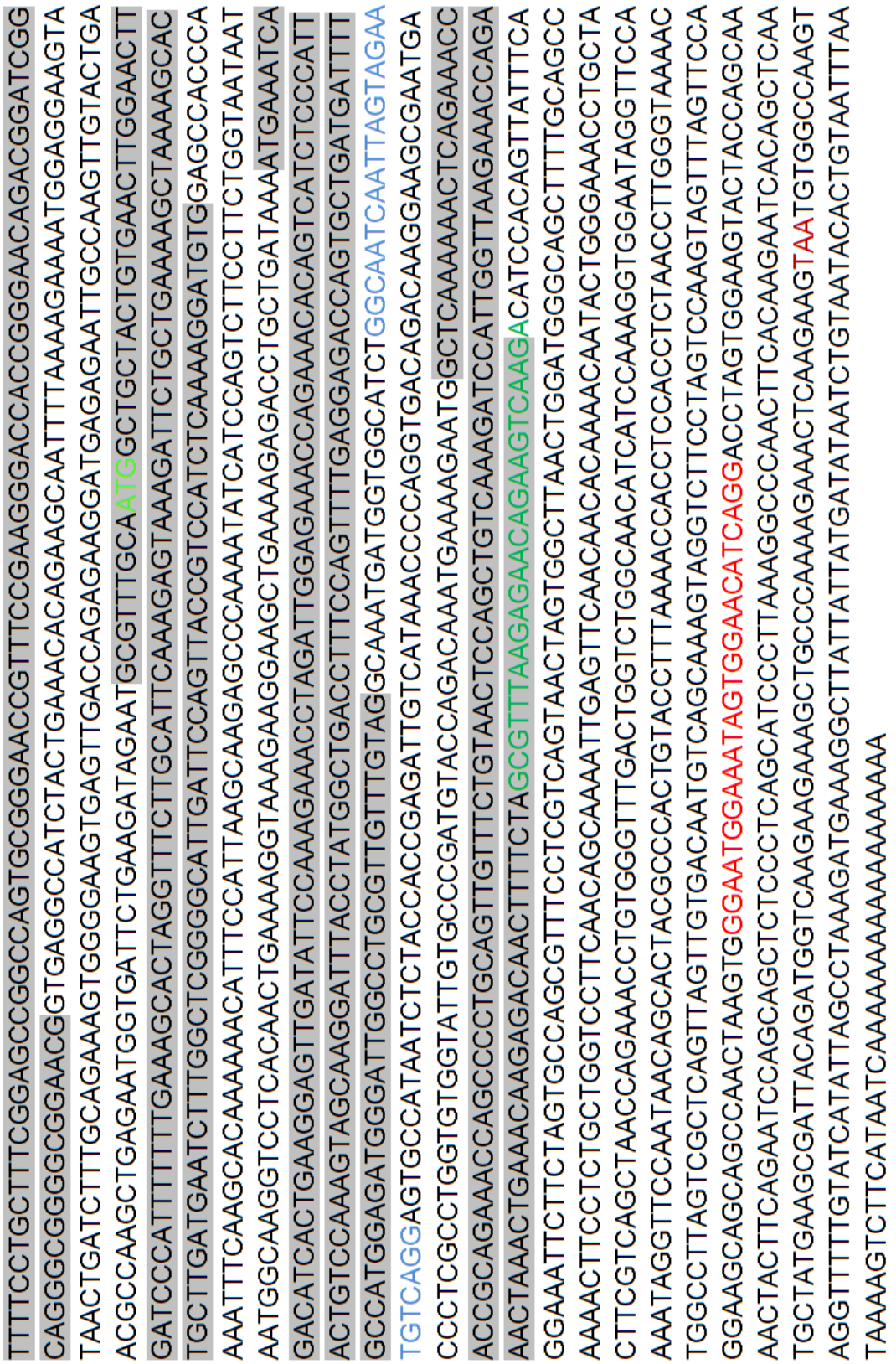


Figure 2.1: INTS12 variant 1 mRNA sequence (NM_020395.3) showing exon-exon arrangement, start site (green ATG), stop site (red TAA) and coloured complementary D-siRNA targeting sites (A, B, C).

2.2.2.3 Final optimized gene knockdown protocols

The effects of INTS12 knockdown were examined 48h and 120h post initiation of RNAi in order to test the acute and sustained effects of silencing on cellular function and gene expression. In the case of 48h protocol, cells were transfected at day 0 and lysed at day 2, while in the 120h protocol cells were transfected at day 0 and day 3 with final lysis and/or functional readouts performed at day 5. In all the experiments cells were transfected with 1nM of D-siRNA A and C as well as scrambled non-specific D-siRNA control (see section 2.2.2.2). In addition, some cells were left un-transfected. Transfections were performed with INTERFERin© at 1μl/ml.

2.3 Fundamental molecular biology methods

2.3.1 Gel electrophoresis

Electrophoresis is the movement of charged molecules in an electric field used to separate DNA or RNA molecules based on their molecular size. The electrophoretic mobility depends on a number of variables among which the most important are net molecular charge, size and shape of the molecules, strength of the electrical field as well as density of sieve-like matrix. Anions, i.e. negatively charged molecules, move towards positive anode while cations, i.e. positively charged molecules, migrate towards negative cathode. Highly charged molecules move faster than those with lesser charge. Smaller molecules migrate faster than large molecules due to frictional resistance of the matrix. Shape of the molecule also affects its migratory rate, e.g. linear DNA has a lesser mobility than circular DNA. Mobility also increases with the increasing field voltage but there are practical limitations in using high electrical field strength because of heating effects. DNA size standard ladder is typically run in parallel to other samples to roughly estimate the size of the separated molecules. For visualization purposes, gels contain a DNA and RNA intercalating ethidium bromide (EB) which highlights their location when exposed to ultraviolet light (UV) (Lodish et al. 2004).

Depending on the expected molecular weight of electrophoresed molecules, electrophoresis was carried out in 1 to 2% (w/v) agarose gel. Every agarose gel contained 5 x 10^{-3} % (v/v) of stock EB (Gibco, cat. num. 15585-011) resulting in its final concentration of 0.5µg/ml. In order to be able to load the samples and visualize the progress of electrophoresis, 6 x orange-g dye was mixed with samples in a 1:6 ratio. The final volume of the loaded sample was 15µl. Electrophoresis was run between 50mA and 150mA and stopped when orange-g dye migrated across ~80% of gel. Gel was exposed to UV light using Gel Doc system (Syngene).

2.3.2 RNA extraction and deoxyribonuclease I treatment

Total RNA was extracted from cells using GenEluteTM Mammalian Total RNA Miniprep Kit (Sigma-Aldrich© cat. num. RTN70). First, cells were lysed and homogenized in a lysis buffer containing guanidine thiocyanate and β-mercaptoethanol (β-ME) to ensure thorough denaturation of macromolecules and inactivation of RNases. Throughout this project this initial step was performed by adding 300µl of lysis buffer per each well of 6-well plate and lysates were pooled from two wells into single 1.5ml Eppendorf tube representing a particular experimental condition. Lysates were stored for a minimum of 24h before proceeding to the next step of the procedure. After thawing, the lysates were spun through a filtration column for 2min at 13200 (13.2k) rotation per minute (rpm) which removes the major contaminants as well as cellular debris and equal volume of 70% ethanol was added to the filtrate. The addition of ethanol causes RNA to bind when the lysate is spun through a silica membrane column in a microcentrifuge tube. Contaminants were washed away during a series of washes and then silica membrane column was treated with 1 in 8 diluted neat deoxyribonuclease I (DNaseI) (Sigma-Aldrich©, conc. 1unit/µl, cat. num. DNASE70) in order to aid the removal of genomic DNA (gDNA). Finally, RNA was eluted in 50µl elution solution. RNA yield was determined on NanoDrop 2000 UV-Vis Spectrophotometer (Thermo-Scientific©). Prior to total RNA extraction

the workspace and pipettes were thoroughly cleaned with RNase Away solution (Invitrogen©, cat. num. R60001).

2.3.3 Quality control of total RNA

For critical experiments RNA integrity was assessed on Agilent 2100 bioanalyzer using RNA LabChip® kit. Briefly, this instrument allows electrophoretic separation of RNA macromolecules on microfabricated chips. After electrophoresis the RNA fragments are detected via laser induced fluorescence detection, generating an electropherogram and a gel-like image. The process of determining RNA integrity is algorithmically standardized in order to remove individual interpretation in RNA quality control (QC). Software algorithm takes into account the entire electrophoretic trace into account and classifies eukaryotic total RNA into RNA Integrity Number (RIN) based on a numbering system from 1 to 10, with 1 being the most degraded profile and 10 being the most intact. For some experiments RNA QC was performed by running RNA through 1.5% (w/v) agarose gel at 55-100mA (see section 2.3.1 for more details on gel electrophoresis). The 2:1 ratio of 28S to 18S ribosomal RNA (rRNA) bands and lack of smearing was considered to represent intact RNA.

2.3.4 Complementary DNA synthesis by reverse transcription

Principally, total RNA was converted into complementary DNA (cDNA) to measure gene expression by leveraging a viral reverse transcriptase (RT) enzymatic reaction. For this purpose, the SuperScript™ First-Strand Synthesis System for RT-PCR kit (Invitrogen, cat. num.11904-018) was used. Prior to the reverse transcription reaction, RNA was treated with DNaseI (Invitrogen, cat. num. 18068-015) for a second time to ensure complete removal of any remaining traces of gDNA. DNaseI digestion reaction was performed for each RNA sample in 0.5ml Eppendrof tube in a final volume of 10µl containing 1µl of digestion buffer, 1µl of DNaseI, and various quantities of RNase/DNase-free water (Sigma, cat. num. 95284), and RNA samples depending on RNA concentration. Generally, it was ensured to have 1µg of total RNA per

RT positive (RT+) reaction and equivalent amount per RT negative (RT-) reaction. All the samples had the same amount of RNA in order to normalize differences in gene expression driven by differences in RNA yield. DNaseI digestion was carried out at RT for 15min. DNase I was inactivated by the addition of 1µl of 25 mM ethylenediaminetetraacetic acid (EDTA) followed by incubation at 65^{o}C for 10min. cDNA synthesis reaction was initiated by using random hexamers priming, and therefore the generated cDNA represents the total RNA content rather than mRNA pool only, as it is the case for polyT oligo priming. In addition to 1µl of random hexamers the final reaction mix contained 1µl of 10mM deoxynucleotides mix (dNTP), 2µl of 10X RT reaction buffer, 4µl of 25mM $MgCl_2$, 2µl of 0.1M dithiothreitol (DTT) and 1µl of ribonucleaseOUT (RNaseOUT). The RT+ samples contained 1µl of RT enzyme and 2.5µl of RNase/DNase-free water while RT- samples contained 3.5µl of RNase/DNase-free water. Prior to the addition of RT enzyme, samples were equilibrated on thermocycler at 25^{o}C for 10min. cDNA synthesis reaction was started at 25^{o}C for 10min followed by 42^{o}C for 50min, and 70^{o}C for 15min. Finally, 1 µL of RNase H was added and incubated for 20min at 37^{o}C. Samples were stored at -20^{o}C.

2.3.5 Quantitative real time and end point polymerase chain reactions

Polymerase chain reaction (PCR) is a molecular method whereby a fragment of DNA can be amplified more than trillion fold. In the end point PCR this amplification is performed using a pre-determined number of cycles and the final PCR product is visualized by gel electrophoresis. The size of this product can be estimated based on the relative-to-ladder location of the DNA band and sequenced following purification from the gel. It may be possible to compare quantities of the amplified fragment by end point PCR but this is not recommended because after certain number of cycles the quantity of DNA reaches plateaux and therefore is not comparable between the conditions. This issue is addressed by the real time PCR where the quantity of the amplified product is monitored

in a real time. Therefore, the differences in the starting DNA material can be quantified based on the rates of the amplifications (Bustin et al. 2009).

2.3.5.1 Principles of polymerase chain reaction

The basic set up of PCR includes DNA template containing region to amplify, two primers that are complementary to the 3' ends of each of the sense and anti-sense strand of the DNA target, *Taq* polymerase capable to withstand temperature as high as 98^{o}C, deoxynucleoside triphosphates (dNTPs) which are the building-blocks from which the DNA polymerase synthesizes a new DNA strand, bivalent cations such as Mg^{2+} or Mn^{2+}, monovalent K^{+} and buffer providing a suitable environment for the amplification. PCR consists of series of repeated temperature changes whereby DNA copying can occur.

There are three main steps in the PCR reaction: denaturation, annealing and extension. In the denaturation step the DNA is heated to 94 – 98^{o}C causing DNA melting by disrupting hydrogen bonds resulting in the separation of complementary strands into single strands. During the annealing step the temperature is lowered to 50 – 65^{o}C in order to allow the forward and reverse primers to hybridize to the appropriate DNA templates. PCR assay specificity is conferred thanks to primers' complementarity to target sequences. Finally, extension occurs at 72 - 80^{o}C during which polymerase adds dNTPs complementary to the template in 5' to 3' direction. These steps are repeated in a cyclical fashion allowing coping of specific section of DNA (Figure 2.2). PCR works not only with gDNA but equally well with cDNA.

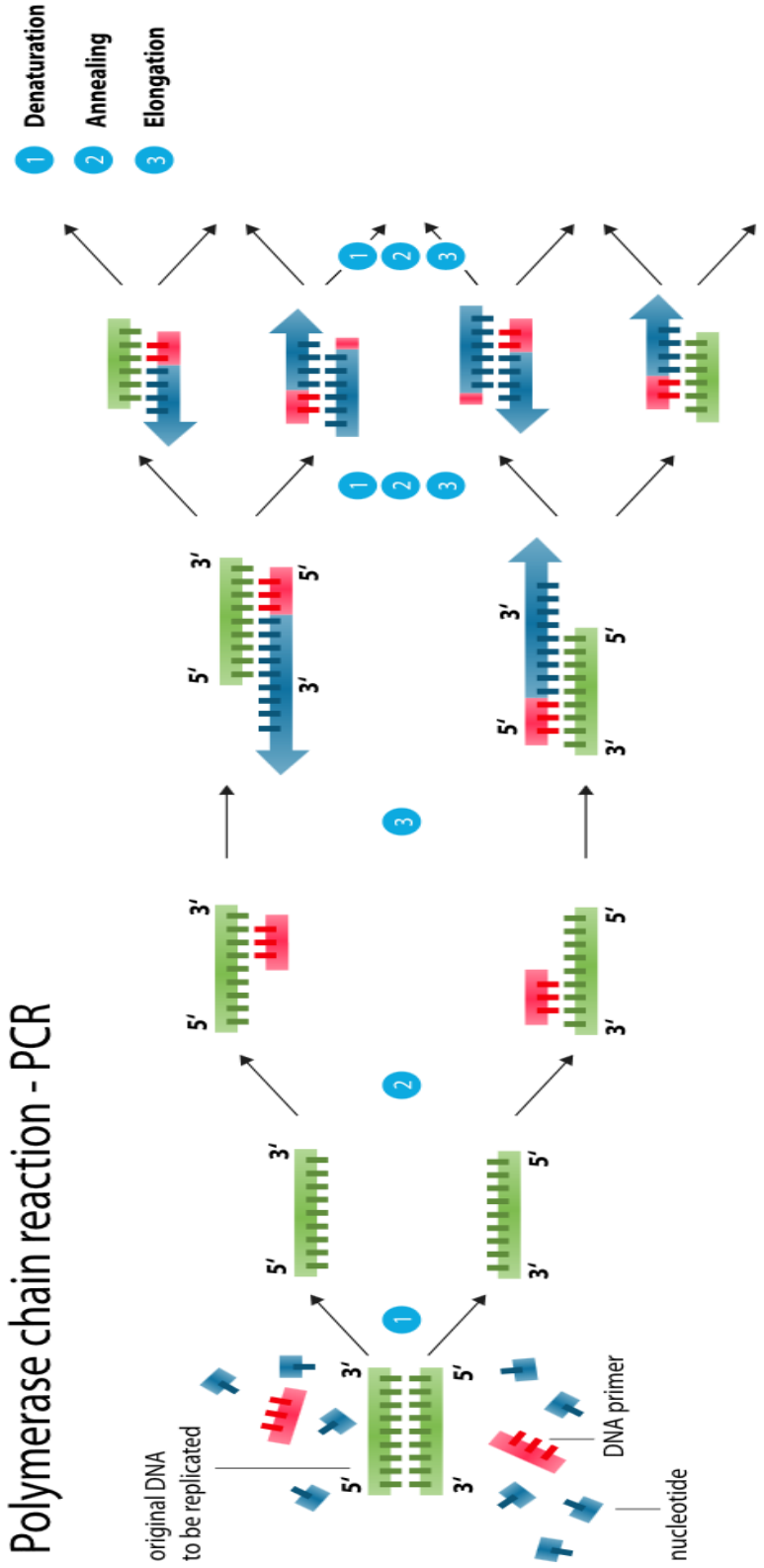


Figure 2.2: PCR is possible by a repeated process of denaturation, annealing and elongation yielding millions of copies of amplified DNA fragment.

2.3.5.2 Chemistries of real time polymerase chain reactions

Real time PCR is referred to as quantitative PCR (qPCR) because it allows for either absolute or comparative quantification of starting DNA material. There are two main qPCR chemistries that make continuous monitoring of produced DNA copies possible, each with its advantages and disadvantages: the DNA binding dyes and probe-based designs.
DNA binding dyes such as SYBR® Green have a very low level of fluorescence when unbound which increases by 1000 fold after binding to the double stranded DNA (dsDNA). As PCR amplification increases the quantity of dsDNA, the fluorescence signal increases proportionally (Figure 2.3). The advantage of this approach is the ease of implementation, whereas the main disadvantage is the fact that reaction specificity is determined solely by the utilized primers, as fluorescence intensity increases regardless of the sequence identity of amplified DNA. Thus primers should be designed to avoid non-specific binding as much as it is possible.

This limitation of SYBR® Green qPCR assay is addressed by running a dissociation curve at the end of PCR run. The purpose of the dissociation curve is to determine if anything other than the gene of interest was amplified in the qPCR reaction. In the dissociation curve analysis amplicon is subject to an increase in temperature peaking at 90oC, with fluorescence measurement taken throughout. As SYBR® Green dye binds exclusively to dsDNA, continuous dissociation of the two DNA strands will result in a decrease of the fluorescence. A plot of the negative first derivative of the fluorescence versus temperature displays distinct peak corresponding to the melting temperature (Tm) of DNA product or multiple peaks if multiple PCR products of varying length were generated. A single peak provides the evidence for qPCR reaction specificity. Therefore, the underlying assumption of this method is that different PCR products have different rates of dissociation from SYBR® Green during a temperature increase.

On the other hand, in the case of probe-based assay, such as TaqMan, in addition to the primers, a probe annealing between forward and

reverse primers is designed. This probe is conjugated with fluorochrome plus a quencher. When the probe is intact, the fluorochromic wavelengths are not released because quencher prevents this from occurring. However, when probe is degraded by DNA polymerase as it moves along the DNA template synthesising the daughter strand, the fluorescent signal can be detected (Figure 2.3). Hence, the main advantage of TaqMan qPCR is increased level of specificity, provided by not just primers but also by probe. Also, multiple probes can be labelled with different reporters allowing for parallel profiling of multiple targets in a single PCR reaction (multiplex qPCR). The disadvantage of probe-based approach is the need for probe design and lower time and cost efficiencies when compared to SYBR® Green qPCR.

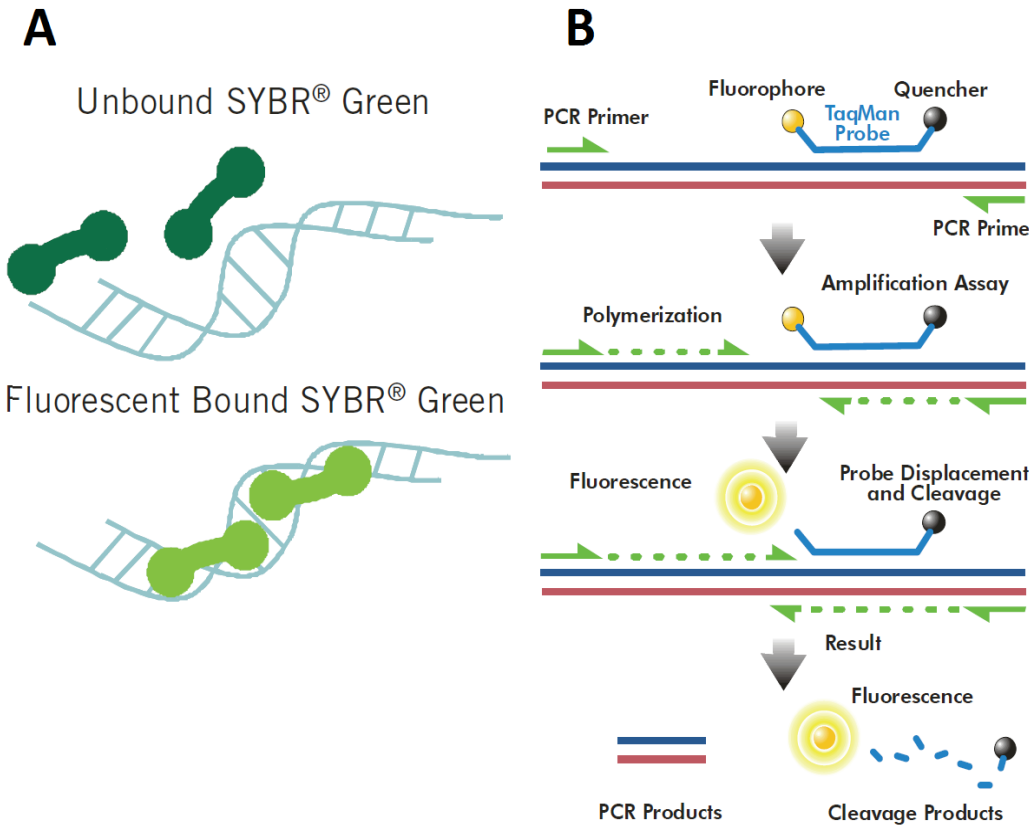


Figure 2.3: In SYBR® Green the dsDNA binding dye, binds to the DNA as amplification progresses increasing the fluorescence signal (A), while in TaqMan qPCR fluorescence is produced after DNA polymerase degrades a probe with attached fluorochrome and quencher (B).

2.3.5.3 qPCR data analysis

In the process of qPCR analysis, a fluorescence threshold is established in order to compare the PCR cycle numbers (Ct) between the samples. Throughout this thesis the qPCR thresholds were initially established either algorithmically by the analysis software (Stratagene®) or manually, and kept the same in any subsequent experiments for consistency. The fundamental idea behind the analysis is that samples with lower quantity of starting target material would reach the threshold after more cycles then the samples with higher quantity and thus will have lower Ct values (Figure 2.4).

There are two main qPCR data analysis approaches: the absolute quantification and the relative quantification. In case of absolute quantification, the absolute quantity of the target can be quantified by using a calibration curve. Calibration curve can be generated by using samples of known quantities of target cDNA and measuring their equivalent C_t values. The relationship of template quantity to C_t values is linear and the best fit curve can then be used to determine the quantity of target cDNA in unknown samples. However, the more popular method for estimating differential gene expression is the relative quantification method also known as $\Delta\Delta C_t$ method (Livak, 2001). Briefly, the C_t value of a housekeeping gene ($C_{t\ housekeeper}$) is subtracted from the C_t value of target gene ($C_{t\ target}$) yielding ΔC_t for each considered sample ($\Delta C_{t\ sample\ A} = C_{t\ target\ sample\ A} - C_{t\ housekeeper\ sample\ A}$). Then one of the experimental conditions is considered a control with an assumed average expression of 1, where the rest of the samples are compared to this control. To do this the average of ΔC_t values of the control samples is calculated ($\Delta C_{t\ average\ control}$) first. Then $\Delta\Delta C_t$ value for each individual sample is calculated by subtracting $\Delta C_{t\ average\ control}$ from $\Delta C_{t\ sample\ A}$ ($\Delta\Delta C_{t\ sample\ A} = \Delta C_{t\ sample\ A} - \Delta C_{t\ average\ control}$). The relative expression is finally calculated as $2^{-\Delta\Delta Ct\ sample\ A}$. Thus according to this formula the two underlying assumptions are that the PCR assay efficiency is ~100% and that the housekeeper expression is relatively constant across the experimental conditions. It is important to test these assumptions because if not

fulfilled may result in flawed results. Housekeeper normalization is particularly important in *in vivo* animal model experiments where gene expression is compared between different animals as RNA input normalization for cDNA synthesis reaction is not enough to compare differences in gene expression. Because $\Delta C_{t\ average\ control}$ is subtracted from each control ΔC_t values the $\Delta\Delta C_t$ for controls is ~0 and therefore relative expression $2^{-\Delta\Delta Ct}$ is ~1 while relative expressions of the rest of samples represent fold changes versus control.

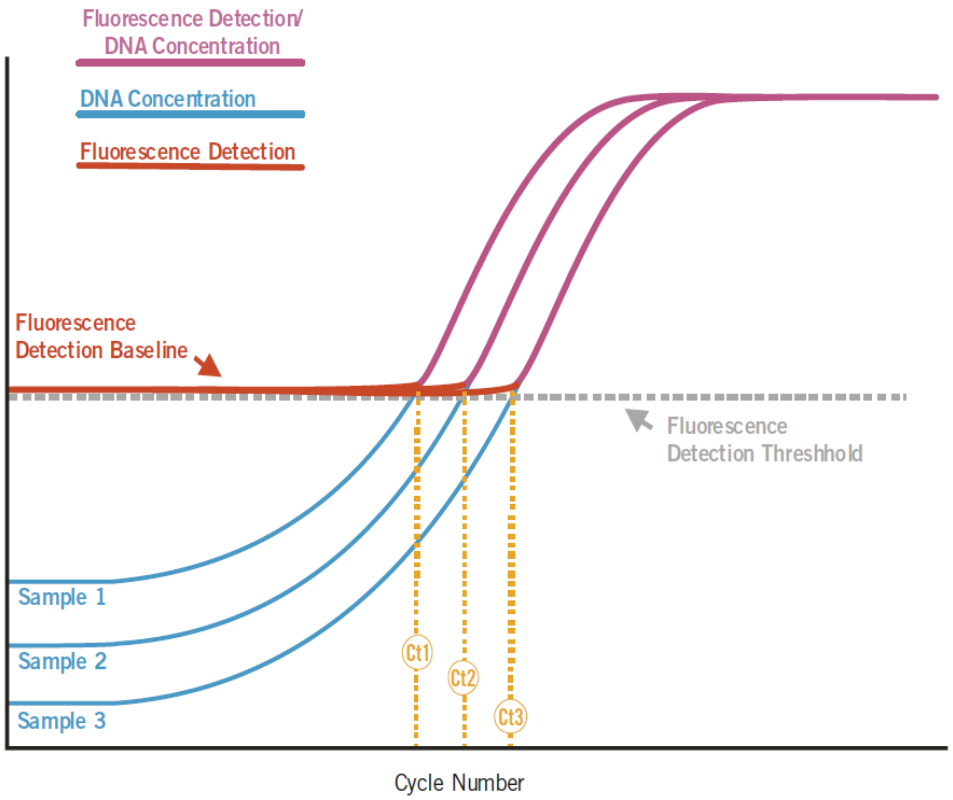


Figure 2.4: In qPCR the sample with a lower quantity of starting template (sample 3) reaches the threshold after more cycles than the sample of higher quantity of starting template (sample 1).

2.3.5.4 Design and validation of SYBR® Green and TaqMan qPCR assays

Both SYBR® Green and TaqMan qPCR assays were used in this project. Some assays were designed in-house and thus were tested and validated for ~100% amplification efficiency. Housekeeping gene assays were pre-designed and commercially obtained TaqMan oligos (Life Technologies; Table 2 of Appendix). Some SYBR® Green assays were also pre-designed and commercially obtained (Sigma-Aldrich; Table 3 of Appendix).

When designing the primers, the following rules (Dieffenbach et al. 1993) were followed where possible to produce optimum qPCR primers and probes:

- T_m should be 55-80^{o}C. T_m was calculated according to the approximation formula $T_{m=}4(G+C) + 2(A+T)$.
- The annealing temperature should be approx. 5^{o}C less than the lowest T_m of the primer pair.
- Each primer should be between 17 - 28bp long.
- The content of G and C bases should be between 35 - 60%.
- The 3' end of the primer should end with G, C, CG or GC to increase priming efficiency.
- Runs of more than 3bp of C or G at the 3' end of the primer should be avoided as this could result in formation of primer dimers.
- Self-complementarity score should be less than 7 as higher scores may result in secondary structures such as hairpins.

The designed primers were checked on Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) to check for primer length, T_m, proportion of GC content and self-complementarity indices. NCBI Primer-Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) was used to test primers specificity by blasting against NCBI Transcript Reference Sequences.

In this thesis the relative qPCR quantification method was used in determining gene expression. Assay efficiency was calculated using a calibration curve analysis on Stratagene® analysis software where qPCR measurement was performed over 2-fold dilution series cDNA template, starting from 1μg of neat cDNA obtained from un-treated P3 HBECs obtained from donors 195307 and 7F3206 (D195307 and 7F3206 respectively). Because in an ideal PCR reaction there is a perfect doubling of target amplicon every cycle, an assay efficiency of 100% means that there is roughly 1 C_t difference between samples of the 2-fold dilution series. Formally, the program performs a simple regression to find the slope of line relating cDNA quantitates to the C_t

values. Because of this relationship the slope can be directly used in assay efficiency calculation leveraging 2-fold dilution series as follows:

$$\% \text{ efficiency} = (2^{(-1/\text{slope})} - 1) * 100\%$$

In a 10-fold dilution series, efficiency is calculated according to the same formula with the exception of number 2 which is substituted for number 10.

qPCR amplifications and monitoring were carried out on 96-well plate using Stratagene® fluorescence detection machine (model Mx3005P). For TaqMan assays the final volume of qPCR mix per single well was 20µl consisting of 2µl of cDNA template, 6.4µl of DNase and RNase free water, 0.6µl of forward primer (final conc. 0.3µM), 0.6µl of reverse primer (final conc. 0.3µM), 0.4µl of probe (final conc. 0.1µM), and 10µl of x2 TaqMan master mix (Applied Biosystems, cat. num. 4369542). For SYBR® Green assays the final volume of qPCR mix per single well was 25µl consisting of 5µl of cDNA template, 6.4µl of DNase and RNase free water, 0.5µl of forward primer (final conc. 0.25µM), 0.5µl of reverse primer (final conc. 0.25µM), and 12.5µl of x2 Brilliant III Ultra-Fast SYBR® Green master mix (Agilent, cat. num. 600882). The precise quantity of cDNA used in these reactions depends on the total RNA yield from the RNA extraction step, as equal amounts of RNA ought to be added to cDNA syntheses tubes. Typically, in an ideal experiment the neat concentration of cDNA used for qPCR is 1µg/ml. In TaqMan assays, the 2µl of template is prepared by 1:5 dilution, yielding 0.001 % (w:v) final cDNA concentration in a single PCR well. Similarly in SYBR® Green assays, the 5µl of template is also prepared in a 1:5 dilution making the final concertation to be 0.008 % (w:v). RT+ samples were always run in triplicate while RT- samples were run in triplicates and duplicates. Every qPCR ran had a water only control to account for possible contaminations. TaqMan qPCR reactions were run according to the following thermal programme:

- 2 minutes at 50˚C
- 10 minutes at 95˚C
- 15 seconds at 95˚C (40 cycles)

- 1 minute at 60˚C (40 cycles)

On the other hand, SYBR® Green reactions were run according to the following programme:

- 3 minutes at 95˚C — 1 cycle
- 10 seconds at 95˚C
- 20 seconds at 60˚C — 40 cycles

- 1 minute at 95˚C
- 30 seconds at 55˚C
- 30 seconds 95˚C — 1 cycle

2.3.6 Automated dideoxy DNA sequencing

Dideoxy DNA sequencing, also known as Sanger sequencing, is a molecular method in which the identity and order of DNA bases can be determined (Sanger et al. 1977). According to this method DNA sequencing in essence is a primed DNA synthesis reaction carried in the presence of dNTPs and dideoxynucleotides (ddNTPs) in a ratio of 4:1. Because ddNTPs are missing the 3` hydroxyl group, which is required for DNA chain extension, the synthesis reaction is prematurely ended. Therefore, after the reaction, multiple DNA fragments are generated depending on the DNA template and the incorporated ddNTP. In the automated version of the technique, each of the four ddNTPs is labelled with a different fluorophore and therefore can be differentiated and detected by laser exposure following capillary gel electrophoresis (Figure 2.5). Sanger sequencing has been superseded by NGS (see section 2.5) but is still widely used for validation of NGS results and for obtaining long contiguous DNA sequence reads (>500 bp).

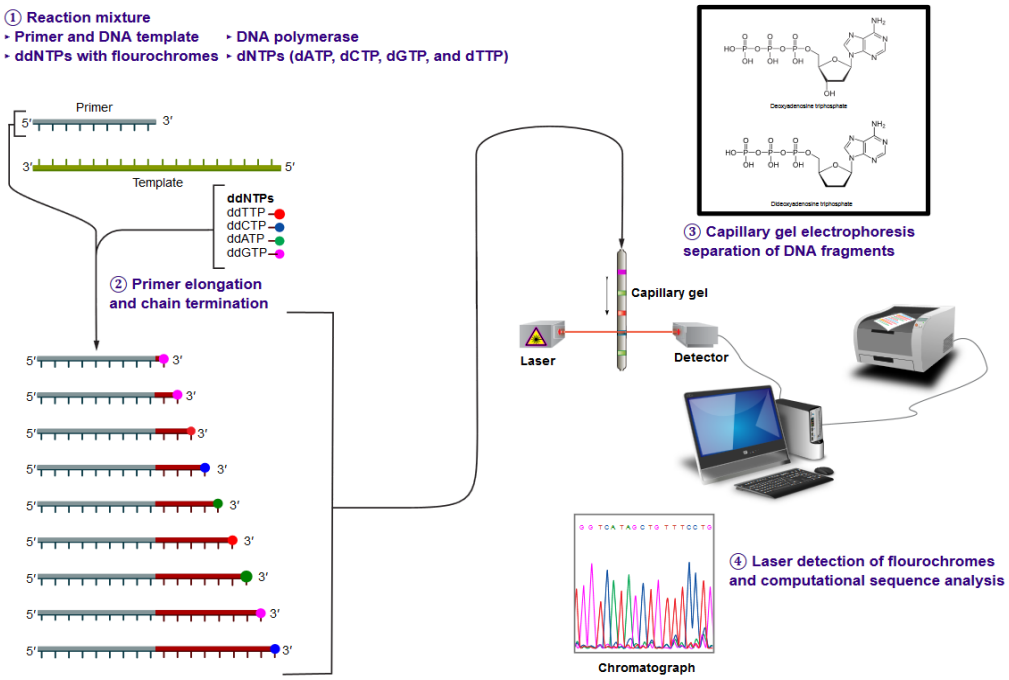


Figure 2.5: Automated Sanger sequencing procedure.

2.3.6.1 PCR amplicon purification and automated dideoxy DNA sequencing procedure

In this thesis automated Sanger sequencing of cDNA was used to validate RNA sequencing by NGS (RNAseq; see section 2.5.1) findings and to check the sequence identity of cloned DNA constructs. Electrophoresed PCR products were visualized on UV station, cut out of the agarose gel using a sharp scalpel and were purified using a StrataPrep® DNA Gel extraction Kit (Agilent Technologies, cat. num. 400766) following manufacturer's instructions. The Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, cat. num. 4337456) was used in the sequencing reactions. The following reagents were added per sequencing reaction: 1.5μl Big Dye termination mix, 1μM primer, 2μl sequencing buffer, 30ng purified PCR product template and DNase and RNase free water up to 10μl total volume. Samples were incubated at PCR thermal cycler according to the following program:

- 96°C for 30s
- 50°C for 15s
- 60°C for 4min

} 25 cycles

- 4°C hold to stop the reaction

Finally, DNA was precipitated, dried and sequenced at core sequencing facility using 3130xl ABI PRISM Genetic Analyzer (Life Technologies). For precipitation 62µl of DNA precipitation mix composed of 50µl of ethanol (100%), 10µl of DNase and RNase free water and 2µl of Sodium Acetate (NaOAc; 3M) was added to each tube. The samples were mixed by flicking and left on the bench at room temperature for 20min. The samples were centrifuged at 13.2 x 1000rpm for 1 hour. The supernatants were removed and discarded, by setting the pipette to 100 µl and entering the pipette into the tube on the opposite side to the DNA pellet. 250 µl 70% ethanol was added to each tube and the samples were centrifuged at 13.2 x 1000rpm for 10min. The resulting supernatants were removed and discarded in two lots by setting the pipette to 200µl. The DNA pellets were dried at 95^{o}C with the lids of the tubes open and, finally, samples were submitted for sequencing.

2.3.7 INTS12 construct cloning and DNA plasmid transfection optimizations

Molecular DNA cloning refers to a set of experimental methods used to assemble a recombinant DNA molecules and to direct their replication, transcription and, if encoding a protein, translation inside the cells. In this project transient INTS12 overexpression was used in parallel to its knockdown in order to test the effect of this manipulation on gene expression. NCBI's INTS12 mRNA variant 2 (NM_001142471.1) as well as novel naturally-occurring variant 3 with premature stop codon (Obeidat et al. 2013) were PCR amplified, sequenced by Sanger reaction and cloned into pcDNA3.1 backbone plasmid with expression driven from CMV promoter. Variant 2 encodes canonical INTS12 protein sequence, while variant 3 encodes a smaller version of the protein that misses serine rich compositional bias. Constructs were maxi prepped (Qiagen, cat. num. 12302) in order to have enough DNA material for cell biology experiments. The constructs were named pcDNA3.1-INTS12_v2 and pcDNA3.1-INTS12_v3 for the variants 2 and 3 respectively (Figure 2.6).

Figure 2.6: Naturally occurring INTS12 variants 2 and 3 were cloned into overexpression constructs. The constructs were named pcDNA3.1-INTS12_v2 and pcDNA3.1-INTS12_v3 respectively.

pEGFP-N1 plasmid containing GFP open reading frame (ORF) was used to optimize FuGENE6® mediated transfection. Cells were quantitatively tested for transfection efficiency by treating with 2µg/ml, 1µg/ml and 0.5µg/ml of plasmid DNA in 3:2 and 3:1 FuGENE6® to construct ratios. The ratios are volume-to-total-mass and therefore if the quantity of DNA used was 2 µg, than 3:1 ratio implies using 6µl of FuGENE6®. Appropriate volumes of FuGENE6® were added to BEGM media and incubated for 5min at RT followed by addition of DNA. The mixture was incubated for 15min and added to freshly replaced media as appropriate. After 48h cells were washed with PBS (Oxoid, cat num. BR0014), fixed with 4% formaldehyde (Sigma-Aldrich, cat. num. 252549) and stained for 5-10min with 1µg/ml of 4',6-diamidino-2-phenylindole (DAPI) for dsDNA staining. Cells were imaged epifluerescently using DAPI and GFP exposures with 100x magnification (see section 2.4.1). Mean GFP fluorescence was quantified per a field using Volocity image measurement tool. The optimal transfection conditions of 2µg/ml of DNA using 3:1 ratio were used for INTS12 overexpression experiments.

2.4 Cell microscopy

Cell microscopy refers to various imaging techniques used to visualize cells or biological molecules. In this thesis light microscopy was used primerily to check cells for overall confluency. Immunofluorescence (IF) and epifluorescent microscopy were used to determine the subcellular protein localization as well as to validate gene knockdown on the protein level. IF is a semi-quantitative technique which may be used to compare levels of protein expression between experimental conditions, albeit it is of lower sensitivity then Western blotting (WB) which is better suited for this purpose. Also, by IF alone it is not possible to ascertain the protein's molecular weight and therefore there is less certainty in the assay's specificity. However as outlined in the Introduction, according to the ENCODE, IF in combination with RNAi-mediated knockdown can be

used to demonstrate antibody specificity as well as successful protein silencing (Landt et al. 2012).

2.4.1 Immunofluorescence and epifluorescent microscopy

There are two classes of IF: direct and indirect. In direct IF the antibody used to detect a particular protein of interest is attached to fluorophore detected on epiflurescent microscope. On the other hand, in indirect IF which was used in this project, the antibody used to detect the protein is known as primary antibody and it is not conjugated. A secondary antibody carrying the fluorophore is applied to bind onto the primary antibody. This is possible because an antibody consisting of four polypeptide chains has two predominant parts: a variable region, which recognizes the protein's antigen, and constant region, which makes up the structure of the antibody molecule and which can be recognized by another antibody. In some cases, cell membrane permeabilization is required to allow antibody's access to intracellular proteins. A procedure of blocking with milk or goat serum is applied in order to minimize the degree of non-specific binding.

Primary antibodies can be either monoclonal or polyclonal. Polyclonal antibodies are produced by different B cell lineages raised against a specific antigen, each identifying a different epitope. In contrast, monoclonal antibodies are produced by identical clones of B cells that came from the same parent cell thus generating antibodies of monovalent affinity targeting the same epitope (Lipman et al. 2005). Imaging cells transfected with fluorescent molecules or stained with fluorescent antibodies can be achieved with epifluorescent microscope. In epifluorescent microscopy, the sample is shined with a specific, desired, bandwidth of wavelengths and then the weaker emitted light is used detect the excited molecules.

2.4.1.1 Immunofluorescence procedure

Indirect IF often requires experimental optimization of primary and secondary antibodies for optimal imaging results. After a series of optimizations, a final procedure was developed. Cells were grown on 8-chamber glass slides seeding 8000 cells onto each chamber and were

left un-treated or were transfected with INTS12 and scrambled D-siRNAs as described previously (see section 2.2.2). After reaching 50-100% confluence (observed confluency depending on the condition) cells were washed x3 with PBS (Oxoid, cat num. BR0014) and fixed with PBS diluted 4% formaldehyde (Sigma-Aldrich, cat. num. 252549) for 15min at RT. Cells were washed x3 with PBS and permeabilized with 0.15% Triton X-100 for 10min at RT. Then, cells were washed x3 with PBS and incubated in block solution consisting of 10% goat serum diluted by 1% bovine serum albumin PBS solution for 30min at RT. Finally, cells were washed x3 with PBS and incubated for ~24h at 4^{o}C with 1µg/ml of raised in rabbit IgG polyclonal anti-human-INTS12 antibody (Sigma Prestige Antibodies, cat. num. HPA035772) and 1µg/ml of general rabbit IgG isotype control (Abcam, cat. num. 171870). Next day, after PBS wash 15µg/ml of secondary goat anti-rabbit-IgG conjugated with rhodamine (TRITC) fluorophore was incubated for 1h at RT. Cells were left on agitating rack and protected from light during this time. After that cells were washed x3 with 0.05% (v/v) PBS-Tween and incubated for 5-10min with 1µg/ml of DAPI for dsDNA staining. Lastly, cells were washed x3 with 0.05% (v/v) PBS-Tween, mounted and epifluorescently imaged on epifluorescent microscope using DAPI and TRITC exposures. It was ensured to keep the same exposures across the conditions to avoid differences in the fluorescence intensity driven by different exposures. Cells were magnified 200 times unless specified differently.

2.5 RNA next generation sequencing

RNAseq by NGS is thought to have revolutionized the field of transcriptomics because of the wealth of analytical information generated, including transcriptome assembly, differential gene expression, differential expression of individual mRNA isoforms, differential splicing, and differential promoter use (Wang et al. 2009). The critical advantage of RNAseq over tilling array approaches is a greater dynamic range in estimating gene expression, i.e. accurate estimation of highly and lowly expressed genes, and less technical variability between

biological experiments (Majewski and Pastinen, 2011). Microarray and RNAseq derived gene expression correlate fairly well for medium levels of expression (correlation coefficient of 0.5), but correlation is low for genes with either very high (correlation coefficient of 0.2) or very low (correlation coefficient of 0.1) expression (Wang et al. 2009).

In typical RNAseq experiment, the first step is isolation of RNA species of interest (such as mRNA) and its conversion into a library of cDNA fragments. mRNA enrichment by poly(A) selection is frequently employed for the estimation of expression of protein coding genes, however ribosomal RNA depletion method has increased in prominence due to its ability to assess the entire transcriptome and not just a subset of protein coding genes. Sequencing adaptors are subsequently added to each cDNA fragment and its sequence is recovered by NGS.

In NGS, vast numbers of short reads are sequenced in a single stroke (Illumina®). Briefly, reads are attached to the sequencing platform by using the attached adaptors. PCR is carried out to amplify each read, creating a spot with many copies of the same read. This step is necessary as otherwise base calling is hardly detectable. Fluorescently labelled nucleotides are added to the slide together with DNA polymerase. Complementary strands of the reads are denatured by heating the platform and polymerase then adds the labelled nucleotides generating coloured fluorescent signal detected by imaging, indicating which base has been added. As the bases have a terminator, the extension cannot occur continuously but rather happens one at the time to allow for the monitoring of the process. The terminators are chemically removed allowing the next base to be added and the whole process is repeated until the entire sequence is retrieved (Figure 2.7). The depth of sequencing refers to the total number of sequenced reads with greater depth implying greater number of sequenced reads. Ribosomal RNA depleted library typically requires greater depth of sequencing because of the substantial number of reads derived from ribosomal RNA attenuating the signal of other genes.

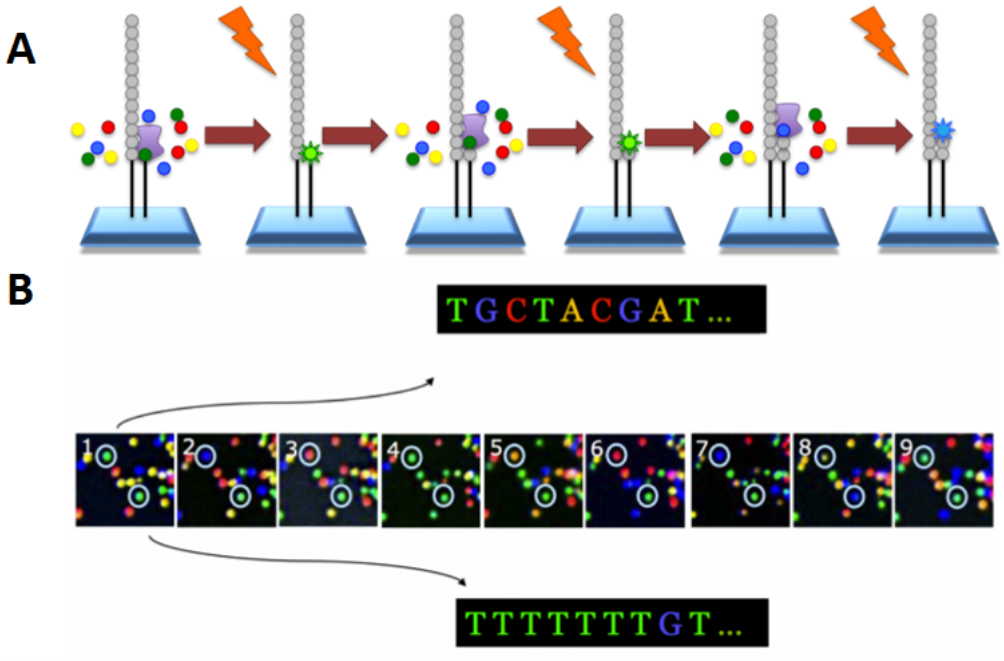


Figure 2.7: Illumina® NGS procedure. After sequencing library preparation, reads are attached to the sequencing platform and PCR amplified. Sequencing is commenced by the addition of fluorescently labelled nucleotides and DNA polymerase which adds the bases one at a time allowing for monitoring of sequencing (A). An example of continuous imaging (9 bases) of part of NGS platform. T, G, C, and A nucleotides are labelled with greed, blue, red and orange dyes respectively. The imaged colour indicates the identity of the added nucleotide (B).

2.5.1 RNAseq experiments

INTS12 knockdown was initiated and total RNA was extracted as described before (see sections 2.2.2.3 and 2.3.2 respectively). Sequencing samples were ensured to have RIN scores ≥ 8 (see section 2.3.3). There were four experimental conditions performed in three independent biological replicates (i.e. using different cell vials of the same donor and at different times):

- Un-transfected P3 HBECs
- P3 HBECs transfected with scrambled D-siRNA
- P3 HBECs transfected with INTS12 D-siRNA A
- P3 HBECs transfected with INTS12 D-siRNA C

Transcriptomic profiling was performed at 48h post initiation of knockdown using D7F3206 cells and 120h post initiation of knockdown

using D195307. Sequencing library was prepared with Illumina TruSeq RNA Sample Prep Kit v2. mRNA was poly-A selected by capturing total RNA samples with oligo-dT coated magnetic beads. The mRNA was then fragmented and randomly primed. cDNA was synthesised using random primers. Finally, ready-for-sequencing library was prepared by end-repair, phosphorylation, A-tailing, adapter ligation and PCR amplification. Paired-end sequencing was performed on the Illumina© HiSeq2000 platform using TruSeq v3 chemistry over 100 cycles yielding approximately 40 million reads per sample stored in raw FASTQ files used for subsequent analyses.

2.6 Chromatin immunoprecipitation

ChIP is a method used to assess the binding profile of proteins interacting with gDNA. It can be used to decipher protein's binding sites in basal conditions or to study differential binding due to experimental manipulations. It can also be used to study specific histone modifications. This technology has been used for accurate and high-resolution mapping of the protein-gDNA interaction loci that are important in the understanding of many processes in development and disease. Briefly, in ChIP experiments the first step is the fixing of proteins to their cognate gDNA sequences by crosslinking by formaldehyde treatment. Then gDNA is IPed with antibody targeting a protein of interest. After precipitation DNA is isolated and can be tested by PCR or NGS approaches thus being ChIP-PCR and ChIPseq respectively. ChIPseq, in contrast to ChIP-PCR allows for genome-wide profiling of gDNA interacting proteins at ~10-50bp resolution accuracy (Figure 2.8; Park et al. 2009). A plethora of downstream analyses can be performed on ChIPseq dataset e.g. identification of candidate genes regulated by the protein, co-occupancy with other regulatory elements and conservation analyses (see section 2.8.2). INTS12 ChIPseq was performed, primarily but not exclusively, to determine whether INTS12 binding is enriched upon genes identified as differentially expressed following INTS12 knockdown. ChIP-PCR was utilized to technically

validate the binding profiles identified in ChIPseq by testing the enrichment at three positive sites and one negative site.

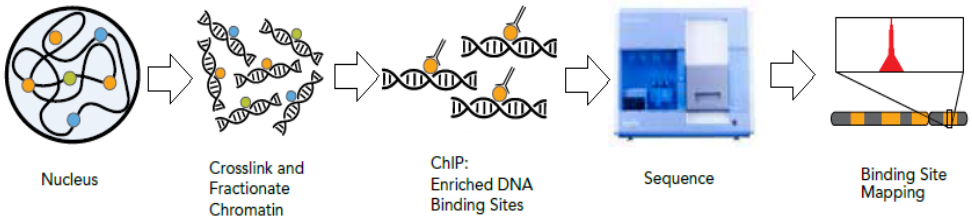


Figure 2.8: ChIPseq experiment procedure: protein fixing, IP, sequencing, and read mapping.

2.6.1 INTS12 chromatin immunoprecipitation sequencing procedure

HBECs from two different donors (D195307 and D7F3158) were fixed with formaldehyde solution for 15 min. Formaldehyde solution contained 11% formaldehyde (Sigma cat. num. F-8775), 0.1M sodium chloride (Sigma cat. num. S5150-1L), 1mM EDTA (pH 8.0; Sigma cat. num. 03690-100ML), 50mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES; pH 7.9; Applichem cat. num. A6906,0125). Fixation was quenched with 0.125 M glycine (Sigma cat. num. G-7403-250G). Chromatin was isolated by the addition of lysis buffer (Active Motif ChIP-IT® ChIPseq kit, cat. num. 53041), followed by disruption with a Dounce homogenizer (Active Motif ChIP-IT® ChIPseq kit, cat. num. 53041) to allow for efficient chromatin preparation. Lysates were sonicated and the DNA sheared to an average length of 300-500bp. Genomic DNA regions of interest were isolated using 4 µg of antibody against INTS12 (Sigma cat. num. HPA03577) following manufacturer's specifications (Active Motif ChIP-IT® ChIPseq kit, cat. num. 53041). Complexes were washed, eluted from the beads with SDS buffer, and subjected to RNase and proteinase K treatment. Crosslinks were reversed by incubation overnight at 65^{o}C, and ChIP DNA was purified by phenol-chloroform extraction and ethanol precipitation. Pellets were re-suspended and the resulting DNA was quantified on a NanoDrop 2000 spectrophotometer (Thermo-Scientific©). Extrapolation to the original chromatin volume allowed quantitation of the total chromatin yield. 30µg chromatin of each

sample was precleared with protein A agarose beads (Invitrogen cat. num. 15918-014). Unprecipitated genomic DNA (i.e. input control) was prepared from a pool of equal aliquots of the 2 samples.

Illumina© sequencing libraries were prepared from the ChIP and Input DNAs by the standard consecutive enzymatic steps of end-polishing, dA-addition, and adaptor ligation. After a final PCR amplification step, the resulting DNA libraries were sequenced on Illumina© NextSeq 500 Illumina© sequencing machine yielding approximately 40 million single-end 75bp raw reads FASTQ files per two ChIP samples from each donor cells and one input control of both donors. Raw files were used for subsequent bioinformatic analyses.

2.6.1.1 The choice of the antibody

The same antibody used in IF was also used in ChIP protocol. Therefore, ENCODE criteria were followed for its validation by combining IF with gene knockdown (see section 3.6.1.3.2). Disappearance of nuclear signal in cells treated with D-siRNA indicated its specificity (Figure 2.22). Moreover, antibody was validated for ChIPseq due to passing a pilot experiment by generating sufficient number of unique reads were sequenced using it (see section 6.3.1).

2.6.2 INTS12 chromatin immunoprecipitation polymerase chain reaction

INTS12 peak regions used for qPCR validation were prioritized based on ChIPseq signals observed on the genome browser. The following three positive and one negative binding regions were chosen for ChIP-PCR validation, based on hg19 (GRCh37) sequence and RefSeq annotation:

- ACTB TSS -145 (145 bps upstream gene's TSS)
- NBPF1 TSS +108 (108 bps downstream gene's TSS) } positive
- POR TSS -154 (154 bps upstream gene's TSS)
- Untr12 (region Chr12: 61667747 – 61667824) negative

PCR primers were designed as described before (see section 2.3.5.4 and Table 4 of Appendix) to span the above regions. qPCR reactions were carried out in triplicate on above specified genomic regions upon 500 times diluted precipitated gDNA (i.e. final DNA quantity was 12.5 ng) from each donor and input control using SYBR Green Supermix (Bio-Rad, cat. num. 1708880). The raw qPCR C_t values were converted into the number of binding events detected per 1000 cells according to the manufacturers of ChIP-PCR kit specifications (Active Motif ChIP-IT® qPCR analysis kit, cat. num. 53029).

2.7 Statistical considerations

Data were grouped from multiple experiments and are expressed as the mean ± standard error of the mean (SEM). Unless otherwise specified, statistical significance was assessed by ordinary one-way ANOVA followed by Fisher's Least Significant Difference test (Hayter, 1986). Results were considered significant when $P<0.05$ (Table 2.1). For high throughput analyses the nominal P values were corrected for multiple comparisons using Benjamin-Hochberg false discovery rate (FDR) correction to minimize the risk of type I error (Benjamini and Hochberg, 1995).

P values	Donation
>0.05	ns
<0.05	*
<0.01	**
<0.001	***
<0.0001	****

Table 2.1: Used star indications of statistical significance.

2.8 Bioinformatic analyses

2.8.1 RNAseq analyses

Once the cDNA fragments have been sequenced, the first task in the RNAseq data analysis is to evaluate the quality of sequenced reads. For that purpose FastQC is often used

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Although Illumina© high throughput sequencing provides highly accurate sequencing data, poor quality reads are possible. The primary errors are substitution errors and these errors rise in frequency at the 3' ends of the reads. One way to investigate sequence quality is to visualize the quality scores (Q-scores) which can be generated by FastQC. Q-scores express error probability. In particular, it serves as a compact way to communicate small error probabilities. Mathematically, the probability that is A not true, P(~A), is expressed by a Q-score , Q(A), according to the relationship:

$$Q(A) = -10 \log_{10}(P(\sim A))$$

Therefore the relationship between quality score is as follows:

- If quality score is 10 then error probability is 0.1
- If quality score is 20 then error probability is 0.01
- If quality score is 30 then error probability is 0.001
- If quality score is 40 then error probability is 0.001

Q-scores are assigned by sequencing machine and it is generally acceptable to use raw reads data for subsequent analyses without any data trimming if the average Q-scores per base are above 28 (Conesa et al. 2016). RNAseq tools fall into three categories: (1) programs used for read alignment, (2) programs used for transcriptome assembly, which as far as end result is concerned, is equivalent to the process of genome annotation, and (3) programs used for individual transcript and gene quantification.

After initial data QC, the next step is to map (i.e. align) short reads to the reference genome or to assemble reads to contigs before genomic alignment. Several mapping programs have been developed with an aim to identify, for each short read in the dataset, all the locations in a reference genome that show perfect or near perfect matching. The differences among the alignment programmes lay in the algorithm design and therefore computational efficiency. Bowtie is among the top fastest short-read aligners (Langmead et al. 2009), Maq can make the use of reads quality scores (Li, Ruan et al. 2008), while SeqMap

considers insertions and deletions (Jiang and Wong, 2008). Read mapping is a relatively straightforward task for reads derived from non-splice or non-polyadenylation sites as, for each read, there is only one correct result given the reference genome and number of mismatches allowed. These reads produce un-gapped alignments. Poly(A) tails can be identified simply by the presence of multiple As or Ts at the end of some reads. For large transcriptomes, alignment is complicated by the fact that some reads map to multiple locations of the genome. This can be addressed by proportionally assigning them based on the number of reads mapped to their neighbouring unique sequences (Mortazavi et al. 2008). Alternatively reads aligning to multiple locations can be discarded if they map to more than previously specified threshold. This threshold is determined on an arbitrary basis.

The greatest computational challenge is related to aligning reads derived from mRNA splice junctions. Because spliced alignment is critical in RNAseq analysis this task has attracted much research effort in recent years (Engström et al. 2013). Bowtie is an example of splice-aware aligner capable of producing spliced alignments in addition to un-gapped alignments. Splice alignment is implemented in a two-step approach in which initial read alignments are analysed to infer splice junctions and these junctions are used to guide the final alignment (Langmead et al. 2009). Bowtie can use existing gene annotation to inform spliced-read placement and in this case alignment results are highly specific to a particular version of the transcriptome. Spliced alignments are critical for correct placing of exon-intron boundaries. Alignment information is used for transcriptome assembly, again, typically in the context of reference annotation (i.e. it is a Reference Annotation Based Transcript). Although it is possible to perform differential gene expression without transcriptome assembly, it has been suggested that the failure to look for new transcripts can bias expression estimates and reduce accuracy (Trapnell et al. 2010). Predicted annotation and genome sequence are used to compare read counts between loci for differential gene expression analysis.

2.8.1.1 Tuxedo pipeline

Tuxedo tool kit for RNAseq analysis contains all the categories of programs used in RNAseq analysis (Trapnell et al. 2012). This pipeline can serve multiple purposes but the main one is to compare transcriptome profiles between two biological conditions such as wild-type versus mutant or control versus knockdown experiments (Figure 2.8). First, reads from the considered conditions are aligned to the reference genome by TopHat. TopHat uses Bowtie for read mapping and therefore can produce spliced alignments. The next step in the workflow is the transcriptome assembly with Cufflinks. The algorithm takes mapped reads information and assembles overlapping reads into contigs as it would have been done for de novo genome assembly. Smaller contigs contained within larger contigs and derived from different exons can be differentiated thanks to the spliced alignments (Figure 2.9). Assembled contigs are used to identify 'incompatible' fragments that must have originated from differently spliced isoforms. These incompatible fragments are determined based on irreconcilable spliced alignment. Fragments are then connected if they are compatible and their spliced alignments overlap in the genome. Therefore, the paths through the fragments are sets of mutually compatible fragments that could be merged into complete isoforms. Isoforms are finally assembled from the overlap graph. In this process, Cufflinks implements Dilworth's Theorem which states that the number of mutually incompatible fragments is the same as the minimum number of isoforms needed to explain all the fragments.

Transcriptome assembly is performed on individual sample basis and all the assemblies are merged into one unified and final genome annotation using Cuffmerge utility (Figure 2.9). When using reference annotation to guide assembly, Cuffmerge performs Reference Annotation Based Transcript assembly. If an isoform did not receive enough coverage, Cufflinks may not have recovered it and contigs derived from that isoform will not be linked due to other fragments being missed. However, when Cufflinks assemblies are unified the entire isoform may be reconstructed.

Cuffmerge will merge contigs if they overlap, and agree on splicing (i.e. are not mutually incompatible).

In order to perform differential gene expression, the alignment file and the merged predicted assembly are fed to Cuffdiff, which calculates normalized expression levels and tests the statistical significance of observed changes in read counts between the loci specified in the annotation. Gene expression values are normalized for gene length, as raw read counts are higher for longer genes and therefore are not comparable unless normalized for gene size. Expression is also normalized for library size because differences between the conditions in library size can yield non-biological differences in gene expression. Gene expression is represented as fragments per kilobase per million reads (FPKM) units for paired-end sequencing which includes adjustments for both the gene length and library size rendering expression values comparable. Reads per kilobase per million reads (RPKM) is used for singe-end sequencing but it is the same as FPKM. The mathematical formula for FPKM was introduced by Mortazavi et al. (Mortazavi et al. 2008) and it is as follows:

$$FPKMfeature = \frac{X\ feature}{N\ library * L\ feature} * 10^9$$

Thus FPKM value per considered genomic feature ($FPKM_{feature}$) is calculated by dividing the total mappable reads aligning to the feature divided by a product of total reads in sequencing library ($N_{library}$) and the number of bases constituting the feature ($L_{feature}$) multiplied by 10^9. Cuffdiff performs quantifications and differential expression analyses on individual isoform level and gene expression is simply the sum of FPKM values of its individual isoforms. Because a read from a shared exon could have come from one of several isoforms a simple counting procedure will not suffice for that purpose. Therefore both Cufflinks and Cuffdiff implement a linear statistical model to estimate an assignment of abundance to each transcript that explains the observed reads with maximum likelihood (Figure 2.9; Trapnell et al. 2010; Trapnell et al. 2012).

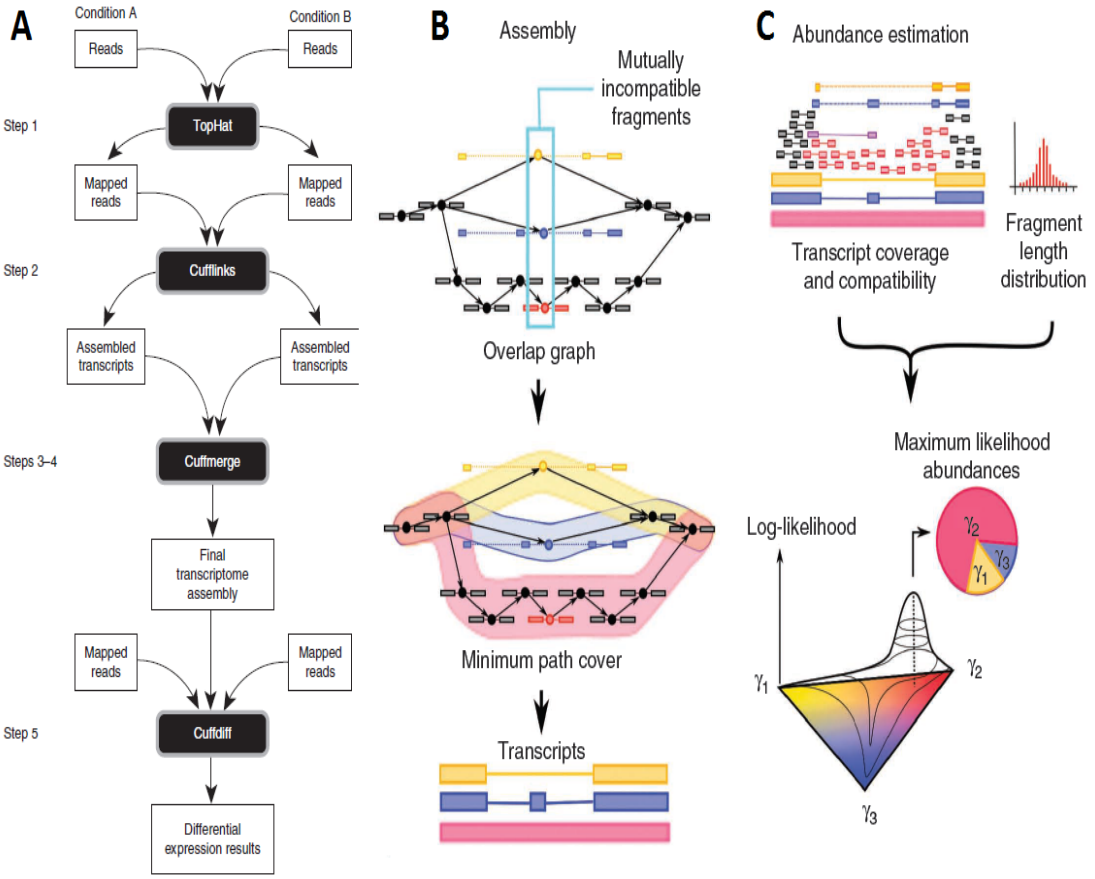


Figure 2.9: An overview of the tuxedo pipeline for RNAseq analysis approach. Reads are first mapped to the genome with TopHat. These mapped reads are provided as input to Cufflinks, which produces one file of assembled fragments. Assembly files per each of the available conditions and replicates are then merged using reference transcriptome annotation as a guide into a unified annotation for further analysis. Predicted annotation is then used for differential read count analysis performed by Cuffdiff (A). For transcriptome assembly overlapping reads are assembled into contigs which can be considered to be exons. Contig assembly considers spliced alignments and therefore can separate between fragments. Fragment pairs are tested for compatibility based on relative locations of spliced alignments, i.e. whether spliced alignments are within other spliced alignment indicating a separate isoform origin (B). For isoform abundance estimation, gene expression can be determined at individual isoform resolution level based on a statistical model in which the probability of observing each fragment is a linear function of the abundances of the transcripts from which it could have originated explaining the observed reads with maximum likelihood (C). Reproduced from Trapnell et al. (Trapnell et al. 2010 and Trapnell et al. 2012).

Cuffdiff can also perform additional analyses beyond mere differential gene or isoform expression. For example, by grouping transcripts into biologically meaningful groups, Cuffdiff identifies genes with evidence of differential promoter usage or differential expression of coding sequences and can also identify genes with varying isoform abundances providing evidence of differential splicing per gene.

2.8.1.1.1 Tuxedo pipeline datatypes and commands

RNAseq analyses were completed on Linux Ubuntu 12.04 LTS (64-bit) run on remotely-accessed computer with 126 GB of RAM memory. All the program tools of Tuxedo pipeline were installed according to the developer instructions.

i) Reads alignment to reference genome and initial QC

Raw reads are saved as a FASTQ file which in essence is a text file containing all the reads' sequences. Raw files were tested for quality on java implementation of FastQC. There were two FASTQ files per sample because sequencing was pair ended. TopHat alignment was executed with the general command shown below. In it, the `-p` parameter specifies the number of cores to be used in the alignment (it is arbitrary and depends on the available computer processing units), `-G` parameter specifies the path to genome annotation, `-o` parameter specifies the name of output folder, `genome` is the base name for Bowtie indexed genome, and there are two compressed raw read files per RNAseq sample because the sequencing is paired. Reference file is saved in GTF format. Reference genome used was GRCh37 (also known as hg19) and its equivalent annotation was taken from Ensembl database. Both the reference genome sequence and annotation were downloaded from Illumina's iGenomes repository of model organisms (https://support.illumina.com/sequencing/sequencing_software/igenome.html).

```
tophat -p 2 -G <path_to_annotation>/genes.gtf -o
                    output_folder
  <path_to_Bowtie_indexed_genome>/genome_prefix
```

```
<path_to_compressed_raw_data>/_1_reads.fastq.gz
<path_to_compressed_raw_data>/_2_reads.fastq.gz
```

This command generates a folder with binary `accepted_hits.bam` alignment file and other files including read alignment statistics.

ii) Individual transcriptome assemblies

Cufflinks transcriptome assemblies on individual RNAseq samples were performed by executing the common shown below. The critical element of the Cufflinks assembly is provision of binary BAM file alignment containing un-gapped and spliced alignment information. This alignment file was created by TopHat and is found in the TopHat output folder. Thus alignment file is per sequenced RNAseq sample. The purposes of `-p` and `-o` parameters are the same as specified in the TopHat alignment stage.

```
cufflinks -p 2 -o output_folder
<path_to_alignment_file>/accepted_hits.bam
```

This command generates a folder with `transcripts.gtf` predicted transcriptome annotations per RNAseq sample.

iii) Final transcriptome merging

Final transcriptome assemblies were merged into a final and unified assembly by merging the predicted GTF annotations generated on individual sample basis. First, a list of paths to predicted GTF assemblies (`transcripts.gtf` files that were generated per sample by Cufflinks) was typed in a column-wise fashion in a text file named `gtf_out_list.txt`. Then Cuffmerge command shown below was executed. The `-g` parameter specifies the path to the reference annotation which was also used during the alignment. The `-s` parameter specifies the path to the fasta-type sequence of reference human genome (i.e. not the indexed genome sequence).

```
cuffmerge –p 10 –g
<path_to_reference_annotation>/genes.gtf –s
<path_to_fasta_genome_sequence>/genome.fa
gtf_out_list.txt
```

This command yields final `merged.gtf` transcriptome assembly file.

iv) Differential analyses

Differential RNAseq analyses were performed using `merged.gtf` assembly and `accepted_hits.bam` alignments. The general command is shown below. Alignments representing biological replicates of the same condition are grouped together by typing the replicate alignments one after another separated by comma. The two blocks of conditions that are being compared are separated by a space. Predicted genome annotation generated by Cuffmerge is supplied to the Cuffdiff differential analyses.

```
cuffdiff –p 7
<path_to_predicted_annotation>/merged.gtf
<condition_X_sample_1>/accepted_hits.bam,
<condition_X_sample_2>/accepted_hits.bam,
<condition_X_sample_3>/accepted_hits.bam
<condition_Y_sample_1>/accepted_hits.bam,
<condition_Y_sample_2>/accepted_hits.bam,
<condition_Y_sample_3>/accepted_hits.bam
```

The output files contain the information about differential gene expression analysis as well as other information such genes with evidence of differential splicing or differential promoter usage. Comparisons were performed between un-transfected cells and cells transfected with scrambled D-siRNA (UTvsNC), cells transfected with scrambled D-siRNA and INTS12 D-siRNA A (NCvsA), and cells transfected with scrambled D-siRNA and INTS12 D-siRNA C (NCvsC).

2.8.1.1.2 Identification and visualization of novel splice variants

The Cuffmerge generated novel gene transfer format (GTF) annotation file was compared to Ensembl GTF annotation by using Cuffcompare. The command used is shown below.

```
cuffcompare -r
<path_to_reference_annotation>/genes.gtf -i
<path_to_predicted_annotation>/merged.gtf
```

The generated `refmap` file was used to ascertain which of the predicted mRNA isoforms were matching the isoforms annotated in the reference annotation and which isoforms are novel. All the predicted variants were visualized by SpliceGrapher using `gene_model_to_splicegraph.py` and `plotter.py` functions as specified in the SpliceGrapher user's guide (http://splicegrapher.sourceforge.net/userguide.html).

2.8.2 ChIPseq analysis pipeline

The aim of ChIPseq experiment analysis is to gain an adequate number of mappable reads aggregated at the binding regions. In that context, a mappable read is defined as a read that maps (i.e. aligns) to a unique location in the genome. The popular alignment programs typically allow for two mismatches when aligning these reads. On the other hand, non-redundant reads are the mappable reads that occur only once in the entire dataset. Redundant reads are also known as duplicates and in contrast to RNAseq where such reads may have been derived biologically because of differences in gene transcription, in ChIPseq they are presumed to be generated artificially as a result of PCR amplification on NGS platform. Before any large-scale production run of a ChIPseq experiment it is desirable to conduct a pilot experiment where a small number of reads is generated (Ma and Wong, 2011). This dataset can be used to assess the overall success of pilot ChIPseq. For a pilot experiment to be deemed successful (a) percentage of uniquely mappable reads ought to achieve at least $\frac{1}{3}$ of total reads, and (b)

percentage of non-redundant reads should be greater than 50% of the total mappable reads (Ma and Wong, 2011).
As it is the case with RNAseq analysis, the first step in ChIPseq analysis is mapping reads to the reference genome. This can be achieved with any of the available read alignment tools that were mentioned before (see section 2.8.1). The next step is concerned with background estimation. Although in ChIPseq a considerable fraction of reads would have originated from ChIP fragments, a significant proportion are nonspecific. These could have originated for a number of reasons such as library contamination, PCR amplification selection yielding redundant reads, nonspecific antibody binding or image processing sequencing errors.

Knowledge about the background rate is pivotal for the assessment of statistical significance of the enrichment of binding regions. There is no obvious way to ascertain which read is derived from a true ChIP fragment or is part of background noise but a signal read is considered as such if it falls within enriched regions, while background read would fall outside enriched region. Therefore, peak calling, i.e. identification of binding sites, is a signal-over-noise detection problem. In each called peak region, the fold change of the ChIP signal intensity to the control signal intensity is used as a local estimate of the signal-to-noise ratio (Zhang et al. 2008). The input (i.e. no-ChIP) control sample is often used to differentiate between true read distribution enrichment and random noise.

The general consensus highlights the importance of conducing biological replicates of ChIPseq assays to assess biological reproducibility and thus to have confidence in the identified binding sites and to, potentially, perform subsequent analyses on the subset of reproducible binding sites (Landt et al. 2012; Figure 2.10). Alternatively, if peak regions show good biological reproducibility, one may use one ChIPseq sample as the representative (Landt et al. 2012), and this method was followed in this thesis. After obtaining the list of binding sites, the biological implications of these bindings can be preliminarily grasped by asking what are the

genomic annotations associated with the binding and thus get an idea about potential functions of peak regions. The considered annotations can be for example promoter annotations or phylogenetic conservation. Binding sites may also be used for de novo discovery of candidate DNA motif which may act as molecular signatures for a binding event.

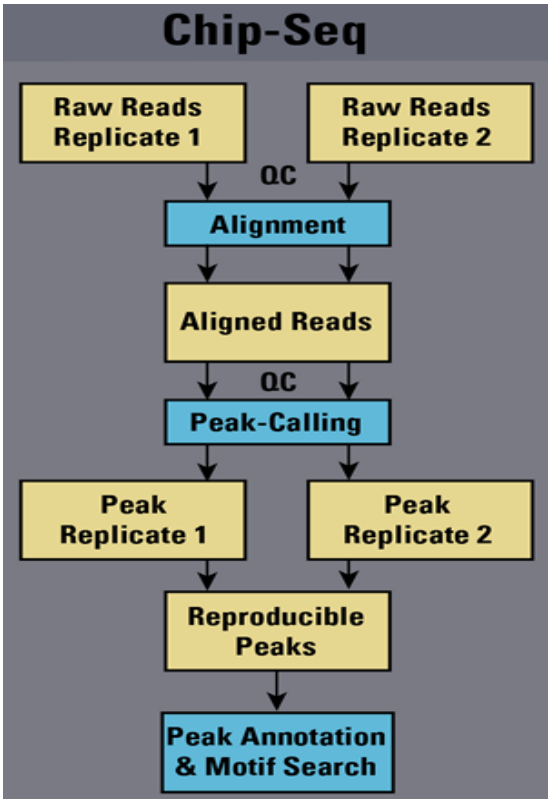


Figure 2.10: An outline of ChIPseq analysis pipeline with two biological replicates.

2.8.2.1 Commands and data types

As with RNAseq analyses, ChIPseq analyses were completed on Linux Ubuntu 12.04 LTS (64-bit) run on remotely-accessed computer with 126 GB of RAM memory. The necessary programs were installed as specified by their developers.

i) Reads alignment to reference genome and initial QC

As it was the case for RNAseq samples, ChIPseq raw FASTQ file were first quality evaluated using FastQC. Reads were aligned to the human genome version hg19 using the BWA (alignment via Burrows-Wheeler transformation) algorithm version 0.6.1-r104 with default settings (Li and

Durbin 2009). Duplicate reads were retained. Reads aligning to more or equal to two locations were held as well, with a single location decided randomly. The general commands used for the alignment are shown below and `prefix` is the bwa-indexed reference genome. `Bwa samse` command was run after `bwa aln` command using the `sai` file generated in the first step as argument in the second step.

```
bwa aln prefix in.fastq

bwa samse prefix in.sai in.fastq
```

The information about the alignment was obtained with Qualimap version 2.1. The general command used for that purpose was `qualimap bamqc –bam <bam_file>`. The number of reads aligning to multiple locations in the genome (multi hits rate) was obtained by repeating the alignment with Bowtie (Langmead et al. 2009) using `bowtie –m 1 –S –q <path_to_Bowtie_indexed_genome>/genome_prefix in.fastq out.bam` command. The `–m 1` parameter specifies that reads aligning to more than one location are excluded and the output from the alignment specifies the number and proportion of reads of this category.

ii) Read duplicate removal

Read duplicates were removed prior to further analyses due to their artefactual origin during sequencing on NGS platform. Although duplicated reads are removed automatically by the peak caller, duplicates were removed from BAM alignment files in order to draw average plots for different genomic features or genes. Read duplicates were removed using `samtools` with a command shown below.

```
samtools rmdup –s in.bam out.bam
```

iii) Peak calling

INTS12 peak calling was performed using the second generation of model-based analysis of ChIP-Seq (MACS), i.e. MACS2 (Zhang et al.

2008). ChIP samples were compared to input control when determining signal over background noise. Peak calling was performed with a multiple comparisons corrected P value of less than 0.05 considered as significant. Larger dataset of the two submitted BAM files (i.e. one belonging to ChIP sample and the other belonging to input control) was scaled down towards the smaller dataset. Generated fragment pileup signal was normalized per million reads and therefore was normalized for library size. The general command used is shown below.

```
macs2 callpeak -t ChIPsample.bam -c inputControl.bam
-n <name_of_prefix> --outdir <name_of_out_dir> -f BAM
-g hs -q 0.05 -B --SPMR --call-summits
```

iv) Creation of input-and-library-normalized WIG files

Alignment BAM file is too large to be handled by downstream visualization programs. Therefore, is has to be converted into WIG files which contain signal information without read sequence identity significantly reducing file size. Before a WIG coverage file was created, a bedGraph coverage signal file was made using MACS2's bedGraph compare function (`bdgcmp`) with fold-enrichment-above-control normalization to the input sample and total library size. This is possible because during the MACS2 peak calling, the ChIP signal track was generated with library normalization. This was achieved using the general command shown below. In it, `<name>_treat_pileup.bdg` and `<name>_control_lambda.bdg` were generated during MACS2 peak calling procedure while `-m FE` indicates that normalization is based on fold enrichment.

```
macs2 bdgcmp -t NAME_treat_pileup.bdg -c
NAME_control_lambda.bdg -o <signal_track>_FE.bdg -m
FE
```

The above command generated a bedGraph file which was subsequently converted into WIG track at 100bp resolution using signal step with no missing data algorithm. Thus, areas with zero coverage

were written in the output file. This was achieved using Perl language script `bedgraph_to_wig.pl` written by Sebastien Vigneau (https://sebastienvigneau.wordpress.com/2014/01/10/bigwig-to-bedgraph-to-wig/).

v) Peak annotation analysis

Peak regions in BED format were annotated using the HOMER toolkit (Heinz *et al.* 2010) and the Cis-regulatory Element Annotation System (CEAS; Shin *et al.* 2009). HOMER annotation was implemented when peak file was uploaded onto online ChIPseek platform which was used as the ChIP data visualization and manipulation tool (Chen et al. 2014). The CEAS algorithm was implemented not only to precisely annotate the peaks but also to investigate the genome wide distribution of these peaks in comparison to genome-background distribution, which is calculated from the WIG signal and BED region data, and to investigate the genome-wide average INTS12 binding profiles. Therefore, the WIG file had to be inputted into the CEAS programme together with BED regions. The general CEAS command is shown below. In it, `-g hg19.refgene` is SQLite3 database file with pre-compiled UCSC hg19 genome annotation (downloaded from http://liulab.dfci.harvard.edu/CEAS/download.html), `-b` parameter specifies MACS2 outputted BED binding sites, `-w` parameter specifies a continuous input and library size normalized WIG signal, `--bg` parameter indicates estimation of genome background from regions with WIG signal coverage outside binding sites, `--pf-res` parameter specifies the signal resolution in WIG file which in this case is 100 bases.

```
ceas -g hg19.refgene -b NAME_peak_regions.bed -w NAME_signal_track.wig --bg --pf-res 100
```

vi) Average binding plots, heatmaps and motif analyses

The average profiles for all the genes were generated as part of CEAS pipeline. However, class specific average gene profiles were generated with ngs.plot (Shen et al. 2014) using BAM alignments after duplicates

removal. The plots were drown with comparison to input control. The command depends on whether the visualization is centred on TSS or entire gene body and whether an entire gene list or subsets of genes are to be plotted. Motif enrichment analysis was performed with MEME suit (Ma et al. 2014).

vii) Other analyses

Other ChIPseq analyses were performed details of which are specified in results chapters.

2.8.3 Pathway analyses using Gene Set Enrichment Analysis approach

Pathway analysis using INTS12 knockdown RNAseq data was performed in order to aid functional hypothesis generation. In this case what is meant by pathway analysis is identification of physiological pathways that are dysregulated following the experimental manipulation. GSEA method instead of ORA analysis approach was utilized for this purpose due to mentioned limitations of the latter approach. The advantages and disadvantages of GSEA and brief description of the pathway enrichment testing method was described before (see section 1.7.4). The utilized gene ranking algorithm was signal-to-noise based on the following formula:

$$\frac{mean_{conditionA} - mean_{conditionB}}{StandardDeviation_{conditionA} + StandardDeviation_{conditionB}}$$

Thus genes were ordered based on magnitude of change in gene expression normalized by variability in expression.

2.8.3.1 GSEA RNAseq workflow

Traditionally GSEA was used using microarray differential gene expression analyses. However, it is applicable to any quantitative high throughput gene expression method. Therefore, GSEA RNAseq workflow was developed and is outlined below. The workflow makes use of self-written scripts using Python programming language. A number of computational challenges were encountered in the analysis, including

the need to separate genes amalgamated into single loci during transcriptome assembly and dealing with genes occurring multiple times in the expression dataset.

2.8.3.1.1 Determination of normalized gene expression on individual sample basis

GSEA requires the generation of gene expression spreadsheet where columns represent the sample and rows represent the gene. Hence, the first step in the analysis is the estimation of FPKM expressions for each locus in the human genome for each individual sample. This was achieved by leveraging (i) `accepted_hits.bam` per sample read alignments generated by TopHat, (ii) `merged.gtf` predicted genome annotation file generated by Cuffmerge. Alignments and genome annotation file were input into Cuffnorm on sample basis using the general formula shown below. The `sample_sheet.txt` file that is used in the command contains paths to the alignment files and corresponding sample names.

```
cuffnorm -o sample -p 5 --use-sample-sheet
<path_to_predicted_annotation>/merged.gtf
sample_sheet.txt
```

Cuffnorm counted the number of reads that aligned to each locus and normalized the count of reads yielding FPKM expression values for each feature.

2.8.3.1.2 Preparation of expression dataset for GSEA

During transcriptome assembly novel genes were discovered. These were excluded from subsequent GSEA because they do not have gene names as of yet and hence cannot be searched for in pathway analyses. A tab delimited text file was created containing FPKM data per gene locus per biological replicate samples. Separate files were made for each condition. Then Cuffnorm generated gene attribute file was used to assign gene names to each locus id. At this stage the dataset contains multiple genes amalgamated into single loci and the same gene names occurring multiple times in the dataset.

A program named `gene.perXLOC_exp_parser.py` was written to parse and prepare files for GSEA. It reads the gene names in the second column and does the following: splits genes into separate rows if they happen to occur in the same raw separated by comma and assigns the same expression values per each separated gene because they share the same locus id. The rows that do not require separation are simply outputted. Crucially, the program is capable of ignoring the newly discovered genes.

Newly generated expression files were merged into one expression matrix where the first column represents gene name while subsequent columns represent individual samples. This file had to be further processed in order to identify gene names that occurred multiple times. Cuffmerge not only assigns different gene names to the same locus due to close proximity and read coverage over these genes, but also may assign the same gene name to different loci due to distinct TSS of the same gene.

In order to identify gene names that occurred multiple times and create an updated expression matrix with single FPKM per gene a new program called `expression_Table_parser.py` was written. This program identifies genes occurring multiple times in the dataset and sums their respective FPKM values. Summation of expression values is preferable because averaging may significantly skew the data if there is a big difference in FPKM values across loci of the same gene. On the other hand using maximum expression only may result in a loss of important data. The limitation of summation approach is that genes that have multiple locus ids assigned, may be over-represented due to the possibly that the same reads may be counted additional times: the same reads could be contributing to the FPKM values of different loci belonging to the same gene. However, this bias is going to be introduced in all the samples because the same predicted annotation is used in all the samples and thus should cancelled out during differential gene expression analysis.

2.8.3.1.3 Identification of pathway gene names and further expression data preparation

Although the prepared expression dataset can be used in GSEA, it is advisable to further pre-process the file to only include the genes that are present in the pathway database to be investigated for enrichment in the ranked list of differentially expressed genes. In GSEA procedure, if a gene belonging to the pathway is not present in the list then penalty is applied onto the enrichment score (see section 1.7.4) resulting in its reduction. But it is unreasonable to reduce the enrichment score due to the absence of gene in ranked gene list if the gene is not present in the pathway database.

In order to remove those genes from expression matrix that were not present in the pathway database, two programs were written. The first program named `pathway_database_parser.py` takes pathway database information and creates a text file with a list of genes that occur in the entire database. If a gene is part of more than two sets then it is outputted only once. The second program named `genes_extraction_from_my_exp_Table.py` is used to pull out pathways database gene names from the generated gene expression dataset with all the associated expression values. Having generated the final gene expression matrix, the GSEA phenotype file was prepared as required by algorithm developers.

2.8.3.1.4 GSEA analysis

Important GSEA parameters are gene set maximum size and gene set minimum size. If a set would have a size larger or smaller than the specified parameters it would be excluded from the analysis. As all the pathways are to be investigated it was necessary to identify the largest and smallest gene sets. In order to accomplish that another program named `max_min_pathway_genes_counter.py` was written. This program takes gene sets file as input and prints out the maximum and minimum gene set sizes. As of December 2014, the maximum pathway size in curated pathways database (http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C2) has

1,972 genes while minimum pathway size has 5 genes and these parameters were set accordingly. Significance of pathway enrichment was calculated using 1000 gene list permutations. Pathways with Benjamin and Hochberg corrected P-value (Benjamini and Hochberg, 1995) below 0.05 were considered significant.

2.9 Functional assays

Functional assays following gene knockdown were selected in a systematic approach based on the identified dysregulated pathways. Thus the outcomes of interest were *in vivo* protein synthesis and cell proliferation.

2.9.1 Measurement of radioactive amino acid incorporation into protein by a filter-paper disk method

The purpose of radioactive amino acid incorporation assay is to measure the incorporation of radiolabelled ^{35}S-methionine into newly synthesized protein. This method allows for a quantitative measurement of total protein synthesis and has the advantages of being performed without modification of the cells to be studied (Wong et al. 2010). However, its disadvantage lies in the inability to distinguish between the differences in rates of protein synthesis at different stages of translation such as initiation, elongation or termination (Esposito and Kinzy, 2014). Because protein synthesis measurement is affected by cell number it is critical to ensure similar cell densities between the conditions that are being compared. Thus, the number of cells seeded onto the reading plate was optimized to equilibrate the number of cell across the experimental conditions. Cell number is reflected by total protein concentration and therefore Bradford reaction (Bradford, 1976) is carried out together with protein synthesis measurement and its reading used in data normalization.

2.9.1.1 Protein synthesis assay

The effect of INTS12 knockdown on protein synthesis was examined 120h post initiation of RNAi (see section 2.2.2.3) in collaboration with Dr Cornelia de Moor based at the University of Nottingham. Because cell

counts (see section 2.9.2) revealed one cell cycle difference over 120h period in INTS12 silenced cells versus not silenced cells, cell density used for un-transfected and scrambled D-siRNA conditions was 50% less than in both INTS12 D-siRNA conditions in order to minimize the effects of cell density on protein synthesis measurement.

Experiment was performed on 24-well plate, where 6,150 cells were added onto un-transfected and scrambled D-siRNA wells while 12,300 cells were added onto the two anti-INTS12 D-siRNA conditions. After gene knockdown the culture medium was removed and washed with PBS twice, taking off residual PBS with a fine pipette tip. 10 µCi/ml Tran^{35}S label (Perkin Elmer, cat. num. NEG772002MC) was added in warm methionine and cysteine free DMEM with glutamine and placed in the 37ºC incubator for ~15 min. Then, the radioactive medium was quickly removed with a P1000 pipette, washed three times with cold PBS and the plate was placed on ice. The remaining PBS was removed with a P200 pipette and 50µl of 1x passive lysis buffer (Promega, cat. num. E1941) was added per well. The buffer was swirled to distribute the lysis buffer evenly and no scraping was applied. At this point, the 24-well plates were frozen at -20ºC for several days. Rectangles of Whatman® 3MM paper (Fisher Scientific) were cut of roughly 8mm by 12mm. Prepared paper was numbered with pencil according to the experimental condition. The piece of paper was stuck through a pin and mounted on the cardboard well clear of the surface. A 96-well plate suiTable for spectrophotometric absorption reading at 595nm was obtained. Four replicates of 10µl of 1x passive lysis buffer were put to reliably obtain the background reading. The 24-well plates were taken out of -20ºC freezer and left to thaw out. Holding the 24-well plate at an angle, 10 µl of the lysates were collected and spot on the appropriate filter. The same tip was used to collect another 10µl and was put onto 96-well plate for Bradford assay readings as appropriate. Then, 250µl of Coomassie Bradford Protein Assay Reagent (Thermo Scientific, cat. num.1856209) was added to each of the wells in the 96 well spec plate and read at 595 nm. The results were exported to Excel file and saved. For up to forty eight filters, 250ml of 10% (w/v) trichloroacetic acid (TCA) solution was

prepared in water from a 100% stock. A pinch of cysteine and methionine was added to 10% (w/v) TCA and swirled. 125ml of this solution was 1:2 diluted in water to obtain a final 5% (w/v) TCA solution. Whatman filters were dropped into the 125ml of 10% (w/v) TCA solution in a glass beaker and swirled for ~2-3 min. The 10% (w/v) TCA was poured into a waste bottle. 125 ml of the 5% (w/v) TCA solution was poured over the filters and swirl for another ~5 min. 5% (w/v) TCA solution was disposed of in waste bottle. 125 ml of the 5% (w/v) TCA solution was poured over the filters again, swirled for a ~2-3 min and the 5% (w/v) TCA was poured into a waste bottle. The filters were washed three times with 50ml methylated 96% ethanol taking care to wash the whole beaker. The filters were pinned again to dry for ~1h. The emitting radiation was measured on the scintillator. A scintillation counter is an instrument for detecting and measuring ionizing radiation by using the excitation effect of incident radiation on a scintillator material, and detecting the resultant light pulses. 2 ml of Ecoscint liquid (National Diagnostics, nat. num. LS-273) was placed into counting bottles. The dry filters were put in and radioactivity counts per methionine (CPM) counting were performed for 2 min (Esposito and Kinzy, 2014; Wong et al. 2010).

2.9.1.2 Protein synthesis calculations

Bradford assay background was subtracted from sample readings. For each replicate, the CPM readings were divided by the background corrected Bradford measurements yielding a measure of incorporation per amount of total protein (defined as “specific activity”).

2.9.2 An assessment of proliferative capacity

Proliferative capacity was assessed by comparing cell counts at the beginning and at end of the experiment. Cells were subjected to INTS12 silencing using 120h protocol (see section 2.2.2.3). After 120h cells were washed with cell culture grade PBS, treated with trypsin/EDTA at 37^{o}C for ~10min to allow all the cells to detach and were re-suspended in 1ml of culture media. Samples were coded and mixed to perform counting without knowledge of the condition. Conditions were decoded later. Cell counts were performed on haemocytometer (see section 2.1.1) in

triplicate per each condition, averaged and total cell counts estimates derived accordingly.

3. *In silico* approaches and methods development

3.1 Introduction

As described in detail in Chapter 1, there is substantial genetic evidence implying the existence of genetic variation at the *INTS12*/*GSTCD*/*NPNT* 4q24 haplotype as a contributor to variation in lung function parameters and risk of developing COPD (Repapi et al. 2010, Hancock et al. 2010, Castaldi et al. 2011, Wain et al. 2015). Due to the physical linkage of genes in close vicinity, these GWAS are not capable of identifying causal genetic variants and genes. Moreover, correlation does not imply causation (Aldrich, 1995) and therefore additional layers of information are required to prioritize likely genes influencing the considered phenotypes. Claims for causality require functional studies where candidate genes or single polymorphisms are experimentally manipulated with a demonstrable effect on either the lung function trait directly, or on molecular pathways of relevance to pulmonary health. This Chapter sets out to (1) prioritize the likely gene at 4q24 locus whose variable expression contributes to lung function and/or COPD phenotypes, (2) investigate the type of molecular evolution of the identified candidate gene, (3) assign putative functions to this gene via a homology searches, (4) explore the functional annotation of the lung function implicated region, and (5) develop the experimental tools to study the candidate lung function *in vitro.*

3.1.1 Candidate lung function gene prioritization at 4q24 locus

If the leading hypothesis is that altered levels of gene expression in specific allele carriers are responsible for population differences in lung function, then the critical point is whether SNPs associated with lung function are also predicting the candidate gene expression (i.e. whether trait correlated alleles are eQTLs for a particular candidate gene). This is the leading hypothesis for the *INTS12*/*GSTCD*/*NPNT* locus because although premature stop codon variants have been identified (Table 1.3; NCBI gene bank) no non-synonymous mutants are known to exist for *INTS12*, *GSTCD* or *NPNT.* These genes are central in the GWAS signal for lung function and risk of COPD because of their linkage to 4q24 sentinel SNPs used in association studies.

Obeidat et al. have argued that out of *INTS12*, *GSTCD* and *NPNT*, *INTS12* is the most likely gene whose expression contributes to phenotypic differences in lung function due to widespread *cis*-eQTL effects upon *INTS12* expression observed in diverse tissue types (Obeidat et al. 2013). Importantly, INTS12 *cis*-eQTL SNPs are also associated with lung function in SpiroMeta-CHARGE studies (Repapi et al. 2010, Hancock et al. 2010). However, no significant *INTS12 cis*-eQTL effect was detected in lung tissue using a 1,111 lung specimens microarray dataset (Hao et al. 2012). On the other hand, the same dataset provides evidence for *cis*-eQTL on *NPNT* expression. Although some SNPs within the gene bodies of *INTS12* and *NPNT* are in relatively weak linkage disequilibrium (r^2<0.2; Repapi et al. 2010, Hancock et al. 2010) the two genes have been suggested to represent independent lung function signals (Wain et al. 2015).

Therefore, it may be more appropriate to talk about *INTS12*/*GSTCD* and *NPNT* as separate loci and as such, although in the light of the Hao et al. study (Hao et al. 2012) *NPNT* is worth pursuing for functional studies, the nature of the *INTS12*/*GSTCD* signal still requires an explanation and gene prioritization using lung and other eQTL resources. Importantly, it is of pivotal importance to test whether SNPs correlating with poorer lung function also correlate with lower or higher gene expression and vice versa. The lack of observable effect in the data set used by Hao et al. could be due to inherent heterogeneity in microarray datasets (Zhao et al. 2014). As RNAseq has been demonstrated to outweigh microarray in terms of technical reproducibility (Zhao et al. 2014), lung RNAseq eQTL dataset would be advantageous to look for evidence of SNP-*INTS12*-*GSTCD* expression correlations. Moreover, both genes are expressed in a range of human airway cell types making them good candidates for further exploration.

3.1.2 Computational molecular evolution and homology searches

Natural selection plays the fundamental role in shaping the genetic variation on a population level and in speciation. Computational

molecular evolution is the science of evolution of DNA, RNA and protein molecules and studies the causes and mechanisms of molecular adaptions. This discipline has numerous applications and in the context of this thesis some of its methods were used to explore the modes of evolution of candidate lung function gene. Additionally, homology searchers were performed to assign putative gene functions. These approaches have traditionally been used to explore putative biological roles based on the principles of "form dictates function" (Gish and States 1993) and "evolutionary conservation indicates importance".

3.1.2.1 dN/dS ratio test

A set of aligned homologous ORF DNA sequences can be used to infer whether a protein molecule has been evolving positively, negatively or neutrally (Mugal et al. 2013). This is achieved by comparing the rate of non-synonymous changes per non-synonymous sites (dN) relative to rate of synonymous changes per synonymous sites (dS). Figure 3.1 below demonstrates the idea of non-synonymous and synonymous sites due to genetic code redundancy. For example in CTA codon encoding leucine the third position is considered a synonymous site because whether nucleotide A is mutated into C, G or T it will always encode the same amino acid. On the other hand the second position is considered a non-synonymous site as whether nucleotide T is mutated into G, C or A it will always produce a different amino acid. The first position of this codon is considered $\frac{1}{3}$ synonymous and $\frac{2}{3}$ non-synonymous as mutation of C into A or G results in isoleucine and valine respectively whereas mutation into T results in leucine. The total number of non-synonymous or synonymous sites is simply the summation of their respective counts including the fractions. Therefore, e.g. if out of nine bases six are non-synonymous then we can expect 67% probability of random amino acid change.

The numbers of observed non-synonymous and synonymous changes are determined directly from the sequence and thus dN and dS express the significance of non-synonymous and synonymous changes relative to number of changes expected in completely stochastic system (i.e. a

system where changes occur randomly). If a particular stretch of ORF DNA sequence was evolving neutrally during the time between ancestral to the modern versions of the protein then the rate of silent changes should equal the rate of non-silent changes hence dN/dS ratio would be ~1. A dN/dS ratio greater than 1 (dN/dS > 1) implies that there has been more non-synonymous changes than synonymous changes and therefore there has been evolutionary pressure to escape from the ancestral sequence. This typically occurs when a protein adapts to a new environment and advantageous mutation arises and spreads in the population. A dN/dS ratio smaller than 1 (dN/dS < 1) suggests there has been more silent mutations than protein changing mutations, i.e. the protein was constrained due to negative selection pressure. This typically occurs when a protein is required to maintain its function.

dN/dS test can be performed in a statistically stringent way by, e.g. computing the P-value of data given the hypothesis of neutral evolution (Nei and Gojobori, 1986). Critically, when the test is performed over the whole sequence it could result in underestimate of positive selection as variety of domains that constitute the protein may undertake different functions. Thus it is possible to compare regions within the protein, which is known as sliding dN/dS test, or test evolution on a codon-by-codon basis which requires ancestral sequence reconstruction (Pond and Frost et al. 2005). The set of sequences used for dN/dS test ought to share a common ancestral gene or else the results would be spurious.

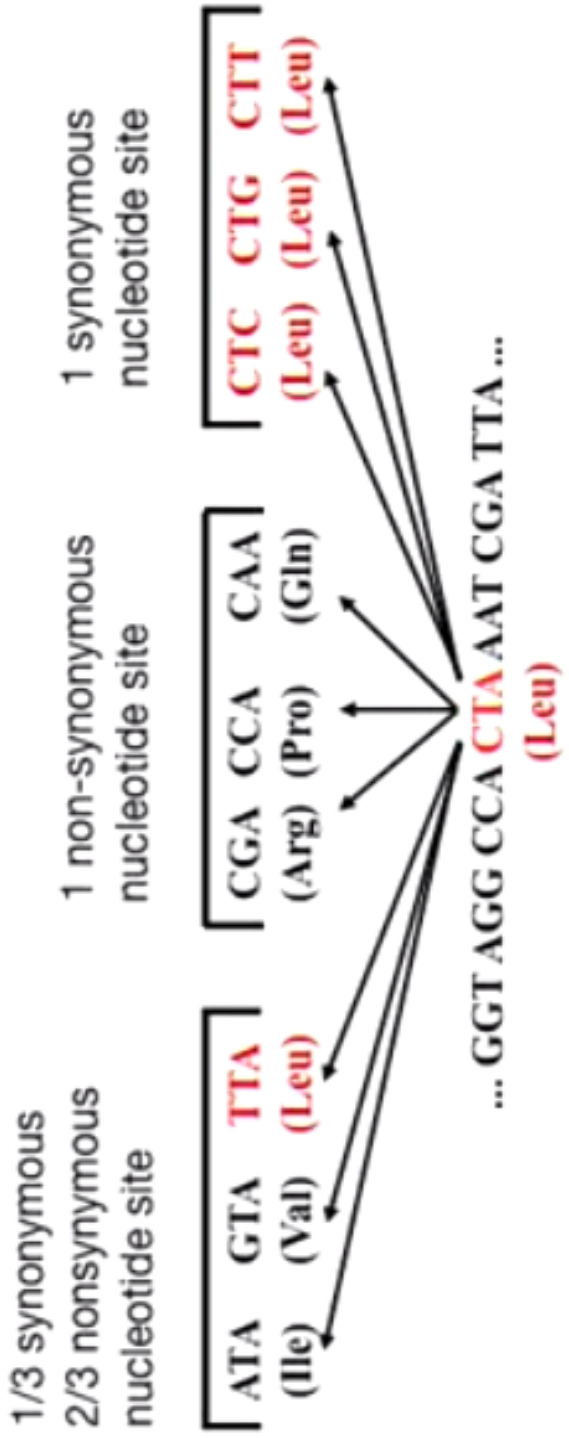


Figure 3.1: Non-synonymous and synonymous sites due to the redundancy of genetic code. As there are 20 naturally occurring amino acids and genetic code contains 61 amino acid encoding codons plus 3 stop codons, some codons are redundant coding for the same amino acid. Based on that some codon sites are defined as non-synonymous and some are defined as synonymous depending on whether their mutation alters the encoded amino acid. A site may be partially synonymous and partially non-synonymous if some changes are protein changing and vice versa.

3.1.2.2 Homology searches

A protein or DNA sequence of unknown role having significant similarity to other sequence of known function is likely to have it as well, because biological activities of the protein depend on its sequence which determines its function. However, it is important to remember that such an observation does not guarantee a particular functionality because among the duplicated genes there is a level of redundancy allowing for otherwise detrimental mutations to be accumulated. As other copies acquire changes, new and different functions can be created (Ohta 2006).

Although these similar proteins may or may not be homologues because non-ancestral proteins could have undertaken a convergent sequence evolution, the principle still holds true. Whether the proteins are truly homologous is determined arbitrary through setting a threshold percent identity or similarity and expected value (E-value) of the BLAST search (Gish and States 1993). With sufficiently significant E-value from BLAST search limited to the same species it may be possible to identify paralogs and hence a set of genes belonging to the same family. Paralogs and non-homologous highly similar proteins are likely to perform similar function.

3.1.3 Aims and Objectives

The aim of this chapter is to "bridge the gap" between a purely observational association to a systematic *in silico* gene prioritization strategy for further functional investigations. This will be achieved firstly by leveraging publically available RNAseq-based lung *cis*-eQTL dataset from the Genotype-Tissue Expression project in which *cis* window was defined as 1MB around the gene's TSS in both directions (Lonsdale et al. 2013). Lung function SNPs predicting the relevant gene expression were analysed on Broad Institute's HaploReg v4.1 (see section 1.5.5) for functional annotation. Similar protein identification was undertaken to assign candidate molecular function to nominee lung function gene. dN/dS test using diverse metazoan sequences was run to understand the molecular evolution of the gene and identify functionally important

protein domains. Finally, the necessary tools to study the prioritized gene function were optimized.

3.2 Lung eQTL analyses

The most recent studies in the UK Biobank population (Wain et al. 2015) have shown that there are at least 3 independent association signals within 4q24 locus: one located over the gene for Tet Methylcytosine Dioxygenase 2 (*TET2*), one over the gene *NPNT*, and a third peak situated over the genes *GSTCD* and *INTS12*. Upon closer examination of originally published regional plots of this region (Figure 1.2) it appears that the linkage between *INTS12*/*GSTCD* and *NPNT* as well as *TET2* is less than 0.2 (i.e. r^2<0.2) reaffirming the independent associations observed in UK Biobank population. Thus each of the three signals ought to be considered separately, each requiring a functional explanation of the genetic contribution of candidate genes to lung function phenotypes. The focus of this thesis is the signal observed at *INTS12*/*GSTCD* and therefore the question is which of the two genes, is the most likely contributor to lung function.

As mentioned, the leading hypothesis is that altered levels of these genes in specific allele carriers are responsible for population differences in lung function. Although these effects may be mediated in *trans*, the more likely scenario is that SNPs at the associated locus control the near-by gene expression and therefore *cis*-eQTLs are the focus of this Chapter. In order to answer the above question, a lung specific RNAseq based dataset was used to test whether lung function SNPs are predictive of *INTS12* or *GSTCD* expression. As discussed in Chapter 1 it is important to use a relevant tissue or cell type as patterns of gene expression differ significantly between various tissues and cell types and hence the reason for using the lung datasets in these *in silico* explorations. Thus although it is possible that e.g. immune cell defect or a blood vessel defect through development may be relevant, the focus of this Chapter is in determining lung specific effects.

3.2.1 Lung function SNPs significantly predict *INTS12* but not *GSTCD* expression in the relevant tissue

In the lung eQTL GTEx resource (n=278), there were 248 SNPs at or near 4q24 that were significant *cis*-eQTLs for *INTS12* expression after multiple comparisons correction. Among these, 30 SNPs showed significant association for lung function in the SpiroMeta consortium study (Table 3.1; Repapi et al. 2010). On the other hand, none of the variants at or near 4q24 showed significant association with *GSTCD* expression (Table 3.1). Upon inspection it became apparent that SNPs correlated with lower FEV_1 were associated with lower *INTS12* expression. In favour of this observation is a similar result presented by Obeidat et al. (Obeidat et al. 2013). However, eQTL datasets leveraged by them were based non-lung tissue profiles, as no significant correlation was observed between lung function SNPs and either *INTS12* or *GSTCD* expression in whole lung microarray dataset (Hao et al. 2012). Therefore, the herein presented data are of added value because are based on the lung specific gene expression.

SNP	FEV_1 P-value	INTS12 eQTL P-Value	INTS12 eQTL FDR	INTS12 effect size	GSTCD eQTL P-value	GSTCD eQTL FDR	GSCD effect size
rs11732650	6.83E-09	3.33E-07	0.000397993	-0.53	0.989632	1	0.00
rs11722225	7.08E-09	3.33E-07	0.000397993	-0.53	0.989632	1	0.00
rs11726124	6.63E-09	3.33E-07	0.000397993	-0.53	0.989632	1	0.00
rs11728716	8.44E-09	3.33E-07	0.000397993	-0.53	0.989632	1	0.00
rs17036090	3.84E-08	1.48E-06	0.000397993	-0.51	0.947098	1	0.01
rs11735851	1.90E-09	1.71E-06	0.000397993	-0.51	0.84487	1	0.02
rs17036225	3.33E-09	1.72E-06	0.000397993	-0.51	0.846924	1	0.02
rs11736859	2.86E-09	1.73E-06	0.000397993	-0.51	0.847786	1	0.02
rs11727745	5.47E-09	1.73E-06	0.000397993	-0.51	0.847958	1	0.02
rs10516528	6.27E-09	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs17036139	1.25E-09	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs11727189	3.38E-09	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs11728044	1.95E-09	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs11733225	2.34E-09	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs11733654	0.0358	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs10516525	1.44E-09	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs11724839	1.79E-09	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs10516526	6.67E-10	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs17036142	1.11E-09	1.73E-06	0.000397993	-0.51	0.850201	1	0.02
rs12374256	1.88E-09	2.97E-06	0.000658031	-0.52	0.816732	1	0.03
rs11097901	6.32E-09	4.58E-06	0.000953622	-0.47	0.849082	1	0.02
rs7676975	6.75E-09	4.93E-06	0.000953622	-0.43	0.821606	1	0.02
rs10050333	7.22E-09	4.97E-06	0.000953622	-0.43	0.819258	1	0.02
rs10050159	7.23E-09	4.97E-06	0.000953622	-0.43	0.819258	1	0.02

Table 3.1: Lung function 4q24 SNPs from the SpiroMeta study (Repapi et al. 2010) are lung *cis*-eQTLs for *INTS12* but not *GSTCD* expressions. Effect size is defined as the slope of linear regression line relative to reference allele normalized as an expression of 1. Data obtained from Genotype-Tissue Expression project (Lonsdale et al. 2013).

3.2.2 In lung tissue *INTS12* expression is higher than *GSTCD* expression

Interestingly, out of 51 tissues *INTS12* expression in the lung is among the top 13 highest tissue expressions (Figure 3.2). On the other hand, lung *GSTCD* expression is among the top 19 tissue expressions implying its higher expression in other tissue types (Figure 3.3). Moreover, the expression of *INTS12* in the lung is five times higher in comparison to *GSTCD* (n=278; P<0.0001, Figure 3.4). To provide validation for the latter observation the analysis was performed leveraging an independent RNAseq lung profiling GEO dataset (Kim et al. 2015) in which *INTS12* expression also appears to be five times higher in comparison to *GSTCD* (n=91; P<0.0001, Figure 3.5).

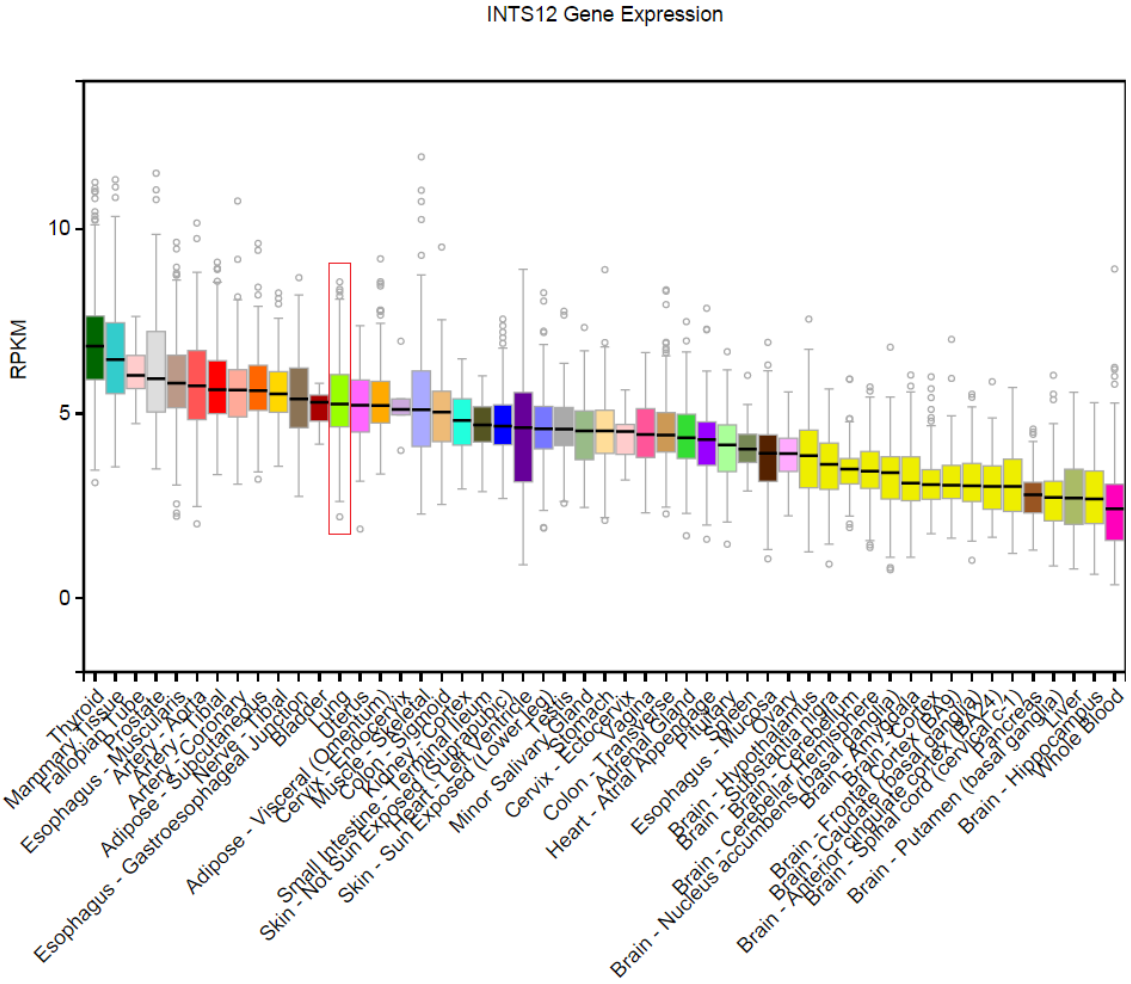


Figure 3.2: *INTS12* expression in the lung is among top 13 tissues with highest expressions. Box plot belonging to the lung data is highlighted in the red box.

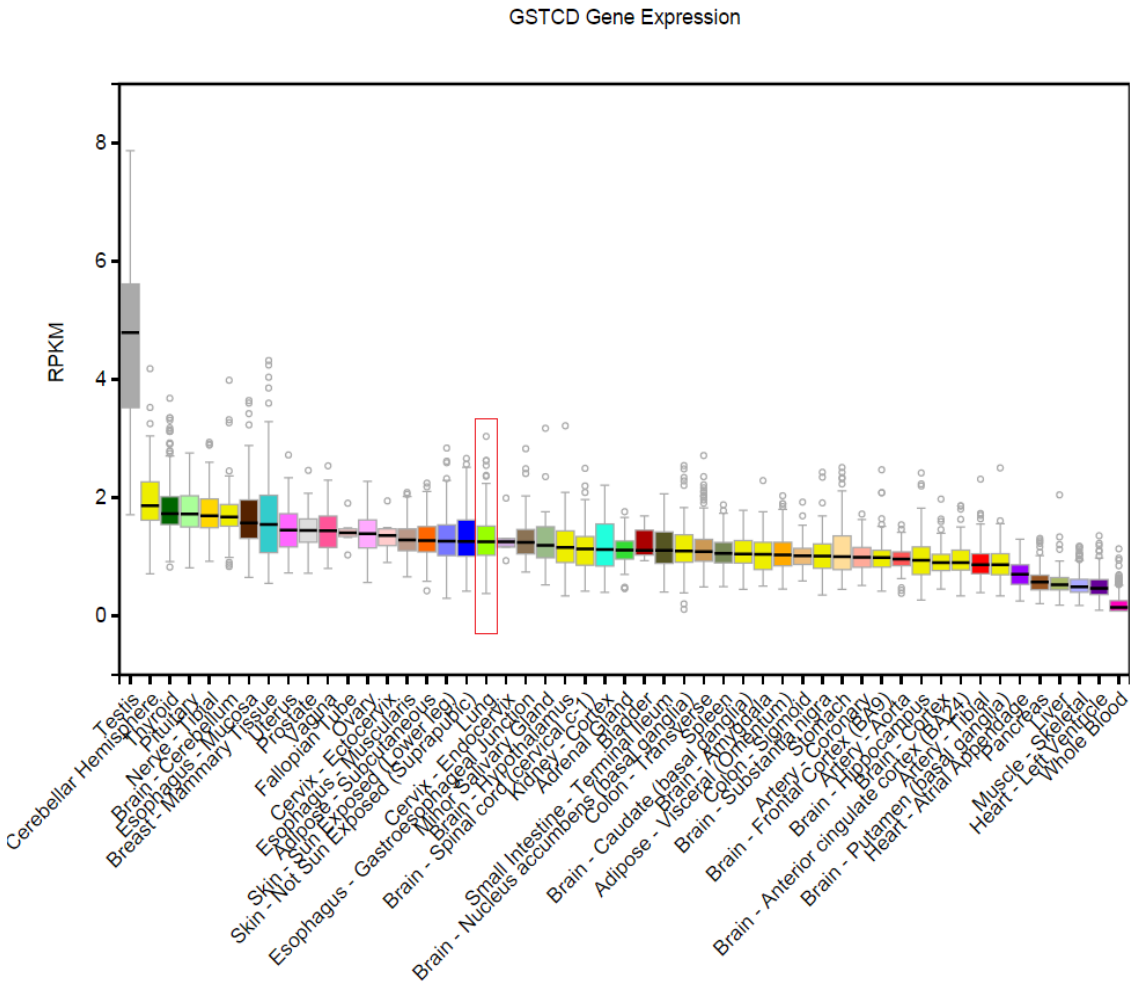


Figure 3.3: *GSTCD* expression in the lung is among top 19 tissues with the highest expression. Box plot belonging to the lung data is highlighted in the red box.

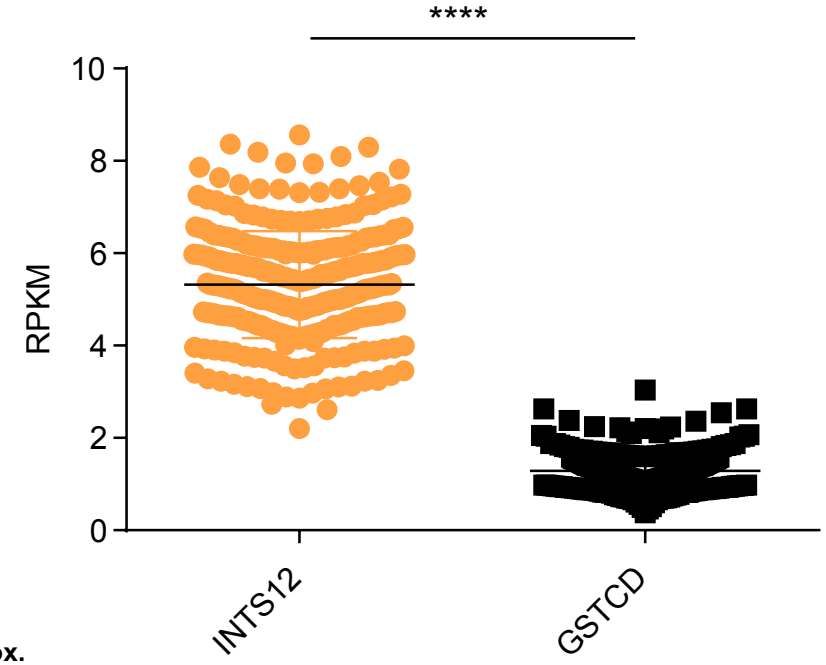


Figure 3.4: Lung *INTS12* expression is five times higher than *GSTCD* (P<0.0001; Mann-Whitney) in Genotype-Tissue Expression dataset.

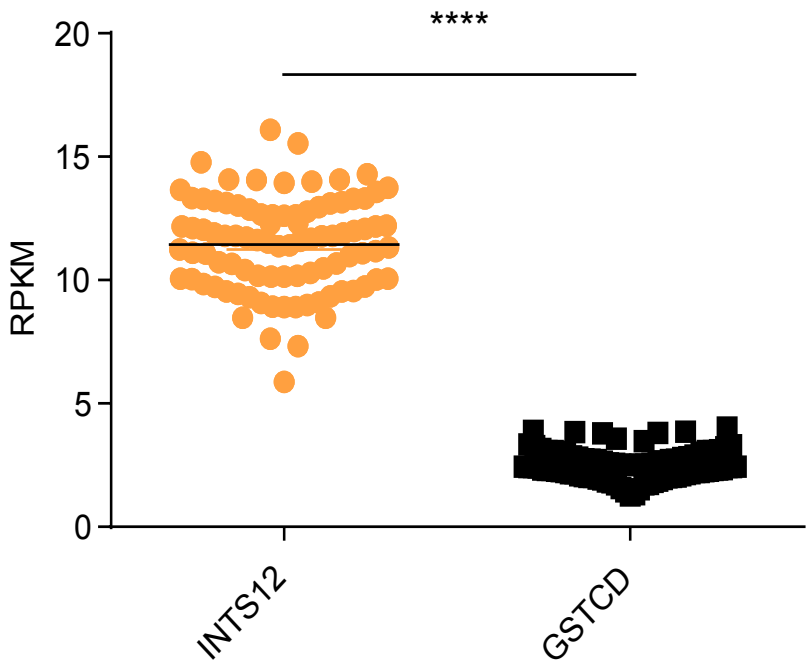


Figure 3.5: Lung *INTS12* expression is five times higher than GSTCD (P<0.0001; Mann-Whitney) in Kim et al. dataset (Kim et al. 2015).

3.2.3 Summary of lung eQTL prioritization strategy

The initial eQTL analyses in multiple non-lung tissues found the strongest evidence supported the hypothesis that altered expression of *INTS12* underlies association signal for lung function at 4q24 (Obeidat et al. 2013). However, no subsidiary evidence for this conclusion was produced using a lung specific gene expression utilising microarrays (Hao et al. 2012). To explore this in more detail, a lung eQTL RNAseq-based dataset (Lonsdale et al. 2013) was used to further test the hypothesis that *INTS12* expression underlies lung function signal at 4q24.

The overall evidence from the RNAseq study appears to be in favour of this hypothesis. Lung functions SNPs significantly predict *INTS12* but not *GSTCD* expression (Table 3.1) and alleles associated with lower expression correlate with worse lung function. The discrepancy between lung microarray (Hao et al. 2012) and RNAseq (Lonsdale et al. 2013; Kim et al. 2015) datasets was probably due to inherently higher technical heterogeneity of the microarrays in comparison to RNAseq, potentially resulting in a loss of statistical significance in the array analyses. This inconsistency was observed despite the fact that the sample size in the

microarray study (n=1111, Hao et al. 2012) was five times larger than in the RNAseq study (n=278, Lonsdale et al. 2013). Moreover, Hao et al. array-based study is from "diseased" lungs of people undergoing surgery which could have been an additional source of heterogeneity. In contrast, the RNAseq eQTL dataset is largely from "healthy" individuals that died in unexpected circumstances, such as victims of road traffic accidents (Lonsdale et al. 2013). Overall, it is possible to say that out of *INTS12* and *GSTCD*, it is the former gene that seems to be the likely contributor to lung function variability and thus was prioritized for functional studies.

3.3 Lung INTS12 *cis*-eQTL focused exploratory analyses

Lung INTS12 cis-eQTL SNPs that are also genome-wide significant for lung function (Table 3.1) belong to the same haplotype. Based on the data obtained from 1000 genomes project, in the Northern and Western European CEU population all of them except rs17036142, rs10516528, and rs17036090 have r^2=1. These three variants are still in strong linkage with the rest of the SNPs (r^2>0.8, Figure 3.6). Due to their strong linkage it is not possible to ascertain which one is potentially causal in relation to *INTS12* expression. Thus the entire haplotype can be said to associate with both lung function and INTS12 levels.

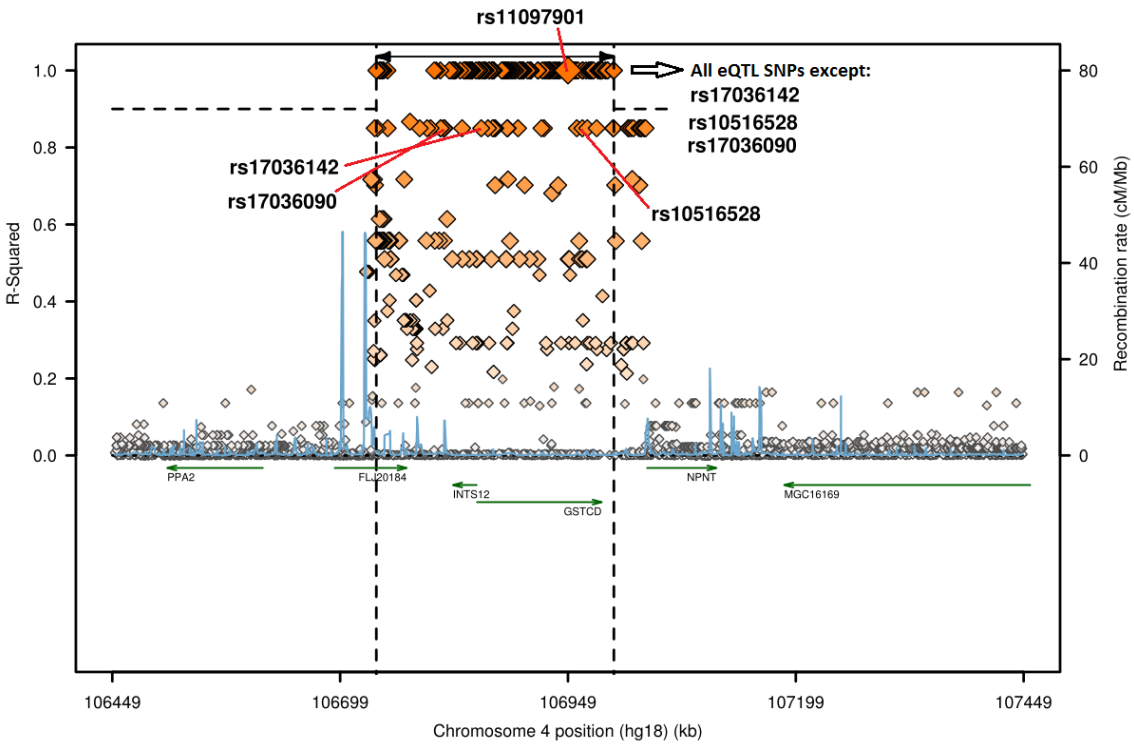


Figure 3.6: Regional LD structure of the INTS12 *cis*-eQTL SNPs. All the SNPs shown on Table 1 except the three indicated have r^2=1 and thus belong to the same haplotype associated with both lung function and *INTS12* expression. This regional plot was generated using Broad Institute's SNP annotation and proxy search tool (SNAP).

3.3.1 HaploReg analysis indicates potential regulation of expression effects of the INTS12 *cis*-eQTL SNPs

To provide initial functional translation of these SNPs, HaploReg tool (Ward and Kellis, 2012) was used with an aim to identify the likely functional variants and to test whether any of the alternative alleles are changing the gene expression by regulating TF DNA binding domains. 90% of the identified 228 INTS12 eQTL SNPs fall within regions enriched for epigenetic marks such as H3K4me1, H3K27ac. 14% of the SNPs are within accessible chromatin region. Crucially, more than 80% of the variants change the molecular signature of protein binding domains previously characterized as regulators of gene expression, such as forkhead box P (FOXP3), or ABI five binding protein (AFP1) (Table 3.2). Overall, these preliminary exploratory analyses give further credit to the hypothesis that lung function and INTS12 *cis*-eQTL SNPs are having an effect on the phenotype via regulation of *INTS12* expression.

SNP	ref	alt	Chromatin_Marks	DNAse	Motifs
rs11732650	G	C	E008,H3K27ac_Enh;E011,H3K2	E059	Crx_1;GR_disc5;Gsc;Obox3;Otx2;Pitx3;ZNF263_disc1
rs11722225	T	C	E003,H3K4me1_Enh;E010,H3K	0	CEBPB_known4
rs11726124	A	G	E003,H3K4me1_Enh;E010,H3K	0	.
rs11728716	G	A	E107,H3K4me1_Enh;E108,H3K	0	SP2_disc1;STAT_known1;STAT_known2
rs11735851	G	A	E006,H3K4me1_Enh;E096,H3K	0	AP-2rep;CEBPB_disc2
rs17036225	A	G	E015,H3K27ac_Enh	0	AFP1;Pou5f1_known2
rs11736859	C	T	0	0	.
rs11727745	T	G	0	0	Arid3a_2;Foxp1;HNF1_6;HNF1_7;Pax-4_5;Pou1f1_1;Pou5f1_disc2;Sox_15;Sox_2;Sox_4
rs10516528	G	T	E014,H3K4me1_Enh;E053,H3K	0	Brachyury_2;Cphx;Eomes;TATA_known3;TBX5_1;TBX5_3
rs17036139	G	A	E004,H3K9ac_Pro;E011,H3K9a	0	AP-3;Pou2f2_known1;Pou2f2_known10
rs11727189	G	T	E037,H3K27ac_Enh;E050,H3K2	0	ERalpha-a_disc4;GR_disc6;STAT_disc7
rs11731417	A	G	E061,H3K27ac_Enh	0	YY1_disc2
rs11727735	A	G	E001,H3K4me3_Pro;E002,H3K	E091;E120	Hoxa10;Hoxa5_2;Irf_known5;Maf_disc2;RFX5_disc3
rs11723225	C	T	E061,H3K27ac_Enh	0	Mef2_known4;NF-kappaB_disc3;STAT_disc7;TATA_disc7
rs10516527	A	G	0	0	IRC900814;Rhox11
rs11733287	G	A	E027,H3K4me1_Enh;E102,H3K	0	Foxj1_1;Foxj1_2;Foxq1;HNF1_7;Mrg_1;Mrg_2;RREB-1_2;Tgif1_2
rs11726569	A	G	E002,H3K4me3_Pro;E039,H3K	0	CDP_1;Fox;Gfi1b;HNF6;Hoxd8;Pbx-1_1;Pbx-1_4;Sox_16;Sox_3;TCF11::MafG
rs11728044	G	C	E013,H3K27ac_Enh;E039,H3K2	0	Cdc5;Nkx3_1
rs11733225	C	G	E027,H3K4me1_Enh;E102,H3K	0	Rhox11
rs11733654	C	A	E004,H3K4me1_Enh;E011,H3K	0	HNF1_7;SRF_known3
rs10516525	T	C	E061,H3K4me1_Enh;E119,H3K	0	Hoxa10;Hoxb13;Hoxb9;Zfp105
rs11724839	T	G	E025,H3K9ac_Pro;E038,H3K9a	0	NRSF_disc1
rs10516526	A	G	E003,H3K4me1_Enh;E007,H3K	E005;E008	.
rs17036142	T	C	E019,H3K9ac_Pro;E023,H3K9a	0	.
rs12374256	G	A	E027,H3K9ac_Pro;E110,H3K9a	0	Foxp3;SIX5_known1;ZEB1_known3
rs11097901	C	T	E001,H3K4me1_Enh;E004,H3K	E006;E017	.
rs7676975	A	T	E015,H3K27ac_Enh;E089,H3K2	0	Cphx;Irf_known4
rs10050333	T	A	E003,H3K4me1_Enh;E012,H3K	0	INSM1
rs10050159	G	A	E003,H3K4me1_Enh;E012,H3K	0	LUN-1;TATA_known5

Table 3.2: Subset of HaploReg exploratory *in silico* analysis output for lung function and INTS12 *cis*-eQTL SNPs indicates potential regulatory effects of these SNPs.

3.4 *In silico* attempt to assign putative INTS12 functions through paralog identification

Identification of INTS12 paralogs was undertaken to try and assign a putative function to INTS12 in order to guide subsequent experimental studies. A full length protein sequence (NP_001135943.1), containing N-terminal, PHD, and serine rich subdomains (Figure 3.7), was blasted against the NCBI's *Homo sapiens* RefSeq protein database. Surprisingly, no strong evidence suggests that INTS12 has any paralogs in the human genome as all the hits have alignment scores below 80 (Figure 3.8). Moreover, hits span only a small fraction of the query protein thus the entire INTS12 sequence appears to be unique. Nevertheless, some hits do show sequence similarity which is confined to the PHD domain (Figure 3.8). Importantly, this domain is annotated as a putative zinc and histone H3 binding site implying possible epigenetic roles for INTS12. Table 3.3 shows the proteins that appeared as top hits in this search. The general consensus molecular function of these proteins is regulation of gene expression and a role in epigenetic modifications. PHD domains can regulate gene expression through regulation of chromatin structure and dynamics and are considered epigenetic effectors (Bienz, 2006).

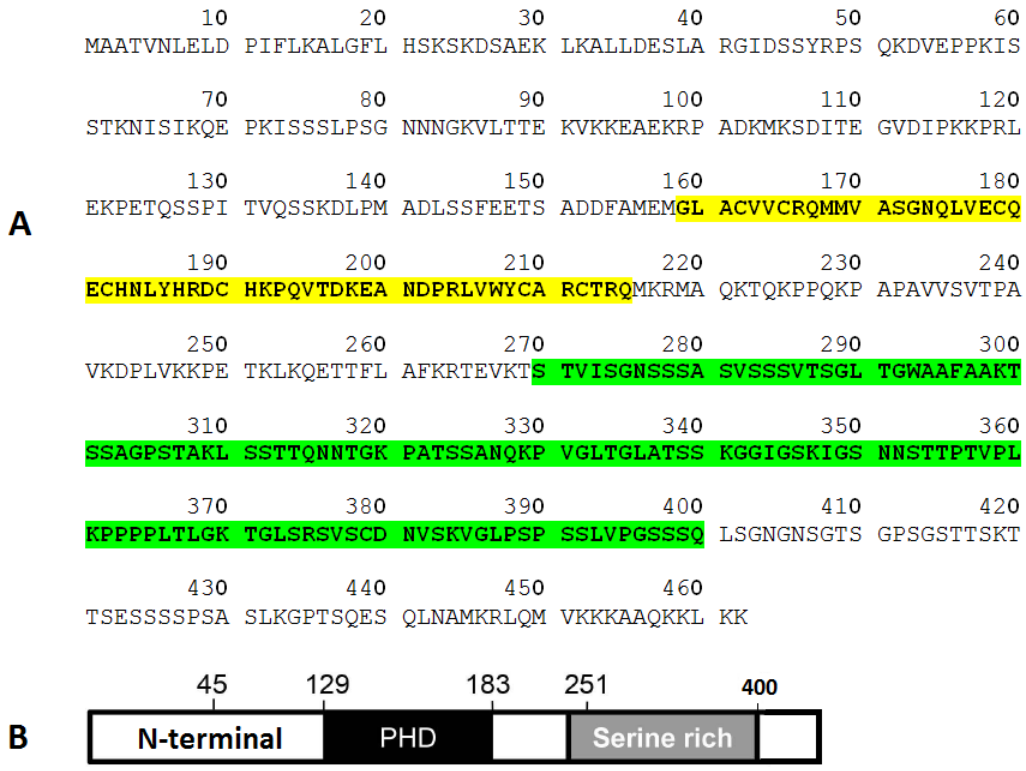


Figure 3.7: Human INTS12 protein sequence (NP_001135943.1) with highlighted PHD finger domain (yellow) and serine rich compositional bias domains (green) (A) as well as features of human INTS12 protein molecule (B).

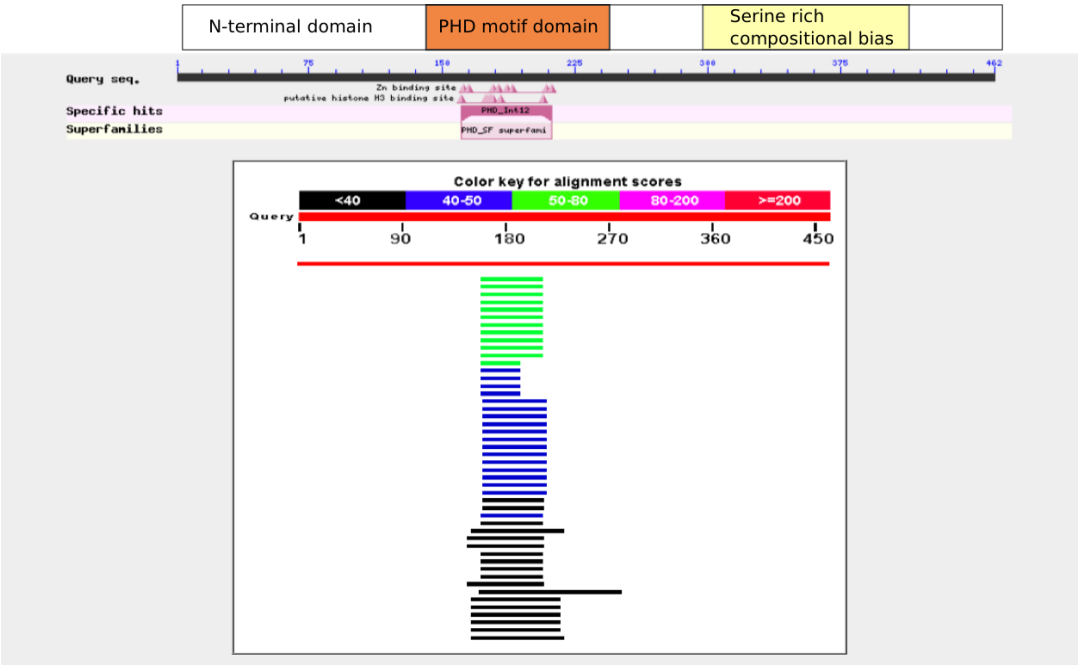


Figure 3.8: Full length INTS12 protein sequence (NP_001135943.1) BLASTP against a database of Homo sapiens non-redundant protein sequences shows the homology to be exclusively within the PHD domain. PHD domain appears as a putative zinc and histone H3 binding site.

INTS12 PROTEIN BLAST HITS SUMMARIES	
PHD finger protein 1 isoform a and b	This gene encodes a Polycomb group protein. The protein is a component of a histone H3 lysine-27 (H3K27)-specific methyltransferase complex, and functions in transcriptional repression of homeotic genes. The protein is also recruited to double-strand breaks, and reduced protein levels results in X-ray sensitivity and increased homologous recombination. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, May 2009]
PHD finger protein 21A isoform a and b	The PHF21A gene encodes BHC80, a component of a BRAF35 (MIM 605535)/histone deacetylase (HDAC; see MIM 601241) complex (BHC) that mediates repression of neuron-specific genes through the cis-regulatory element known as repressor element-1 (RE1) or neural restrictive silencer (NRS) (Hakimi et al., 2002 [PubMed 12032298]).[supplied by OMIM, Nov 2010].
sp110 nuclear body protein isoform a and c	The nuclear body is a multiprotein complex that may have a role in the regulation of gene transcription. This gene is a member of the SP100/SP140 family of nuclear body proteins and encodes a leukocyte-specific nuclear body component. The protein can function as an activator of gene transcription and may serve as a nuclear hormone receptor coactivator. In addition, it has been suggested that the protein may play a role in ribosome biogenesis and in the induction of myeloid cell differentiation. Alternative splicing has been observed for this gene and three transcript variants, encoding distinct isoforms, have been identified. [provided by RefSeq, Jul 2008]
histone-lysine N-methyltransferase 2A isoform 1 and 2 precursor	This gene encodes a transcriptional coactivator that plays an essential role in regulating gene expression during early development and hematopoiesis. The encoded protein contains multiple conserved functional domains. One of these domains, the SET domain, is responsible for its histone H3 lysine 4 (H3K4) methyltransferase activity which mediates chromatin modifications associated with epigenetic transcriptional activation. This protein is processed by the enzyme Taspase 1 into two fragments, MLL-C and MLL-N. These fragments reassociate and further assemble into different multiprotein complexes that regulate the transcription of specific target genes, including many of the HOX genes. Multiple chromosomal translocations involving this gene are the cause of certain acute lymphoid leukemias and acute myeloid leukemias. Alternate splicing results in multiple transcript variants.[provided by RefSeq, Oct 2010]
metal-response element-binding transcription factor 2 isoform a, b and c	No description available
bromodomain adjacent to zinc finger domain protein 2B isoform a	No description available

Table 3.3: Details of proteins showing similarity to human INTS12's PHD domain provide evidence for putative chromatin and gene regulatory roles.

3.5 INTS12 phylogenetic analyses

Orthologous INTS12 protein sequences were obtained from six model metazoan species (*Homo sapiens*, *Drosophila melanogaster*, *Mus musculus*, *Bos taurus*, *Xenopus laevis* and *Danio rerio*) and aligned. As can be seen in Figure 3.9, there appears to be a good degree of overlap between sequence dissimilarity and divergence time since the split from the common ancestor (Table 3.4, Hedges et al. 2006). When the alignment was performed using a richer dataset of 66 metazoan species, the conservation appeared to be more widespread providing qualitative evidence of a negative selection operating on INTS12 protein (Figure 3.10) as evolution was probably preserving some crucial function. Intriguingly, in only *Monodelphis domestica* opossum species INTS12 appears to have fusion sequence attached to the N-terminal domain explaining the gap in conservation observed towards its N-terminus.

Ensembl's phylogenetic reconstruction of orthologous gene sequences generated a tree in agreement with the universal tree of life (Forterre 2015); e.g. rodents and primates formed separate clades. Most of the genes used in this phylogenetic reconstruction are horizontally inherited orthologues from the last common metazoan ancestor. There were two duplications events during INTS12 molecular evolution yielding two paralogs in Rabbit and Microbat species. Moreover, INTS12 gene sequence has not been identified in *Caenorhabditis elegans* worm and *Saccharomyces cerevisiae* yeast genomes.

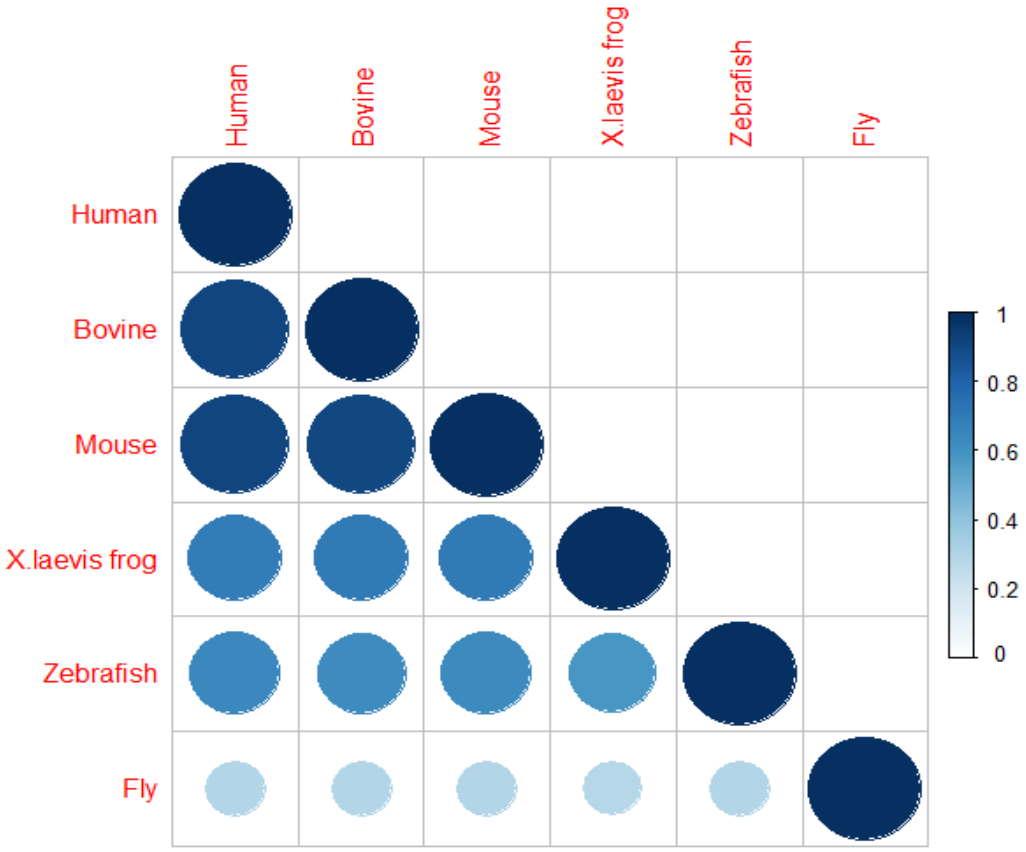


Figure 3.9: Metazoan INTS12 protein conservation in model organisms: the size and shades of blue colour in the matrix indicate percent sequence identity between each of the species' INTS12 protein sequences.

	Homo sapiens	Mouse	Bovine	Frog	Zebrafish	Fly
Human	N/A					
Mouse	90.9	N/A				
Bovine	97.5	97.5	N/A			
X.laevis Frog	355.7	355.7	355.7	N/A		
Zebrafish	429.6	429.6	429.6	429.6	N/A	
Fly	847	847	847	847	847	N/A

Table 3.4: The pairwise time lapse since the split from the common ancestor of the indicated species in millions of years (Hedges et al. 2006) reflects percent similarity between these species.

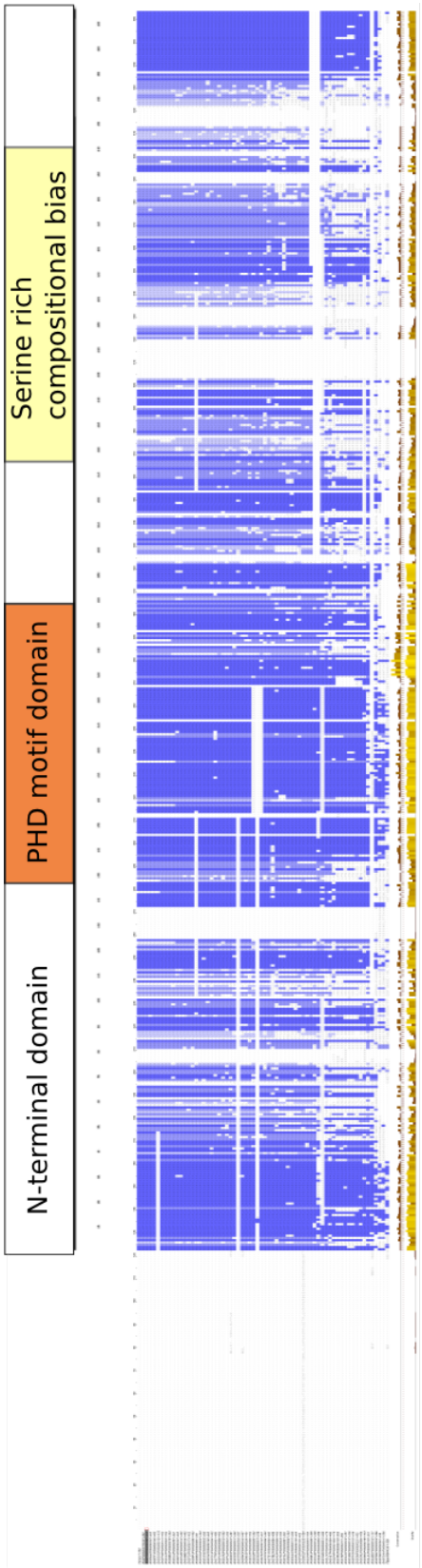


Figure 3.10: Conservation of INTS12 protein sequence in 66 metazoan species. The magenta shades of colour show evidence of negative selection even beyond the annotated protein domain. *Monodelphis domestica* opossum INTS12 appears to have fusion sequence attached to the N-terminal domain.

3.5.1 dN/dS ratio test

To provide quantitatively robust interpretation to the observed INTS12 protein sequence conservation (see section 3.5) a codon-based dN/dS ratio test was performed. In particular, it is an attempt to answer the question of which codon sites in the alignment are subject to positive or negative selection. To achieve that the Single-Likelihood Ancestor Counting (SLAC) method was used (Pond and Frost et al. 2005). Briefly, in this method given a particular phylogeny and maximum likelihood reconstructed ancestral sequence one aims to quantify the dN and dS parameters per each codon via counting procedure similar to the one outlined above (see section 3.1.2.1) by treating the reconstructed ancestral sequence as known. The ancestral sequence is a character state at the root of the neighbour-joining tree (Saitou and Nei, 1987) which is chosen to maximize the probability of the observed multiple sequence alignment (Page, 1999).

Using this approach there was 1 positively selected site and 374 negatively selected sites at $P<0.1$ which is the default statistical threshold in the SLAC programme (Pond and Frost et al. 2005). Thus out of the 374 codons with evidence of purifying selection 10% may have evolved neutrally, and there is 10% chance that the codon with evidence of positive selection is a false positive. Based on these observations and the fact there were 680 codon sites after multiple sequence alignment, it appears that ~55% of INTS12 protein was constrained by natural selection from changing while ~45% was evolving neutrally. The PHD domain, N-terminal and C-terminal subdomains look to be particularly under strong purifying selection (Figure 3.11). The positively selected codon number 191 is present within N-terminus. However, the strength of this selection is only ~1-fold above the neutral evolution hypothesis (dN/dS ~ 2.1). Overall, the average dN/dS value of INTS12 ORF is 0.197 and this quantification agrees with qualitatively detected sequence conservation.

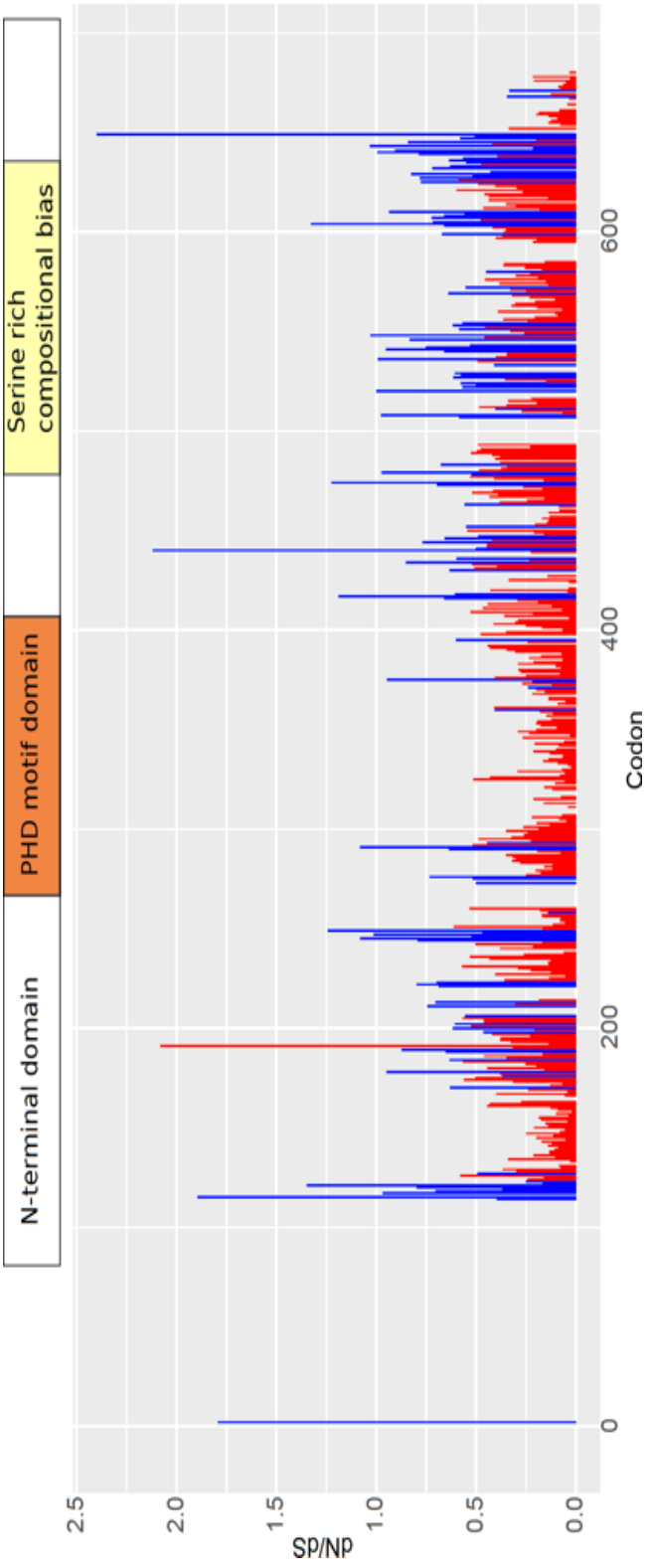


Figure 3.11: Quantitative assessment of INTS12 molecular evolution using a repertoire of metazoan open reading frames. The ratio of non-synonymous changes to synonymous changes (dN/dS) is shown throughout the protein. dN/dS approaching zero indicate strong and significant conservation (n=66 species): red colour P<0.1, blue colour P>0.1. P-value represents the probability of observed dN/dS ratio given the null the hypothesis of neutral evolution.

3.6 Development and optimization of methods to study INTS12 function in *in vitro* HBEC model

A considerable part of this thesis was devoted to the development and optimization of necessary tools to study INTS12 function. Because INTS12 expression is higher in the human bronchial epithelium than other airway structural cells (Obeidat et al. 2013), studies were concentrated on this cell type. Thus an in vitro HBEC model was used and initial experiments focused on the optimization of exogenous D-siRNA transfection into the cells and validation of gene knockdown on mRNA and protein levels. The necessary prerequisite for that was the optimization of quantitative INTS12 qPCR and qualitative immunofluorescence assays. A range of housekeeping genes expression was tested in the utilized model. Additionally, recombinant transient plasmid transfections were optimized for the INTS12 overexpression experiments.

3.6.1 INTS12 targeting D-siRNAs transfection optimization and validation

3.6.1.1 Validation of INTS12 qPCR assay

Prior to the measurement of INTS12 expression in D-siRNA transfected cells, it was necessary to establish a reliable INTS12 qPCR assay as INTS12 levels are the primary outcome measure after silencing. Premiers and probe were designed (see section 2.3.5.4) and their sequences can be found in the Appendix (Table 5 of Appendix). 300000, 30000, 3000, 300, 30 copies of pcDNA3.1-INTS12_v2 construct (see section 2.3.7) were prepared in triplicate. The precise number of 300000 plasmid copies was determined by (i) calculating the mass of a single plasmid molecule, (ii) calculating the mass of plasmid containing the required number of copies. Plasmid mass was calculated according to the following formula where *m* stands for mass and *n* stands for number of plasmid copies (Applied Biosystems):

$$m = n * 1.096 * 10^{-21}$$

The volume of plasmid containing the required mass and thus 300000 copies was determined based on the stock concentration. Stock plasmid concentration was measured spectrophotometrically on NanoDrop 2000 instrument (Thermo-Scientific©). The rest of samples were produced in 10-fold dilution series. The generated samples were used to run qPCR reaction as described in Chapter 2 (see section 2.3.5.4). INTS12 assay fluorescence threshold was established at 0.1 and technical replicate values showed little variability (Figures 2.12 and 2.13). Standard curve gradient was -3.564 thus within the desirable range of -3.1 to -3.6 (Stratagene) and the correlation between C_t values and number of plasmid copies was 0.997 implying predictability of the two variables. The assay efficiency was computed to be 90.8% and therefore the assay could be reliably used in INTS12 expression measurements.

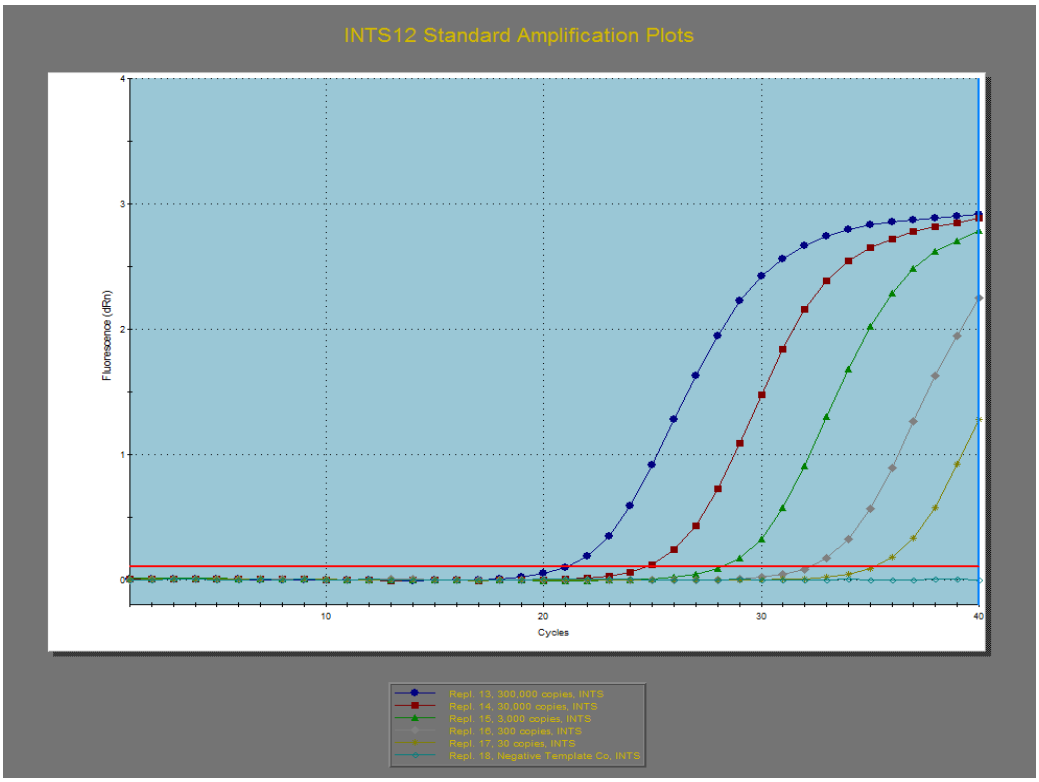


Figure 3.12: qPCR amplification plots showing the accumulation of product in real time. For 100% efficient assay condition the average C_t difference between samples in 10-fold dilution series is $\log_2 10$, i.e. ~3.32. As it can be seen the difference in the plots reaching the threshold is close to this value.

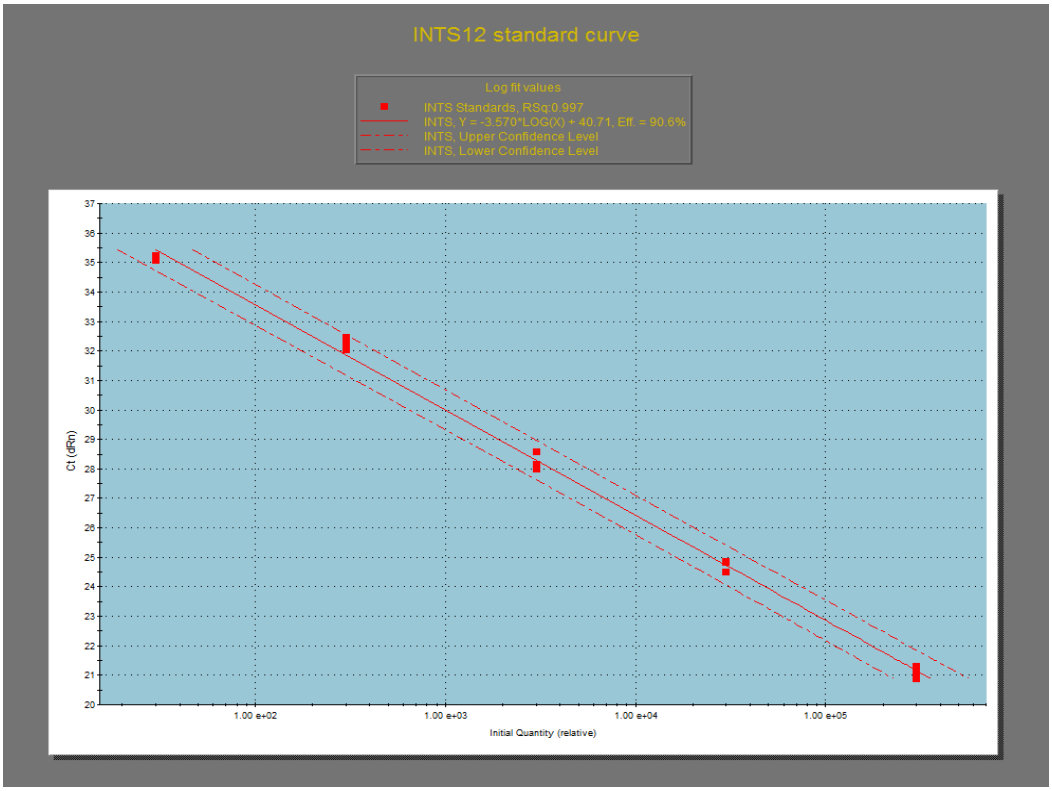


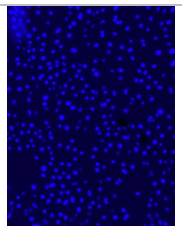
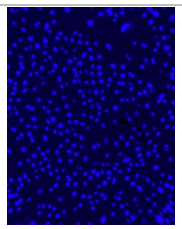
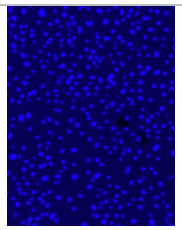
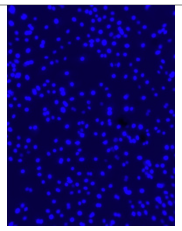
Figure 3.13: INTS12 calibration curve showing the relationship between C_t values and number of plasmid copies. The individual red dots represent the individual technical replicates and appear to be close to each other indicating low technical variability. The intercepted lines above the main slope are assay's upper and lower limits of 95% confidence interval.

3.6.1.2 D-siRNA transfections optimization

Initial attempts to knockdown INTS12 expression by FuGENE6® mediated transfection were unsuccessful (data not shown). Two main variables might have been responsible for this failure: either D-siRNAs were not introduced into the cells or the utilized sequences are not effective in knocking down INTS12 mRNA.

In order to address the first possibility, fluorescently labelled Cy3-D-siRNA was transfected using FuGENE6® and INTERFERin® reagents in three biological replicates. Prior to imaging, cells were DAPI stained. No cells appeared to be fluorescent in the non-treated condition suggesting the absence of auto-fluorescence. No fluorescence was observed in any of the transfection reagents alone. Transfecting with FuGENE6® at 10nM, 50nM and 100nM Cy3-D-siRNA did not result in any detectable fluorescence (n=3; Figure 3.14). In fact, FuGENE6® transfected cells looked identical to non-treated and reagent only conditions. On the other

hand the vast majority of cells transfected with INTERFERin® appear fluorescent. Based on the qualitative assessment of images there seems to be a correlation between each DAPI nuclear staining and red Cy3 fluorescence indicating good transfection efficiency. Although the intensity of fluorescence in the cells was dose dependent, the number of transfected cells appeared to be the same for 10nM, 50nM and 100nM Cy3-D-siRNA concentrations. Importantly the same amount of transfection reagent was used in these experiments. Various concentrations of Cy3-D-siRNA were used in order to test the aforementioned dose dependency. As INTERFERin® showed much higher transfection efficiency than FuGENE6® it was taken forward for subsequent experiments.

Reagent	Condition	DAPI channel	Cy3 channel
FuGENE6®	Non-treated		
INTERFERin®	Non-treated		
FuGENE6®	Reagent only		
INTERFERin®	Reagent only		

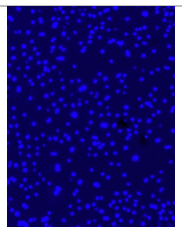
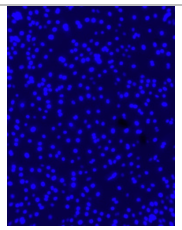
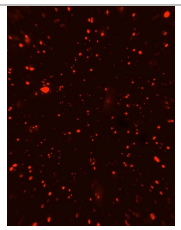
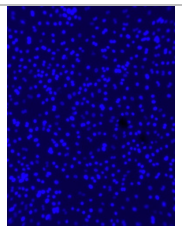
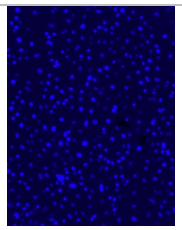
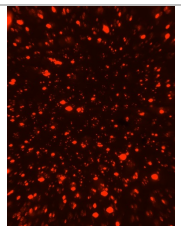
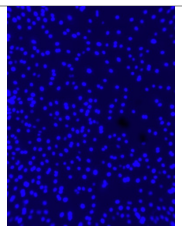
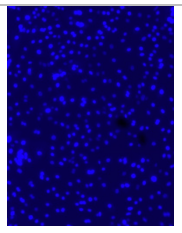
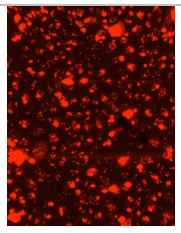
FuGENE6®	10nM Cy3-D-siRNA		
INTERFERin®	10nM Cy3-D-siRNA		
FuGENE6®	50nM Cy3-D-siRNA		
INTERFERin®	50nM Cy3-D-siRNA		
FuGENE6®	100nM Cy3-D-siRNA		
INTERFERin®	100nM siRNA		

Figure 3.14: Representative images of three biological replicates showing DAPI and Cy3 channelled cells transfected with 10nM, 50nM and 100nM Cy3-D-siRNA concentrations using FuGENE6® and INTERFERin® reagents and their respective controls. Cells were imaged 24h after the initial D-siRNA transfections.

3.6.1.2.1 Demonstration of RNAi functionality

Having established INTERFERin® as the transfection reagent of choice and optimized the transfection conditions, a validated HPRT1 positive control D-siRNA was used to demonstrate the functionality of RNAi in *in vitro* HBEC model. HPRT1 D-siRNA transfection was performed at the manufacturer's recommended concentration of 10nM. Cells transfected with HPRT1 D-siRNA had HPRT1 levels attenuated by 88% relative to cells transfected with scrambled D-siRNA control (P<0.001, n=3; Figure 3.15), demonstrating functionality of RNAi in the utilized model. This observation was qualitatively validated by end-point PCR reaction after 28 cycles (Figure 3.16) using primers against canonical HPRT1 transcript sequence

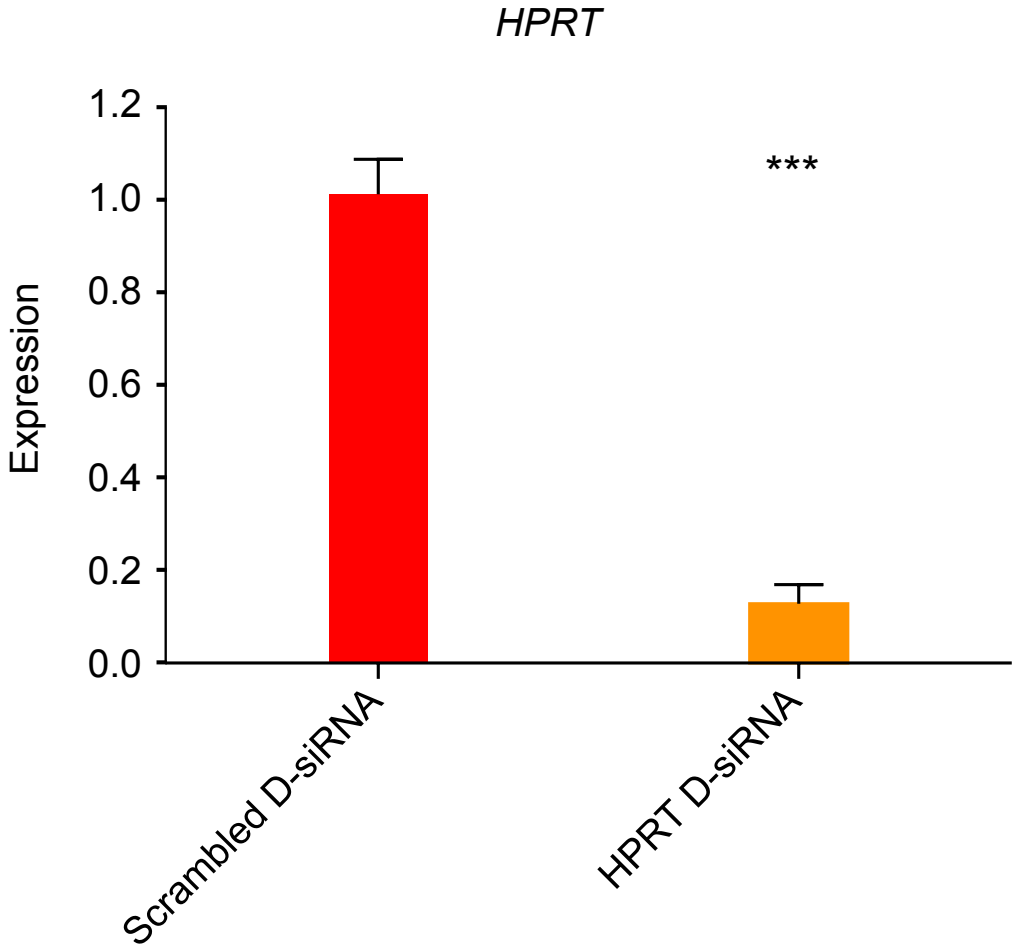


Figure 3.15: HPRT ΔCt expression. HBECs transfected for 48h with 10nM positive control HPRT1 D-siRNA yield 88% reduction in *HPRT1* expression (n=3). Statistical tests were performed comparing to scrambled D-siRNA control: ***P<0.001. Individual ΔCt gene expressions are relative to the mean of scrambled D-siRNA condition. Error bars represent standard error of the mean.

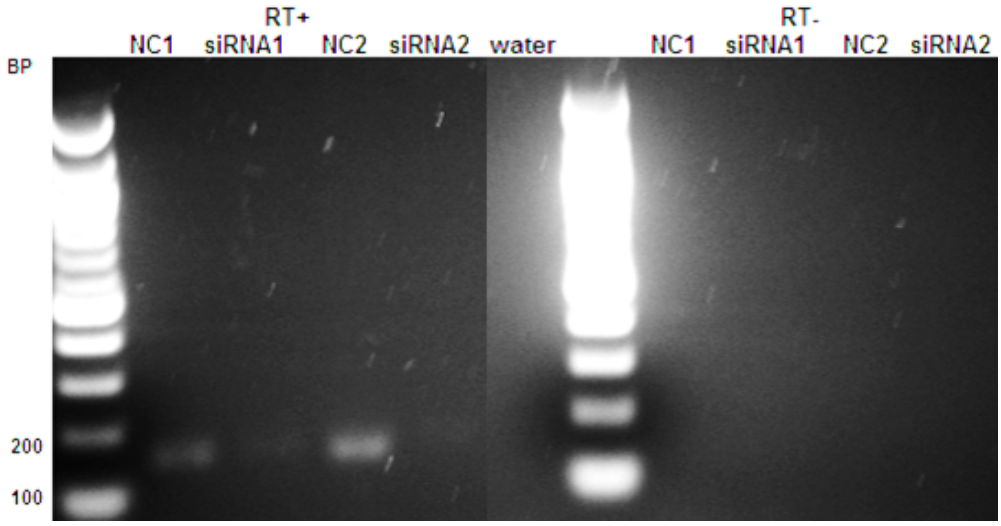


Figure 3.16: HPRT expression by end product PCR and gel electrophoresis. Reduction in HPRT1 expression in D-siRNA transfected cells is apparent in end point PCR analysis after 28 cycles. The expected amplicon size of 141bp is seen in cells transfected with scrambled D-siRNA (NC) but not positive control HPRT1 D-siRNA (siRNA) (n=2).

3.6.1.3 Optimizing INTS12 knockdown

Three different INTS12 D-siRNAs were transfected into HBECs in order to test their respective silencing efficiencies at 10nM dose. Initial studies were performed to identify a suitable housekeeper gene to account for differences in input RNA and differences in cDNA synthesis efficiencies. The genes chosen were glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) and transferrin receptor (*TfR*). Their expression was quantified on the samples from INTS12 knockdown experiments. As *GAPDH* had a more constant expression across the experimental conditions than *TfR*, it was chosen for the gene expression normalization in subsequent experiments. In fact, TfR levels were significantly increased in two out of three D-siRNAs ($P<0.01$, n=4; Figure 3.17) rendering it inappropriate reference for gene expression normalization in this experimental model. GAPDH normalized $\Delta\Delta C_t$ *INTS12* expression in HBECs transfected with three D-siRNAs was decreased by 80±9%, 29±9%, 69±9% ($P<0.05$, n=4; Figure 3.18). As desirable knockdown was defined as >80%, D-siRNA A and C were taken forward for further optimizations.

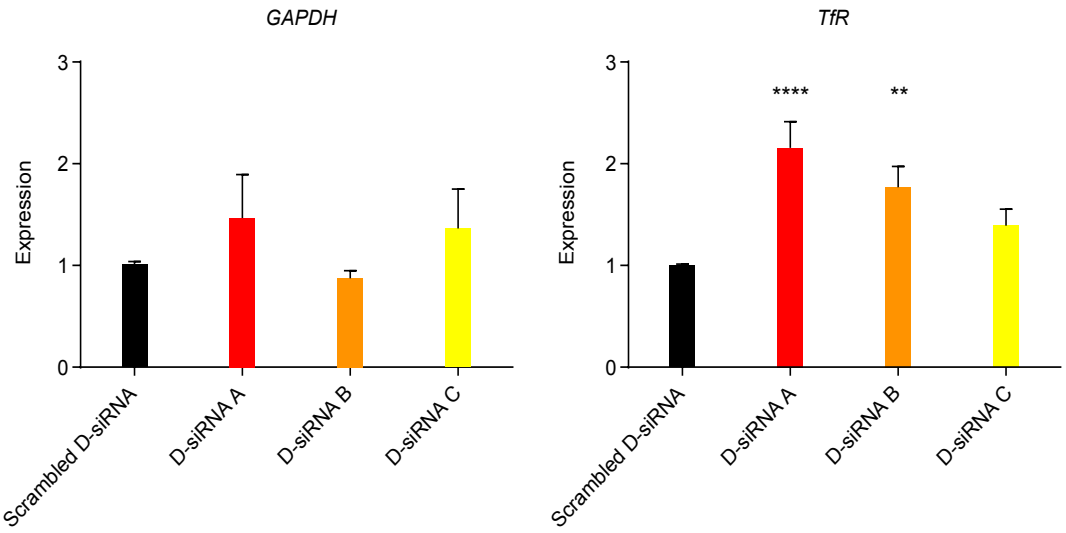


Figure 3.17: *TfR* and *GAPDH* ∆Ct expressions in INTS12 silenced HBECs. Cells were transfected at 10nM D-siRNAs concentration and gene expression was assessed 48h after the start of RNAi (n=4). Statistical tests were performed comparing to scrambled D-siRNA control: **P<0.01, ****P<0.0001. Individual ∆Ct gene expressions are relative to the mean of scrambled D-siRNA condition. Error bars represent standard error of the mean.

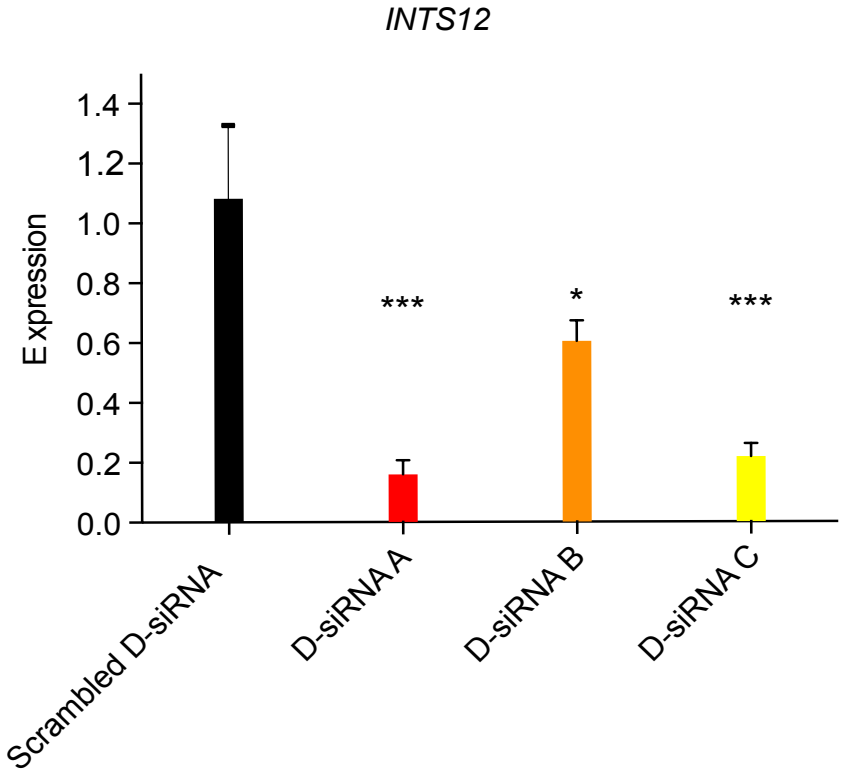


Figure 3.18: *INTS12* ΔΔCt levels in HBECs transfected with three D-siRNAs at 10nM for 48h. Statistical tests were performed comparing to scrambled D-siRNA control: *P<0.05, ***P<0.001. Individual ∆∆Ct gene expressions are GAPDH normalized and relative to the mean of scrambled D-siRNA condition. Error bars represent standard error of the mean.

3.6.1.3.1 Minimizing off-target effects by testing silencing efficiency at a range of concentrations

As described in Chapter 2, the likelihood of off-target effects in D-siRNA induced knockdown experiments increases with the dose of D-siRNA (see section 2.2.1). Therefore, it is of importance to try minimizing off-targeting by using the lowest possible concentration of D-siRNA (Jackson and Linsley, 2010). Thus the experimental optimization of D-siRNA knockdown experiments is concerned with achieving a trade-off between sufficient silencing and low D-siRNA dose, as lower D-siRNA concentrations imply less efficient knockdown.

HBECs were transfected with the chosen D-siRNAs at 0.1nM, 1nM and 10nM concentrations and INTS12 expression was compared relative to scrambled D-siRNA transfected at 10nM. Control condition was performed only at 10nM concentration for the simplicity of experimental design, and although may not be the ideal for like-to-like comparisons against 0.1nM and 1nM doses, the higher concentration in the control is preferable for a more likely non-specific inhibition of INTS12 at the higher scrambled D-siRNA concentration.

Relative to negative control, INTS12 levels were reduced by 67±6%, 77±6% and 78±6% for D-siRNA A and by 63±9%, 73±9%, and 58±10% for D-siRNA C at 0.1nM, 1nM, and 10nM concentrations respectively. ($P<0.01$, $n=3$; Figure 3.19). Considering these results, it is preferable to use 1nM concentration in silencing experiments, as in D-siRNA A transfection 0.1nM dose did not reach the desired degree of knockdown and on average there is only a slight improvement in knockdown between 1nM and 10nM dose. As far as D-siRNA C is concerned, 1nM dose achieved a better knockdown than both 0.1nM and 10nM doses. Therefore, 1nM concentration was chosen for all subsequent experiments.

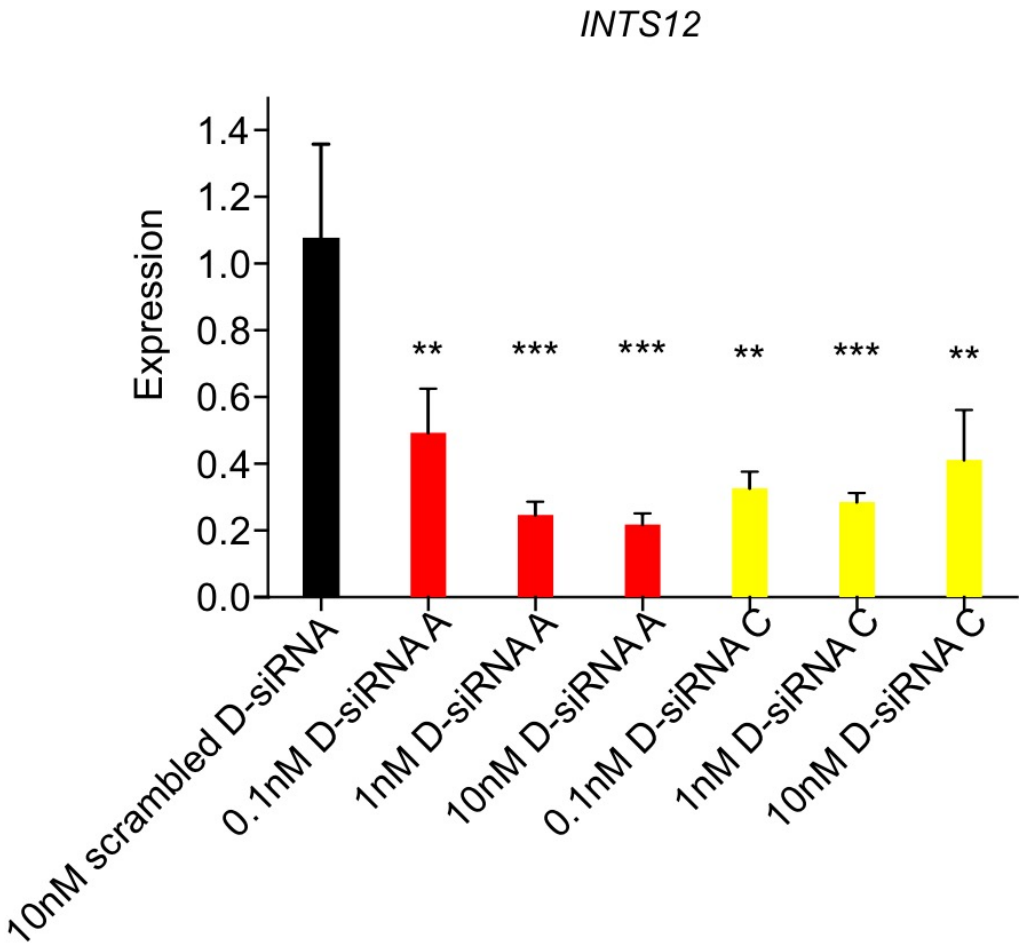


Figure 3.19: INTS12 ΔΔCt levels in HBECs transfected with D-siRNAs A and C at a range of concentrations for 48h. Statistical tests were performed comparing to scrambled D-siRNA control: *P<0.05, ***P<0.001. Individual ΔΔCt gene expressions are GAPDH normalized and relative to the mean of scrambled D-siRNA condition. Error bars represent standard error of the mean.

3.6.1.3.2 Qualitative demonstration of INTS12 protein knockdown

Studies that investigated the relationship between mRNA and protein levels have reported correlation coefficients ranging from 0.4 to 0.8 (Maier et al. 2009). Since the correlation between the two types of molecules is not always strong (i.e. not >0.8), due to multiple factors such as translational efficiency or protein turnover, it is generally preferable to measure levels of protein in addition to mRNA molecules. Although densitometry based quantification via Western blot is a commonly used technique for this purpose, IF may also be used. The advantage of WB over IF is the possibility to ascertain the molecular weight of the detected protein however discrepancy between expected and actual weight could occur due to post-translational modifications. On

the other hand, according to ENCODE criteria, although IF cannot tell the molecular weight of detected protein it can be reliably used for specific detection when combined with D-siRNA knockdown (Landt et al. 2012): the disappearance of signal in knockdown condition indicates specific detection.

In this thesis the latter approach was used to (I) determine the INTS12 sub-cellular localization as well as to (II) qualitatively demonstrate INTS12 protein knockdown in addition to its mRNA-based quantification. The reason IF instead of WB was relied on, is because WB attempts were unsuccessful due to the detection of multiple bands falling outside the required molecular weight. However, there was a dominant band at the expected position (Figure 3.20). The presented WB data do not show differences in band densities between scrambled D-siRNA and D-siRNA A and C probably due to lack of sufficient assay sensitivity to detect change 48h since the knockdown initiation. The detection of multiple bands in addition to the predicted one based on molecular weight of INTS12 is problematic for the confirmation of antibody specificity. However, the subsequent IF data in combination with gene knockdown demonstrate its suitability and specificity for assessment of protein depletion and ChIPseq (Figure 3.21, Figure 6.2, Figure 6.3) as per ENCODE criteria (Landt et al. 2012).

The INTS12 IF procedure described in Chapter 2 (see section 2.4.1) was optimized for primary antibody concentration, secondary antibody concentration and blocking step. The primary anti-INTS12 antibody used in the IF partly recognizes PHD and N-terminal domains. As homology searches against the database of human proteins revealed the uniqueness of N-terminal domain (see section 3.4) it provides supportive *in silico* evidence of antibody specificity.

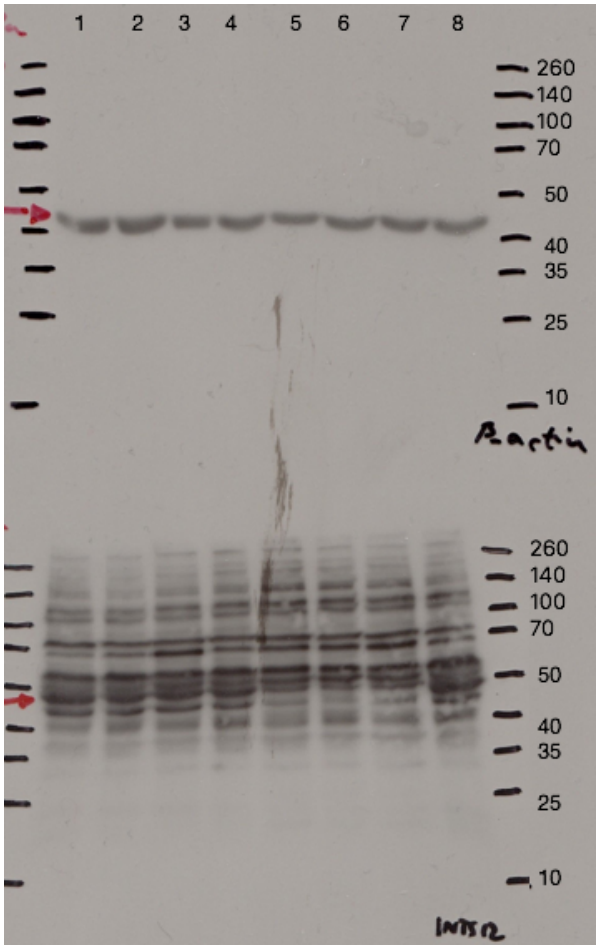


Figure 3.20: WB of INTS12 in HBECs. The top panel represents a housekeeping beta-actin protein expression. The predicted beta-actin band of a molecular weight of 42kDa is indicated by red arrow. Beneath is the panel of WB results for INTS12. Although multiple banks can be seen a band of predicted molecular weight of 49kDa is indicated by red arrow. Each column corresponds to a lysate obtained from HBECs grown under particular experimental condition. Columns 1, 2, 3, 4 correspond to conditions un-transfected, scrambled D-siRNA, D-siRNA A and D-siRNA C respectively using the day 2 protocol. Samples 5, 6, 7, 8 are from a different biological replicate experiment and are presented in the same order.

3.6.1.3.2.1 INTS12 is localized in the nucleus of HBECs and other cell types

As described in section 1.8.5, INTS12 was categorized as being a nuclear protein. In fact, out of all systematically tested INTScom subunits, INTS12 was the only member that was found exclusively in the nucleus, i.e. 100% of tested cells had INTS12 localized in the nucleus (Jodoin, Sitaram et al. 2013). Other INTScom subunits had a more diffuse localization or were found only in the cytoplasm. For example, 100% of tested cells had INTS2 localized in the cytoplasm. These

observations indicate that although INTScom subunits were purified in a complex association (Baillat et al. 2005) it is possible for them to be physically distant from each other (Jodoin, Sitaram et al. 2013) and thus potentially have distinct functions.

Having optimized the INTS12 IF procedure the technique was used to test INTS12 localization in HBECs. In agreement with its nuclear localization observed in HeLa cells (Jodoin, Sitaram et al. 2013), in lung tissue epithelial cells and pneumocytes (Obeidat et al. 2013) as well as in 44 normal human tissues from the human protein atlas dataset (Uhlen et al. 2005), INTS12 appeared to be a nuclear protein based on the agreement between INTS12 and dsDNA DAPI staining (n=4, Figure 3.21).

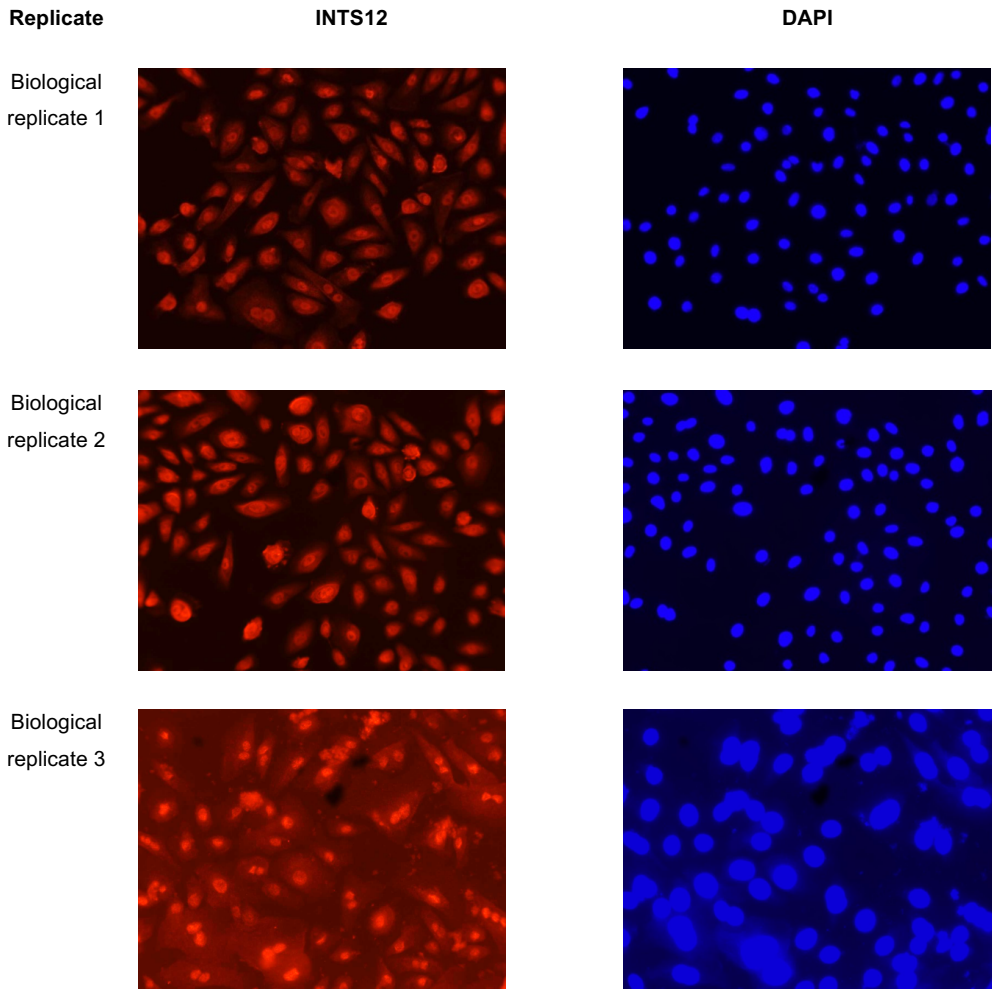


Figure 3.21: Nuclear localization of INTS12. In HBECs INTS12 has a nuclear localization based on the correspondence between INTS12 and dsDNA DAPI staining. Experiment was performed in four biological replicates. Representative images of three biological replicates are shown. Isotype control exposed cells were negative for staining. Cells were imaged at the same magnification.

3.6.1.3.2.2 HBECs treated with INTS12 D-siRNA are depleted of INTS12 protein

In order to test INTS12 knockdown on the protein level, HBECs were treated with D-siRNAs A and C using the same conditions as those used in the functional experiments. As observed before, in un-transfected and scrambled D-siRNA transfected cells, INTS12 appeared localized in the nucleus. On the other hand, cells treated with INTS12 silencing D-siRNAs had a marked reduction in this specific nuclear staining (n=2; Figure 3.22) implying not only a quantified and successful knockdown on the mRNA level, but also qualitatively assessed and demonstrated knockdown on the protein level. In fact, INTS12 depleted cells stained for INTS12 looked like un-transfected cells exposed to isotype control.

As staining was reduced in INTS12 silenced cells this also demonstrates specific affinity of the used antibody, according to the ENCODE criteria (Landt et al. 2012)

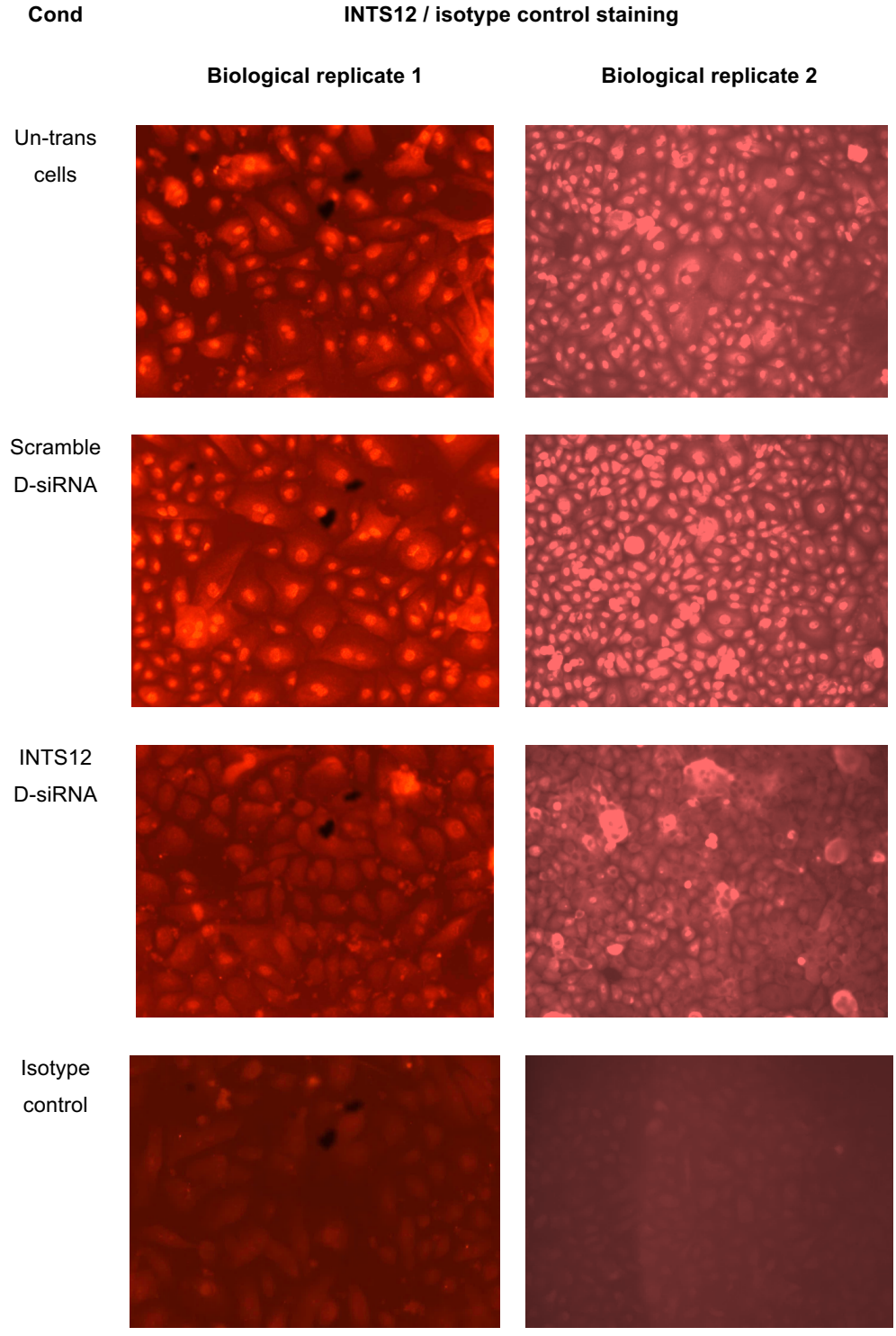


Figure 3.22: INTS12 protein depletion in INTS12 D-siRNA transfected HBECs. Immunofluorescence shows INTS12 to have a nuclear localization in un-transfected and scrambled D-siRNA transfected cells. Nuclear signal is reduced in INTS12 D-siRNA transfected cells and is comparable to isotype control exposed cells.

3.6.2 Optimizing transient recombinant INTS12 constructs transfections

In addition to the knockdown approach for gene function discovery and studies, overexpression approach was used to test a specific hypothesis about the regulatory properties of INTS12 (see section 7.1.1). In order to transfect two recombinant INTS12 constructs, pEGFP-N1 plasmid containing GFP ORF was used to optimize FuGENE6® mediated transfection as described in Chapter 2. Transfection efficiency was qualitatively compared between cells transfected with different plasmid DNA concentrations and reagent volume to total recombinant DNA mass ratios.

Transfection efficiency appeared to increase proportionally with the concentration of recombinant DNA and was much improved when using 3:1 rather than 3:2 ratio (Figure 3.23). Therefore, the optimal DNA concentration and ratio were determined to be 2µg/ml and 3:1 respectively. Thus HBECs were transfected with constructs pcDNA3.1-INTS12_v2 encoding full length canonical protein as well as pcDNA3.1-INTS12_v3 encoding a truncated protein using the optimized conditions and had 710 (not significant) and 995 ($P<0.07$) fold increases in INTS12 mRNA levels respectively (Figure 3.24). It is important to note that although an increase in INTS12 expression on the mRNA level has been detected, it is currently unclear whether this is accompanied by an increase in protein as well. As it is possible that variable levels of mRNA may not correlate with protein, future efforts aiming at further elucidating the effect of INTS12 overexpression should confirm induced protein expression.

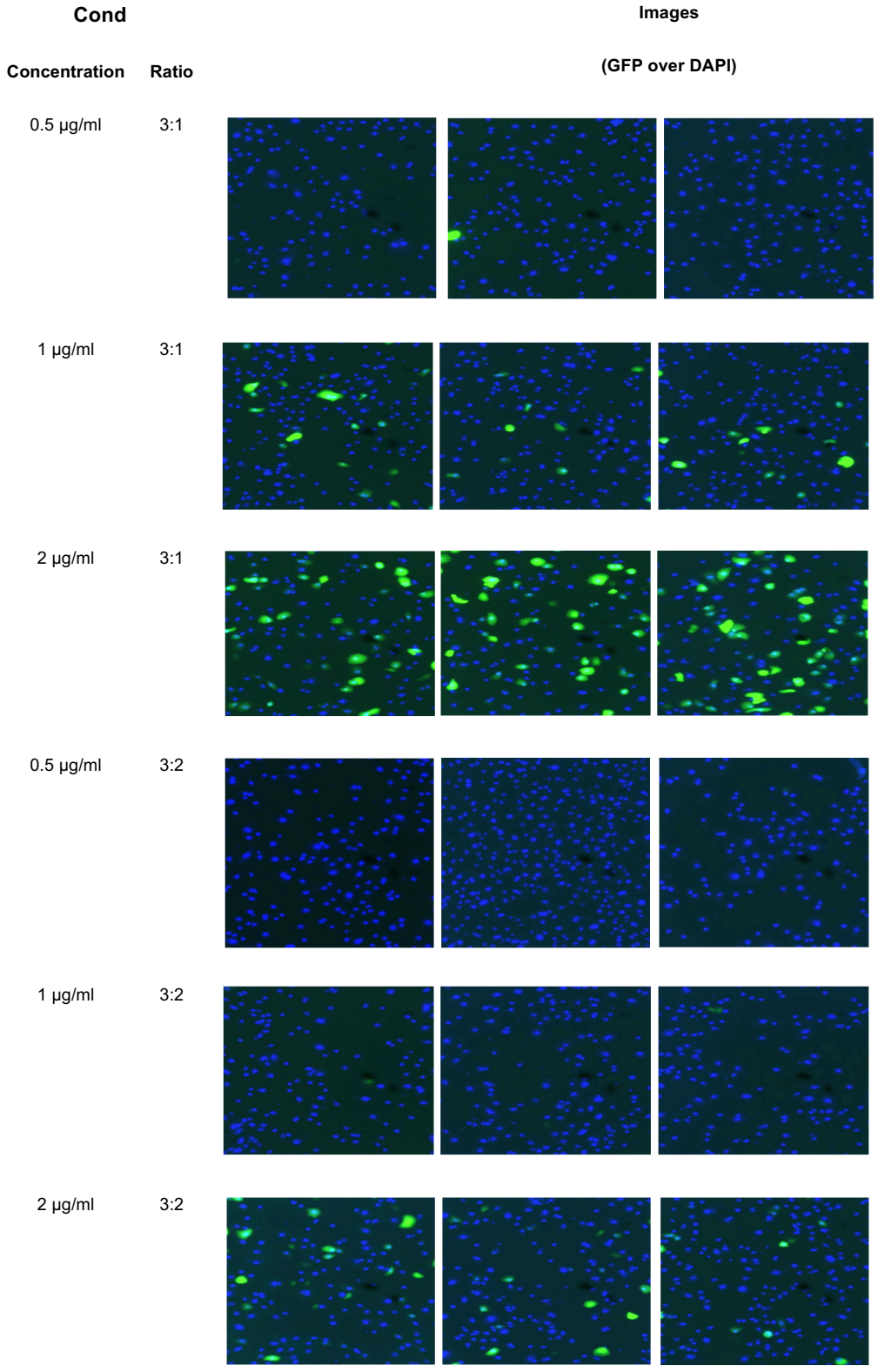


Figure 3.23: Optimizing recombinant DNA transfection into HBECs. Qualitative assessment of DNA concentration and reagent ratio variables determined 2µg/ml using 3:1 ratio to have the optimal transfection efficiency and therefore were used for INTS12 overexpression construct transfections. GFP channel was laid over DAPI channel.

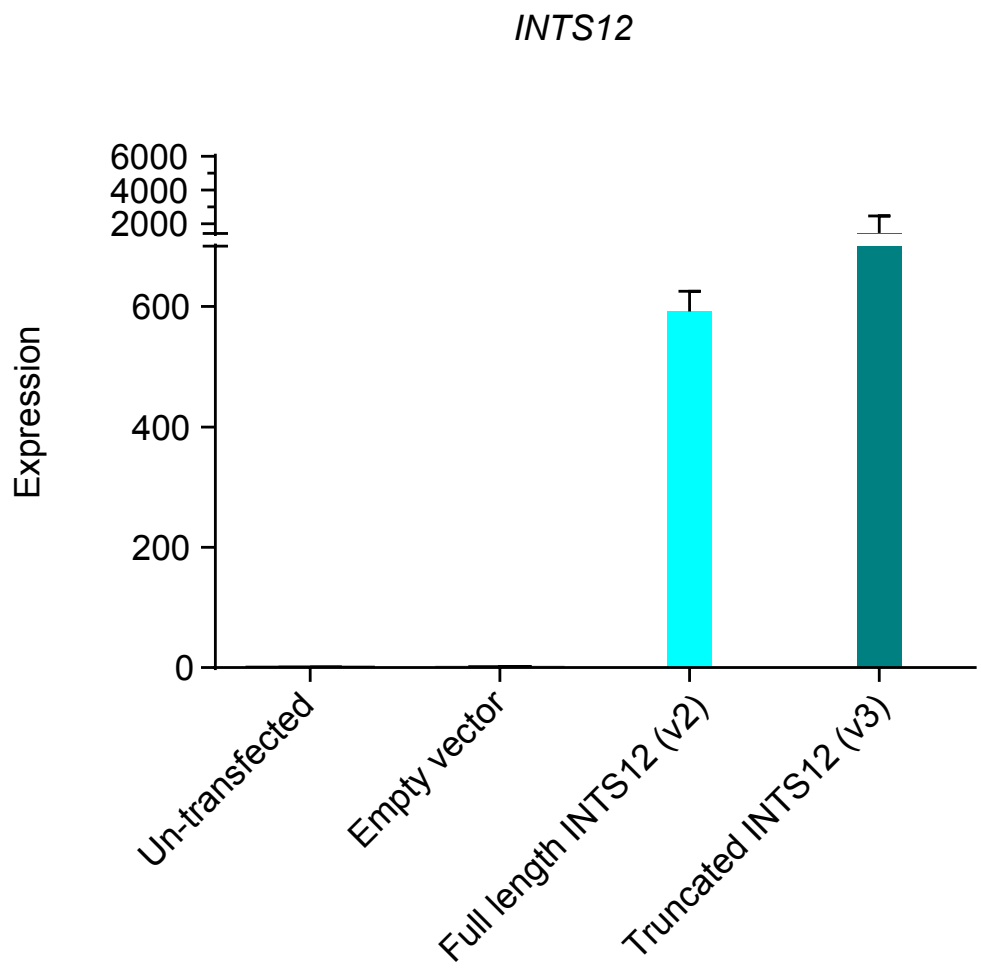


Figure 3.24: Transient transfection of overexpression constructs results in at least 700 fold increases in INTS12 mRNA levels of v2 and v3 variant transfected cells relative to empty vector transfected cells. There was 42% and 7% chance of observing such increase in expression given the null hypothesis of no difference in expression for variants v1 and v2 respectively.

3.7 Discussion

The overarching aims of this Chapter were to identify a likely gene of 4q24 locus contributing to lung function, bioinformatically explore this gene by investigating its molecular evolution and to develop the experimental tools for its *in vitro* study (Figure 3.25).

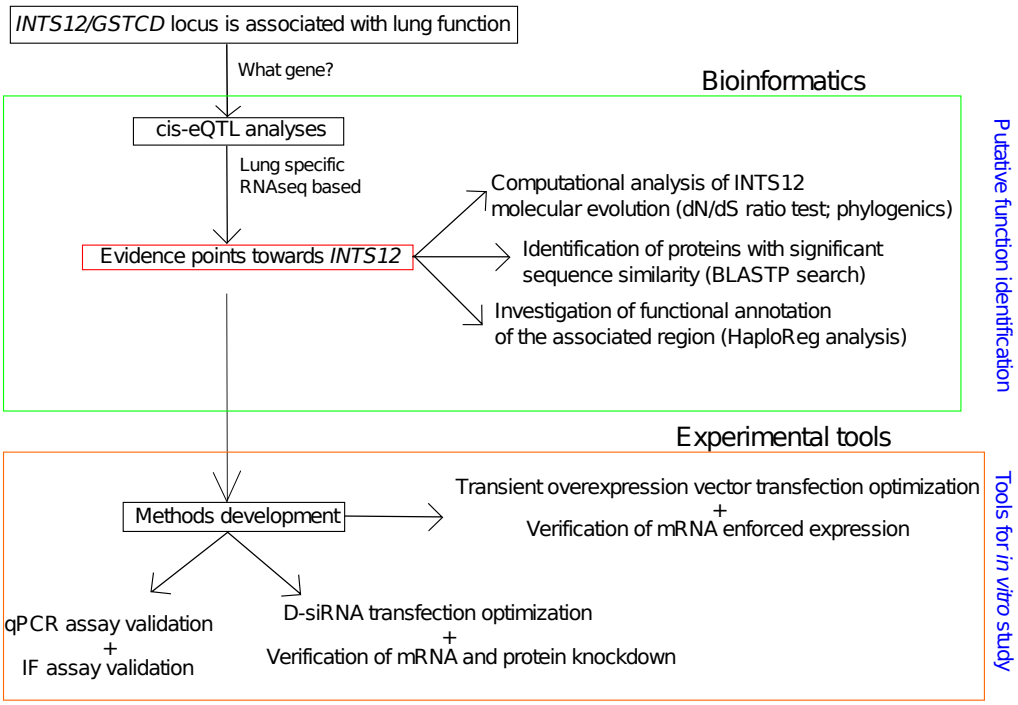


Figure 3.25: An overview of bioinformatics and experimental tools development discussed in this Chapter.

The 4q24 locus has been reproducibly associated with lung function parameters and it is currently believed that there are three independent signals in this region, which include one signal at the *INTS12*/*GSTCD* locus. Relying on systematic *in silico* gene prioritization strategy, this chapter provided up-to-date evidence implying variants which predict *INTS12* expression also contribute to the population variation in lung function. Despite the physical closeness of *INTS12* and *GSTCD* genes, SNPs associated with lung function in the SpiroMeta study (Repapi et al. 2010) are also predictive of *INTS12* but not *GSTCD* expression (Table 3.1). Importantly, SNPs correlated with lower expression were risk factors for lower lung function. Interestingly, lung INTS12 levels are 5 times higher than GSTCD (Figures 3.4 and 3.5) and *INTS12* expression in the lung is among top 13 tissues with highest expressions whereas

GSTCD expression is higher in greater number of other tissues (Figure 2A and 2B). Lung function and cis-eQTL INTS12 SNPs effectively represent a linked haplotype (Figure 3.6) and thus it is not possible to ascertain which ones are causal in relation to either gene expression or lung function.

Although *INTS12* and *GSTCD* have been suggested to be co-ordinately expressed (Obeidat et al. 2013), the observation of eQTL effect on only one gene in this locus contradicts the co-ordinated expression hypothesis. The rationale behind this deduction is that provided two genes are oppositely transcribed and sharing the same promoter, as it has been argued for *INTS12* and *GSTCD* (Obeidat et al. 2013), then these genes are likely to be having common regulatory signatures where causative SNPs occur. Thus the prediction is that *cis*-eQTL effect holds true for both genes. The absence of SNP-to-gene lung expression correlation for *GSTCD* but its presence for *INTS12*, raises the possibility that there are potentially multiple promoters or enhancer elements separately regulating the two genes. Puzzlingly, supplemental analysis revealed that these genes also appear to be co-ordinately expressed in the Genotype-Tissue Expression project ($r=0.3$; $P<0.0001$; Lonsdale et al. 2013) as well as Kim et al. datasets ($r=0.3$; $P<0.01$; Kim et al. 2015) albeit the strength of their correlation is much weaker than previously reported ($r=0.8$; $P<0.0001$; Obeidat et al. 2013). These findings point towards the complexity of *INTS12* and *GSTCD* regulation and provide reason for further exploration.

INTS12 eQTL SNPs fall within regions annotated with DNA binding domains for a range of regulators of gene expression, further supporting the leading hypothesis that altered expression of INTS12 may be responsible for differences in lung function (Table 3.2). A protein BLAST search indicates the lack of INTS12 paralogs in the human repertoire of proteins, although some proteins have sequence similarity to the gene's PHD domain functionally implicated in chromatin and gene regulation (Table 3.3). Indeed, Ensembl database reports the existence of INTS12 orthologues and two paralogs in Rabbit and Microbat only.

It is not fully clear whether PHD fingers have a common function (Bienz, 2006). Interestingly, there are no INTS12 orthologues in neither simple multicellular *Caenorhabditis elegans* nor in unicellular *Saccharomyces cerevisiae,* suggesting this gene to be functionally important in multicellular organisms in which multiple tissues differentiation occurs as INTS12 orthologues have been identified only in those kinds of organisms. An *INTS12*-wide evolutionary analysis revealed that this gene was under negative selection, i.e. constrained from mutating, presumably to maintain a critical cellular or developmental function (dN/dS=0.197). A codon-by-codon selection analysis has shown that PHD domain, N-terminal and C-terminal subdomains are particularly under strong purifying selection. Intriguingly, out of these three regions only N-terminal subdomain was shown to be required and sufficient for INTS12 canonical function of snRNA processing in fly cells (Chen et al. 2013), i.e. PHD and C-terminal domains are dispensable for this activity. Therefore, in light of their strong conservation the snRNA processing dispensable parts of the protein require functional explanation. Moreover, it suggests the existence of additional, possibly multiple, roles for this gene.

As *INTS12* expression is higher in the bronchial epithelium relative to other airway cells (Obeidat et al. 2013), the HBEC model was chosen for the functional studies. A series of experiments were carried out in order to study the prioritized *INTS12* function and have successfully (i) optimized INTERFERin®-mediated transfection of exogenous fluorescent D-siRNAs into HBECs, (ii) demonstrated functionality of RNAi in the model using validated positive control D-siRNA, (iii) demonstrated a >80% INTS12 mRNA knockdown at the lowest possible concentration of two independent D-siRNAs, (iv) qualitatively demonstrated near complete INTS12 protein depletion and its nuclear localization, (v) optimized FuGENE6®-mediated transfection of recombinant DNA construct and (vi) used these transfection conditions to profoundly overexpress v2 and v3 INTS12 construct variants. In conclusion an *in silico* based evidence was produced indicating the probable contribution of INTS12 to pulmonary function and the

necessary experimental tools were optimized and set the scene for all subsequent functional studies.

4. Functional role of INTS12 in human snRNA processing

4.1 Introduction

As described in the Chapter 1, INTS12 protein is a member of the INTScom complex. INTScom was shown to stably accompany POLII and at a molecular level has been implicated in small nuclear RNA (snRNA) and Cajal bodies biogenesis (Baillat et al. 2005, Takata et al. 2012), perinuclear dynein dynamics (Jodoin, Sitaram et al. 2013) and with POLII pause and release (Gardini et al. 2014, Stadelmayer et al. 2014). It is unclear whether all the INTScom subunits or a subset of subunits are involved in the above processes. At the functional level, targeted knockdown and mutagenesis experiments demonstrated INTScom to be necessary for adipogenesis (Otani et al. 2013) and haemopoiesis (Tao et al. 2009).

Because expression of INTS12 is high in the human bronchial epithelium compared to smooth muscle and peripheral blood mononuclear cells (Obeidat et al. 2013), further studies were concentrated predominantly on this cell type. What is known directly about the function of INTS12 is that in *D.melanogaster* S2 cells it is necessary for snRNA processing (Ezzeddine et al. 2011, Chen et al. 2012, Chen et al. 2013), and POLII pause release (Gardini et al. 2014). In HeLa cells, INTS12 was shown to be required for the maintenance of perinuclear dynein (Jodoin, Sitaram et al. 2013). The general consensus regarding canonical INTS12 function is that it is primarily involved in snRNA processing but no evidence of this function has been shown in the human models. Therefore, the aim of the work described in this chapter was to investigate the potential role of INTS12 in the snRNA processing pathway using human *in vitro* HBEC model and compare the results to what was elsewhere observed in *Drosophila* cells.

4.1.1 Integrator Complex subunit 12 contribution to *Drosophila* small nuclear RNA processing

Ezzeddine et al. used the U7-GFP reporter construct to determine the role of INTS12 in snRNA processing (Ezzeddine et al. 2011; see section 1.8.4). There was some ambiguity around the importance of INTS12 in this process. One set of data presented, showed a weak GFP signal, but

another set of data showed a strong signal after INTS12 knockdown, using different controls but otherwise under the same conditions (Ezzeddine et al. 2011). Western blotting has shown the lowest GFP band density in the INTS12 silenced cells compared to the rest of the INTScom subunits knockdowns, except INTS3 and INTS10 that were not implicated in processing at all (Ezzeddine et al. 2011). qPCR results revealed some level of endogenous misprocessed snRNAs in INTS12 silenced cells (Ezzeddine et al. 2011). As an example, the level of fold increase in U1 misprocessed transcript relative to control for INTS12 silenced cells was almost the same as for INTS10, which was not implicated in snRNA processing.

Chen et al. worked on *Drosophila* S2 cells and confirmed the previous finding of INTS3 and INTS10 being dispensable for snRNA processing (Chen et al. 2012). In this study INTS12 was used as a GFP-reporting positive control and a reference for relative GFP levels. Western blots and qPCR analyses of INTS12 depleted samples showed significant misprocessing of endogenous U1 and U5 but not as high as in INTS9 depleted samples (Chen et al. 2012).

In another study, Chen et al. asked which of the INTS12 domains are implicated and which are dispensable in snRNA processing (Chen et al. 2013). First, they have demonstrated that in S2 cells knockdown of INTS12 results in positive signal using U7-GFP and U4-GFP reporters (Chen et al. 2013) as well as in accumulation of endogenous misprocessed U1, U2, U4, U5 snRNAs (Figure 4.1; Chen et al. 2013). Misprocessed U6 was not detected in INTS12 depleted cells because U6 is transcribed by RNA polymerase III (RNAPIII) and INTScom is not recruited onto this polymerase. To identify INTS12 domains responsible for correct pre-snRNA cleavage researchers devised an RNAi-rescue model where endogenous INTS12 is knocked down and snRNA processing is restored through introduction of RNAi-resistant wild type INTS12 (INTS12*). The group was successful in showing that there is a dose-dependent response in processing, depending on the concentration of transfected INTS12* after full knockdown of the endogenous form (Chen et al. 2013).

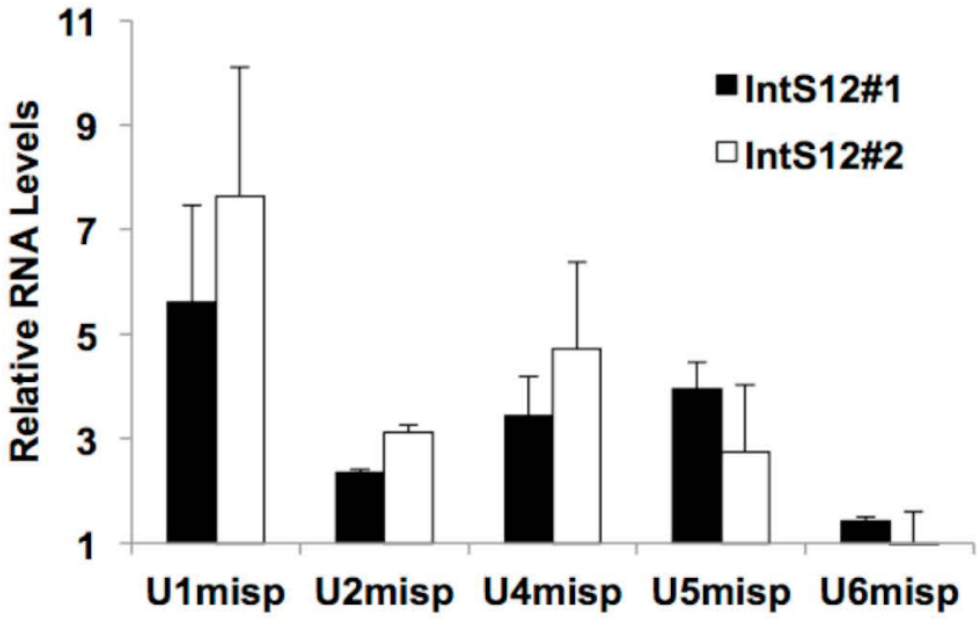


Figure 4.1: INTS12 knockdown resulted in increased levels of endogenous misprocessed U1, U2, U4, and U5 but not U6 snRNAs. Reproduced from Chen et al. 2013.

Having validated GFP reporting of misprocessing assay, Chen et al. produced RNAi-resistant and truncated forms of INTS12 and looked at processing efficiency of these truncations (Chen et al. 2013). Surprisingly, the well-conserved PHD domain was found to be dispensable for snRNA processing and not able to rescue INTS12 depletion. However, an N-terminal microdomain between 15th and 45th residue of *D. melanogaster* INTS12 was required for INTScom snRNA processing activity. This N-terminal microdomain has 83% sequence similarity with its human orthologue counterpart (Figure 4.2).

The fact that INTS12 PHD domain is dispensable for snRNA processing suggests that the protein is more than likely to be involved in other processes unrelated to snRNA processing in *D. melanogaster* cells and probably in another species. Although PHD domains typically bind to histone H3 and the presence of a stable nucleosome between DSE and PSE elements of snRNA genes has been reported (Stünkel et al. 1997), the weight of evidence argues for the lack of histones within snRNA genes (Chen et al. 2013), indicating possible genome binding independent from snRNA genes.

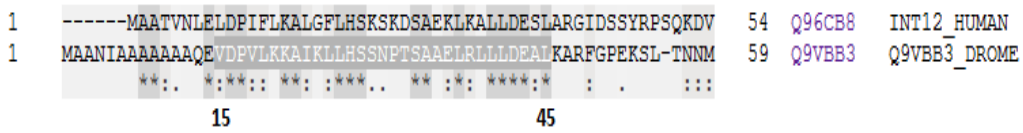


Figure 4.2: Level of human and *Drosophila* INTS12 orthology: pairwise alignment of human and fly INTS12 N-terminal microdomain that was found both necessary and sufficient for exogenous snRNA processing appears to have 50% of residues identical while 85% of residues have similar biochemical properties. The degree of sequence identity and similarity between the full length proteins is 26% and 42% respectively.

4.1.2 Appraisal of data suggesting INTS12 requirement for snRNA processing

Overall the evidence purporting to suggest that INTS12 is required for snRNA processing is not consistent. Considering the data from Ezzaddine et al. the effect of knockdown of various INTScom subunits on endogenous snRNA processing, it is possible to say that INTS12 has a relatively minor role in *D. melanogaster* INTScom activity in comparison to other subunits. When using U7-GFP reporter system for monitoring snRNA misprocessing the data is inconsistent as INTS12 knockdown sometimes resulted in a GFP signal, while sometimes did not result in GFP expression. Crucially, immunoblotting for GFP following the knockdown of various INTScom subunits suggests that INTS12 has a negligible role within INTScom as an important component of snRNA processing machinery.

It is worthy to point that in all available studies that investigated the role of INTS12 in endogenous snRNAs processing the underlying assumption is that the relative levels of misprocessed snRNAs is an accurate proxy of INTScom activity. It is not entirely clear why that may be the case since increased levels of immature snRNAs may be observed due to increased transcription and not just misprocessing. Therefore, there is no way to know from the published data whether investigators have induced the transcription of these snRNAs or observed genuine misprocessing.

However, the general prediction still holds true that if snRNA processing occurs, significant increase in the levels of immature snRNAs is

expected and there are no obvious reasons to suppose that the rates of general gene transcription vary between control and test conditions. If it is accepted that levels of immature snRNAs are a proxy for INTScom activity then it is possible to say that INTS12 play some but moderate role in the INTScom activity and inconsistencies in the available data warrant more investigations. Finally, all the studies looking at the role of INTS12 in snRNA processing were performed on *D. melanogaster* cells and nothing is known about the requirement of INTS12 in snRNA processing in human cells, despite other INTScom members shown to be involved so in HeLa model (Baillat et al. 2005).

4.1.3 Aims and Objectives

The overall objective of this chapter is to assess whether INTS12 is required for endogenous snRNA processing in primary human lung cells. Thus the fundamental question is whether this biological function is conserved between human and *D. melanogaster* species since the split of their common ancestor or whether this function was acquired or lost in each lineage independently. This path of thinking follows the parsimony principle which asserts that simple scientific explanation that fits the evidence is preferable than complex explanation (Farris, 2008). Therefore, preservation of snRNA processing in both species would imply functional inheritance from common ancestral organism as it is a simpler explanation than two independent evolutions of this function. In particular the aim of this chapter was to repeat the INTS12 knockdown experiment of Chen et al. but using a human cell model rather than the *D. melanogaster* model and compare results to Figure 4.1 (Chen et al. 2013).

4.2 Materials and Methods

INTS12 silencing in HBECs was performed according to the 120h knockdown protocol (see section 2.2.2.3). Total RNA was extracted and cDNA synthesised as described before (see sections 2.3.2 and 2.3.3). qPCR assays targeting U1, U2, U4 and U5 snRNA misprocessed transcripts were designed (see section 2.3.5.4) such that forward primer

spans the U coding site (i.e. incorporated into the spliceosome) while the reverse primer spans the U non-coding site (i.e. not incorporated into the spliceosome) which physically is present downstream of the coding site (Figure 4.3). Developed primers (Table 6 of Appendix) were validated bioinformatically and experimentally (see below). SYBR® Green qPCR reactions were carried out (see section 2.3.5.4) using these primers on the test samples in order to assess the functional requirement of INTS12 in processing of U1, U2, U4, and U5 snRNAs.

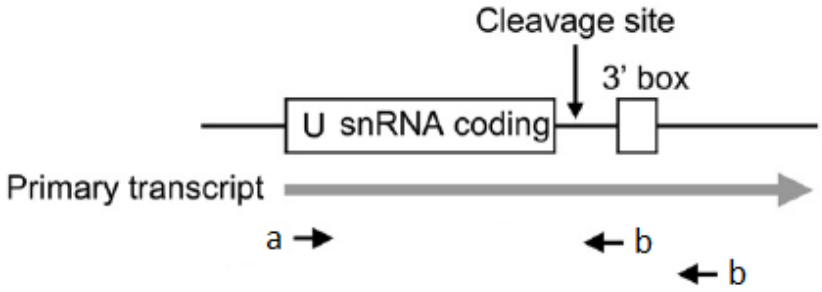


Figure 4.3: Design principles of snRNA processing. Forward primer 'a' anneals to the U snRNA coding site while primer 'b' is complementary to the site downstream of cleavage site rending the assay specific for primary (immature) snRNA transcript. Reverse primer 'b' may partially overlaps with 3'box element or be downstream of 3'box element. Reproduced from Hata and Nakayama 2007.

4.2.1 Development and validation of primary U1, U2, U4, and U5 snRNA qPCR assays

In the human genome there are multiple copies of genes and pseudogenes belonging to the same species of single U snRNA. These genes and pseudogenes are believed to be paralogs that have arisen through gene duplications (O'Reilly et al. 2012). For example, there are at least 14 genes and pseudogenes for U1 snRNA in the human genome. It turns out that most of them, including the pseudogenes, produce fully functional transcripts (O'Reilly et al. 2012). In other snRNA species the pseudogenes are not transcriptionally active. It is difficult to ascertain beforehand which copies of the genes or pseudogenes are expressed and therefore it is advantageous to design the qPCR assay such that it can detect most of the paralogous genes. For illustration, Figure 4.4 shows the multiple sequence alignment of human paralogous

U1 snRNA genes and pseudogenes. As it can be seen, there is a very good conservation of U1 coding site (position 1 to 168 which is incorporated into spliceosome) between all human U1 genes and pseudogenes. However, the conservation is poor beyond the coding sites, having less than 34% of sequence identity (Figure 4.4). The same situation is true for U2, U4, and U5 snRNA loci (data not shown). This simple observation suggests that the multiple copies of the same snRNA species have functional roles and thus are likely to be expressed. Also the fact that the fragments downstream of the coding site are not under strong evolutionary conservation suggests that these sites, including the canonical 3'box, are functionally less important.

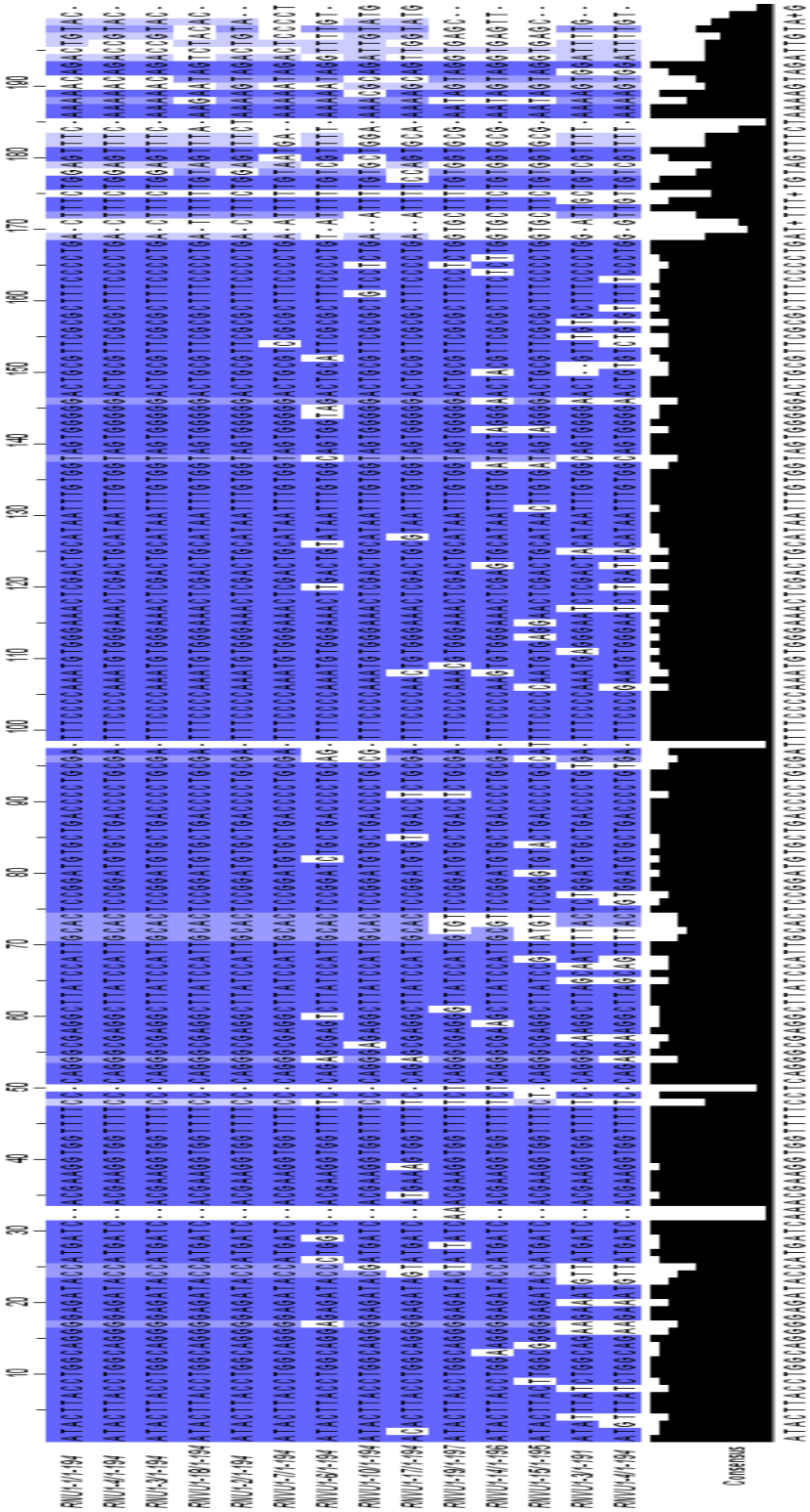


Figure 4.4: Conservation of paralogous human U1 snRNA genes and pseudogenes. Strength of the blue colour and the height of consensus columns per residue show the evolutionary conservation. The coding site ends at the annotated residue number 168. The sequence appears well conserved within (91% sequence identity) the coding site but poorly (33% sequence identity) conserved downstream of the coding site making the design of common assay a formidable challenge.

4.2.1.1 Primer design

Although it may be possible to design a forward primer common to all copies of the considered snRNA, the sequence divergence beyond the coding site means that it is impossible to design a working reverse primer that would anneal to all possible immature transcripts. Nevertheless, primers were designed to detect as much as possible of the immature transcripts (Table 4.1). Primer sequences were blasted on NCBI primer blast site (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) using immature U snRNA sequence as PCR template and NCBI Transcript Reference Sequences limited to *H. sapiens* as database to test for specificity. The rest of parameters were set at default settings. Primer pairs for all snRNAs were found to be specific to input template as no other targets were found in the selected database. Important primer parameters were ensured to be as close as possible to the preferable range (see section 2.3.5.4).

4.2.1.2 qPCR assay validation

The primary U1, U2, U4 and U5 SYBR® Green qPCR assays were validated on 1:2 serially diluted cDNA starting with neat sample originally prepared through cDNA synthesis reaction using 1µg of total RNA obtained from basal P3 HBECs (see section 2.3.4). Thus the expected difference in C_t values between each sample in the series of dilutions is 1 for 100% efficient assay. Assay specificity was tested by dissociation curve analysis (see section 2.3.5.2) and by checking the size of generated PCR product versus the expected size by agarose gel electrophoresis (see section 2.3.1).

4.2.1.3 Assessment of snRNA processing following INTS12 depletion in HBECs

Levels of misprocessed snRNAs were calculated using $\Delta\Delta C_t$ method of qPCR analysis (see section 2.3.5.3). Expression was represented as relative to scrambled D-siRNA condition and was GAPDH normalized (see section 3.6.1.3).

Annotated snRNA gene / pseudogene sequence	snRNA species
>RNU1-1 (NR_004430.2) ATACTTACCTGGCAGGGGAGATACCATGATCACGAAGGTGGTTTTCCCAGG GCGAGGCTTATCCATTGCACTCCGGATGTGCTGACCCCTGCGATTTCCCCA AATGTGGGAAACTCGACTGCATAATTTGTGGTAGTGGGGGACTGCGTTCGC GCTTTCCCCTGACTTTCTGGAGTTTCAAAAACAGACTGTAC	U1 snRNA
>RNU1-4 (NR_004421.1) ATACTTACCTGGCAGGGGAGATACCATGATCACGAAGGTGGTTTTCCCAGG GCGAGGCTTATCCATTGCACTCCGGATGTGCTGACCCCTGCGATTTCCCCA AATGTGGGAAACTCGACTGCATAATTTGTGGTAGTGGGGGACTGCGTTCGC GCTTTCCCCTGACTTTCTGGAGTTTCAAAAACAGACCGTAC	
>RNU1-3 (NR_004408.1) ATACTTACCTGGCAGGGGAGATACCATGATCACGAAGGTGGTTTTCCCAGG GCGAGGCTTATCCATTGCACTCCGGATGTGCTGACCCCTGCGATTTCCCCA AATGTGGGAAACTCGACTGCATAATTTGTGGTAGTGGGGGACTGCGTTCGC GCTTTCCCCTGACTTTCTGGAGTTTCAAAAACAGACCGTAC	
>RNU2-1 (NR_002716.3) ATCGCTTCTCGGCCTTTTGGCTAAGATCAAGTGTAGTATCTGTTCTTATCA GTTTAATATCTGATACGTCCTCTATCCGAGGACAATATATTAAATGGATTT TTGGAGCAGGGAGATGGAATAGGAGCTTGCTCCGTCCACTCCACGCATCGA CCTGGTATTGCAGTACCTCCAGGAACGGTGCACCCCCTCCGGGGATACAAC GTGTTTCCTAAAAGTAGAGGGAGGTAAGAGACGGTAGCACCTGCGGGGCGG CTTGCACGCCGAGTGCCTGTGACGCGCCGGCTTAACTTAACTGCTTCCCTG AAGTACCTTGAGGTTCCTGATGTGCGGGCGGTAGACGGTAGGCTTATGCGG CACGCTGTCG	U2 snRNA
> RNU4-1 (NR_003925.1) AGCTTTGCGCAGTGGCAGTATCGTAGCCAATGAGGTCTATCCGAGGCGCGA TTATTGCTAATTGAAAACTTTTCCCAATACCCCGCCGTGACGACTTGCAAT ATAGTCGGCACTGGCAATTTTTGACAGTCTCTACGGAGACTGAATTTTCTT GCAGTTGAACAACAGAGGCTT	U4 snRNA
> RNU5A-1 (NR_002756.2) ATACTCTGGTTTCTCTTCAGATCGCATAAATCTTTCGCCTTTTACTAAAGA TTTCCGTGGAGAGGAACAACTCTGAGTCTTAACCCAATTTTTTGAGGCCTT GCTTTGGCAAGGCTATATGTGGTAATCCAACAATAGAAATTATTT	U5 snRNA

Table 4.1: The binding sites onto various snRNA genes and their paralogous copies for forward (RED) and reverse (PINK) primers. Mature U1 snRNA is shown in bold. 3'box is shown in green for U1 and U2 as no recognizable 3'box sequence was observed in either U4 or U5 snRNA genes.

4.3 Results

4.3.1 snRNA processing assays validation

On average there was 1 C_t difference between each sample of the 1:2 dilution series. Thus assay efficiencies associated with the designed primers were 89±9%, 75±5%, 104±11% and 95±4% for U1, U2, U4 and U5 snRNA assays respectively (n=3). In general, more diluted samples appear to have greater technical variability as the assay is approaching the limit of reliable detection (Figure 4.5). Dissociation curve analysis of the amplified PCR amplicons produced a single distinctive peak providing evidence for a generation of single amplicon indicating reaction specificity (Figure 4.6). PCR products were electrophoresed with DNA ladder to test their size and compare to the predicted size. All the amplicons appeared to have the predicted size for U1 and U5 snRNAs (within acceptable electrophoresis assay margin). U4 and U2 snRNAs were close to the predicted ladder albeit slightly off (Figure 4.7). This could have been due to redundancy of snRNA genes with multiple transcriptionally active loci presumed to be pseudogenes, as explained in section 4.2.1 and exemplified in Figure 4.4. The U4 and U5 DNA amplicons were isolated from gel and sequenced by Sanger reaction. In correspondence with observed gel electrophoresis result, the generated U5 sequence had 98% identity with reference NCBI sequence (X01691.1) after a BLAST search. On the other hand, U4 presented evidence of presence of more than one set of reaction products that were seen on sequencing chromatogram. This supports the hypothesis that heterogeneity of U4 product comes from sequences beyond the snRNA coding site (similarly to what is seen for U1 in Figure 4.4) and that different U4 loci are being expressed. The U1 and U2 amplicons were not sequenced. Overall, it is possible to say that the overall evidence shows that the developed qPCR assays aimed at measuring misprocessed U1, U2, U4 and U5 snRNAs are reliable for $\Delta\Delta C_t$ method of analysis.

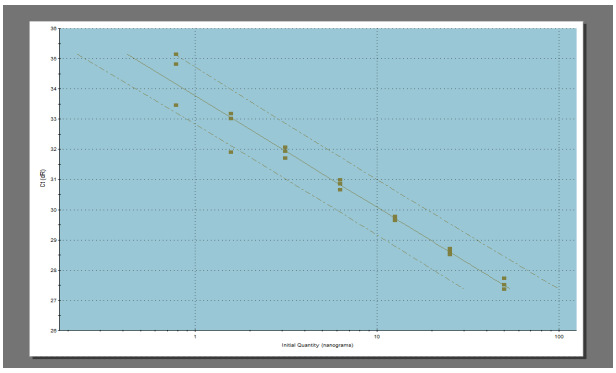
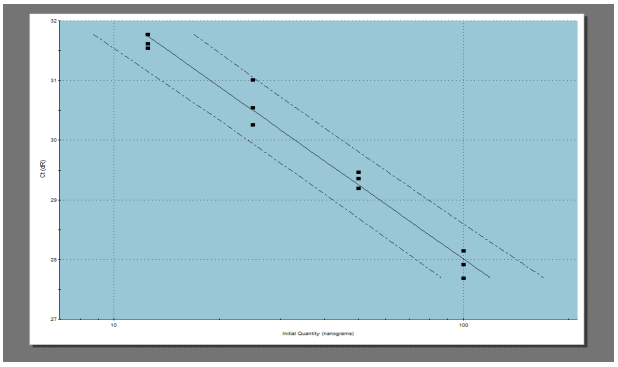
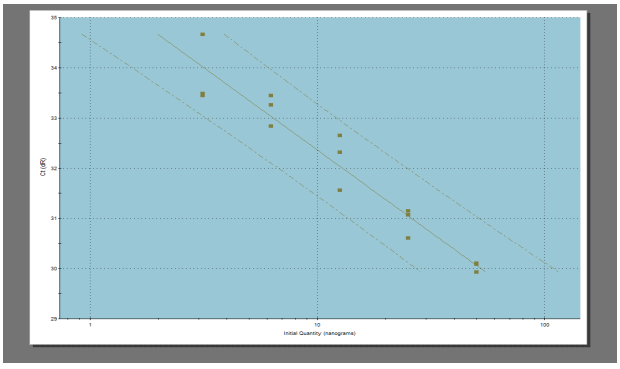
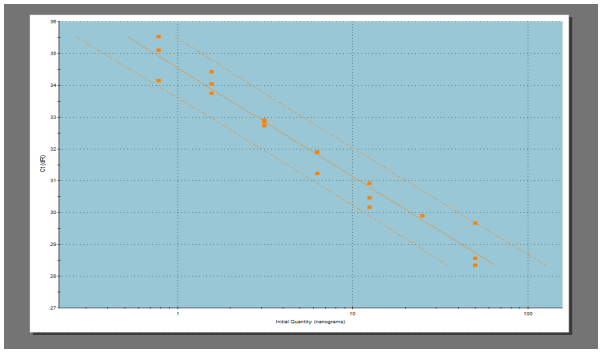
snRNA species	Calibration curve	Assay efficiency
U1		89±9%
U2		75±5%
U4		104±11%
U5		95±4%

Figure 4.5: Experimental validation of the efficiency of the designed primary snRNA primers. The intercepted lines above the main slope are its upper and lower limits of 95% confidence interval. Greater number of samples in 1:2 dilution series is different between some assays and is indicative of assay's range of detection. For example, U1 assay has 7 samples whilst U2 assay has 4 samples as U1 assay can reliably measure target levels with less than 1ng of cDNA, while U2 assay's limit is 10ng of cDNA

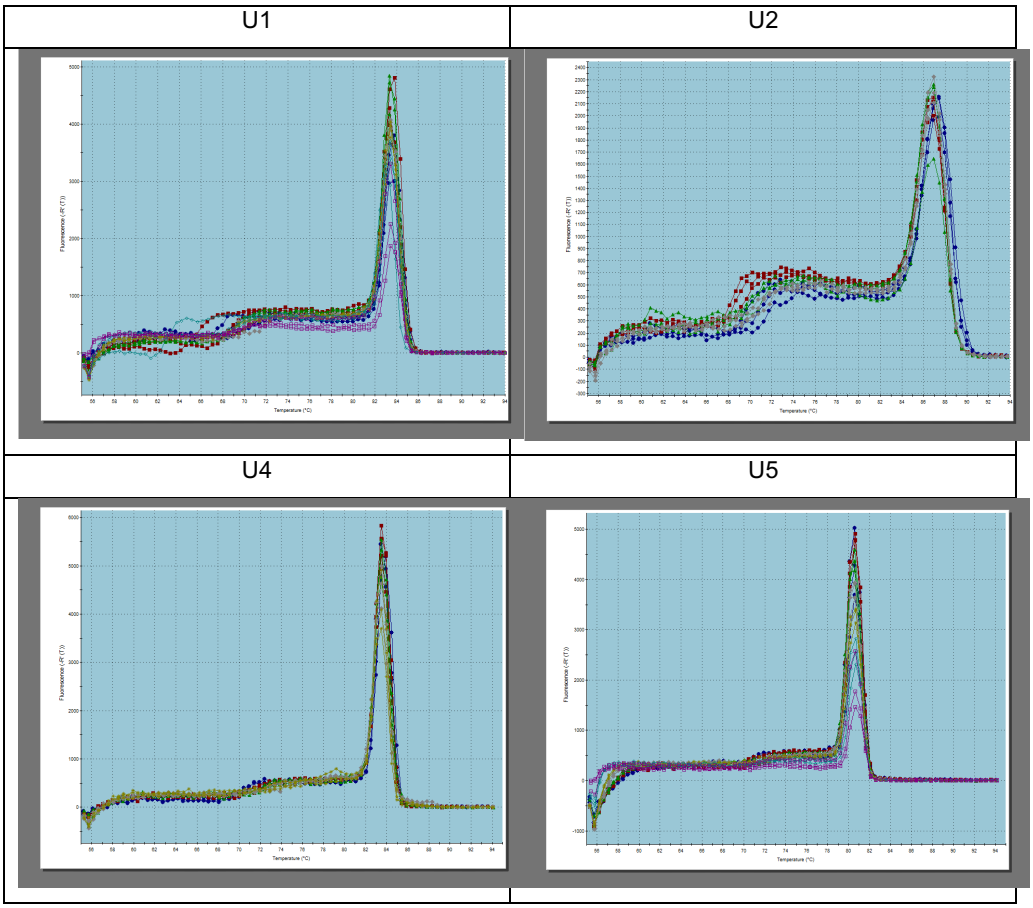


Figure 4.6: Dissociation curve analysis of PCR amplified snRNA amplicons.

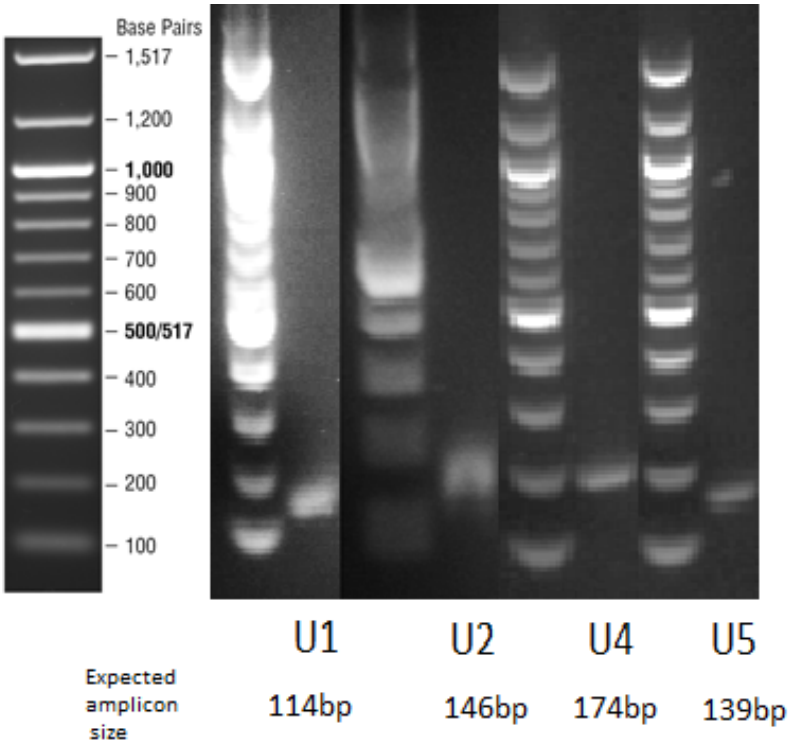


Figure 4.7: The location of amplified primary snRNAs amplicons relative to the DNA ladder after 40 PCR cycles. The cDNA template used was representative of total RNA content of P3 un-treated HBECs.

4.3.2 INTS12 plays a modest role in snRNA processing in human bronchial epithelial cells

As mentioned, given previous observations in *D. melanogaster* suggesting a role for INTS12 in processing of U1, U2, U4 and U5 snRNAs, initial studies set out to determine if the major functional and regulatory role for INTS12 in HBECs involved snRNA processing. Transfection of primary cultures of HBECs with D-siRNAs A and C produced 91±2% and 82±3% knockdown respectively (Figure 4.8; P<0.0001, n=6). In contrast to findings in *D. melanogaster* cells (Figure 5), no significant effects on U1 processing were seen. A role for INTS12 on U2 processing was found, with increases in U2 immature product by 2.58±0.58 fold and by 2.64±0.59 for D-siRNAs A and C respectively (Figure 4.9, P<0.05, n=6). However, in keeping with the lack of effect on U1 processing, no significant effects of INTS12 knockdown on processing of U4 and U5 snRNAs were observed. These data suggest that whilst INTS12 may play a minor role in U2 processing, it does not have a major general role in HBEC snRNA processing.

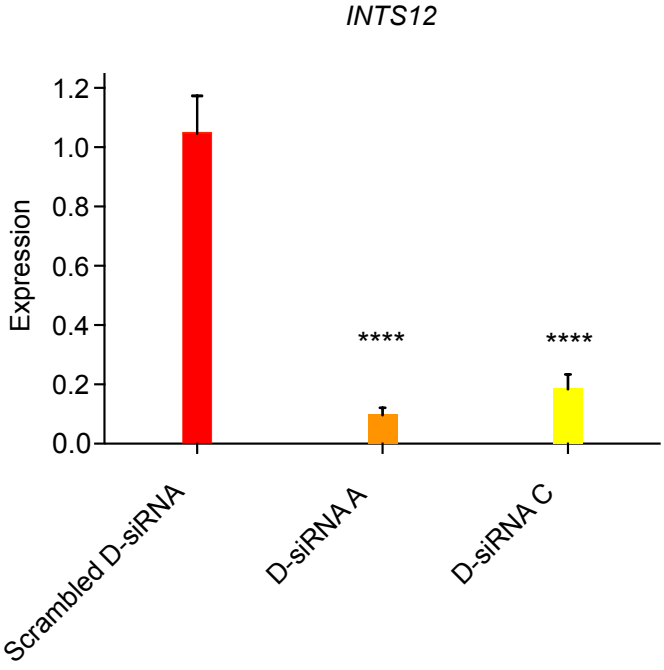


Figure 4.8: *INTS12* ΔΔCt expression in HBECs transfected with D-siRNA A and C. Statistical tests were performed comparing to scrambled D-siRNA control: ****P<0.0001. Individual ΔΔCt gene expressions are GAPDH normalized and relative to the mean of scrambled D-siRNA condition. Error bars represent standard error of the mean.

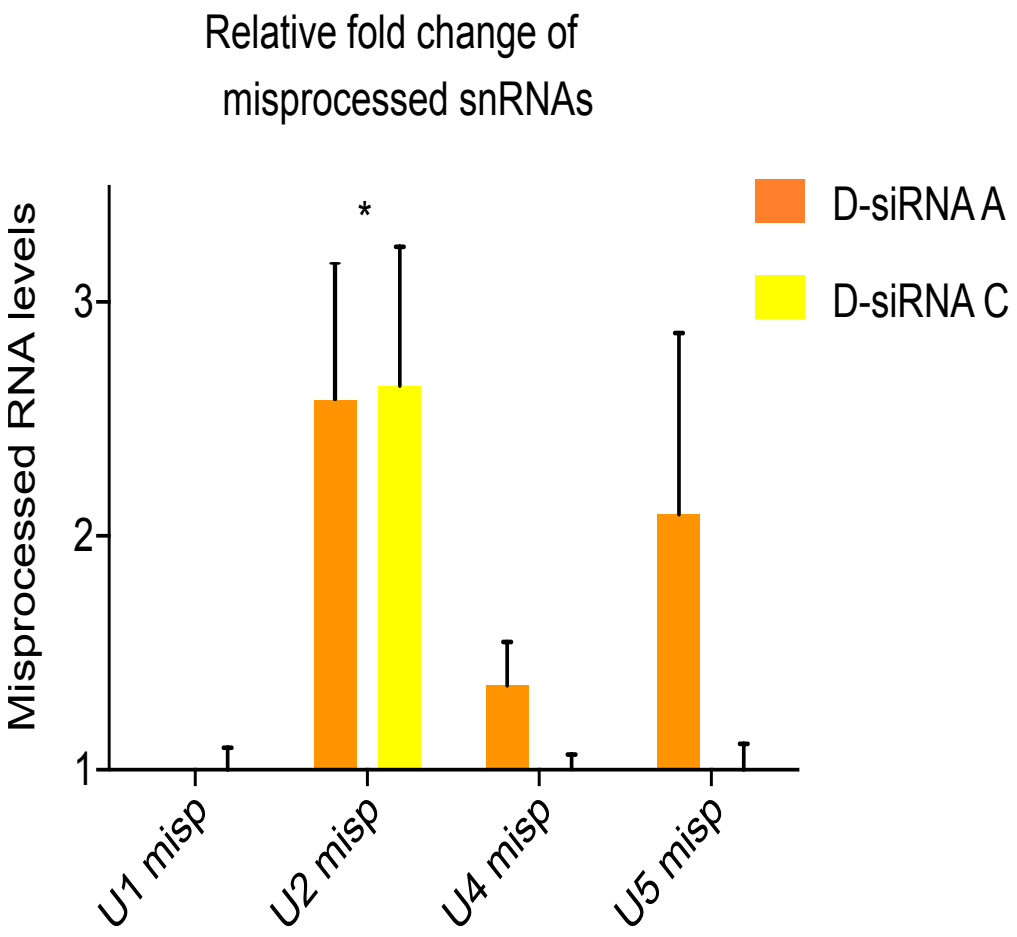


Figure 4.9: $\Delta\Delta C_t$ fold changes of misprocessed snRNAs. Statistical tests were performed comparing to scrambled D-siRNA control: *P<0.05. Individual $\Delta\Delta$Ct gene expressions are GAPDH normalized and relative to the mean of scrambled D-siRNA condition. Error bars represent standard error of the mean.

4.4 Discussion

A direct comparison of effects of INTS12 depletion on snRNA processing in human and *D. melanogaster* S2 cells provides evidence only for a role in U2 snRNA processing. Thus although it appears that in the fly INTS12 plays a general role in processing snRNA, this does not appear to be the case in human cells. Possible explanations for this difference include a loss of a common ancestral function of INTS12 in the lineage leading to human species or snRNA processing activity may have been acquired in the lineage leading to fly species. Both hypotheses are equally parsimonious but in the light of minimal role of INTS12 in U2 snRNA processing in HBECs, the former hypothesis seems more likely. Although the presented evidence shows INTS12 to have a modest role in human snRNA processing, it is not possible to exclude the possibility that other INTScom members are important in delivering this molecular function. This is especially crucial in the light of data showing a more prominent role for INTS1, INTS4 and INTS9 in *Drosophila* snRNA processing (Ezzedine et al. 2011). Human orthologues of these genes could be required more fundamentally for 3'end formation of snRNAs and further studies are warranted to address this question. Moreover, in relation to results in human cell presented in this Chapter, data purporting to imply that the knockdown of *Drosophila* INTS12 impairs snRNA processing are not consistent (Ezzedine et al. 2011, Chen et al. 2012) highlighting the demand to carefully interpret aforementioned fly and human findings.

As mentioned in Chapter 3, cross metazoan sequence analysis of INTS12 proteins revealed high levels of conservation, particularly of the INTS12's PHD motif. The minor role in snRNA processing in human cells and strong conservation suggests the existence of additional functions for INTS12. Furthermore, the evolutionary constrained PHD finger is dispensable for snRNA processing even in *D. melanogaster* implying other possible functions for this protein even in this organism (Chen et al. 2013).

Therefore, it is possible to say that the potential contribution of INTS12 to lung function phenotypes is unlikely to be driven via snRNA processing pathway. At the initial stages of this thesis the guiding hypothesis was that genetically determined deficiency of INTS12 activity leads to compromising splicing of mRNAs enriched for genes known to be critical for pulmonary health (e.g. SERPINA1). This may provide biological explanation of INTS12 contribution to lung function. However, it is improbable that allele carriers with low *INTS12* expression contribute to altered lung function through snRNA misprocessing. As INTS12's PHD domain appeared to be homologous to a large family of PHD fingers (Table 3.3) whose functions lie in the control of chromatin and nucleosomes and where they act as epigenetic regulators of gene expression (Bienz, 2006), the next step in this thesis was to examine the genome-wide regulatory properties of INTS12. RNAseq was used to identify cellular networks whose homeostasis may become disrupted as a result of INTS12 knockdown aiding the generation of gene function hypotheses.

5. Inferring gene and pathway dysregulation in INTS12 depleted HBECs

5.1 Introduction

The dispensability of the evolutionary conserved INTS12 PHD motif domain for snRNA processing strongly suggested existence of other unrealized INTS12-specific molecular and/or cellular functions (see Chapter 3). This is true for *D. melanogaster* S2 cells in which INTS12 was shown to be moderately required for snRNA processing (for critical appraisal of scientific literature of studies that investigated INTS12 requirement for snRNA processing see section 4.1.2). However, it is even more relevant in human cells in the light of the data presented in Chapter 3, where entire endogenous INTS12 appears to be dispensable for snRNA processing among 75% of the tested snRNA species. The key questions are, why does INTS12 show so much evolutionary constraint, and why is it necessary for mammalian development as shown by pre-weaning lethality in homozygous mouse models (Obeidat et al. 2013)?

Beyond the canonical function, in the HeLa cell model INTS12 was shown to be involved in perinuclear dynein dynamics (Jodoin, Sitaram et al. 2013). Interestingly, in a separate study, Jodoin, Shboul et al. reported that INTS12 was also implicated in the maintenance of epithelial cell ciliary function which is thought to arise from a common process controlling both perinuclear dynein as well as primary cilia (Jodoin, Shboul et al. 2013). Dyneins generate force and movement on microtubules in a wealth of cellular processes including cell division (Roberts et al. 2013). Thus a pool of dynein molecules is present in the nucleus and INTScom, including INTS12, was shown to be required for its maintenance near the nuclear membrane (hence the name perinuclear dynein). Jodoin, Sitaram et al. proposed a human relevant cell model in which INTScom mediates 3'-end processing of snRNA, which in turn affects the splicing machinery required for normal processing of mRNAs encoding key regulators of cytoplasmic dynein localization (Jodoin, Sitaram et al. 2013). Thus when INTScom is compromised, the production of critical transcripts is reduced leading to a reduction of perinuclear dynein. However, this seems unlikely to be the

case for INTS12 considering the above mentioned minor role for INTS12 in snRNA processing. In fact, despite finding several thousand genes with evidence of differential splicing in HeLa cells depleted of INTS11, researchers were unable to detect enrichment for gene sets related to dynein–dynactin subunits or adaptor proteins in overrepresentation analysis (Jodoin, Sitaram et al. 2013). This highlighted the need for further studies of INTScom activities.

Although it is possible that what is driving INTS12 conservation is its requirement for the perinuclear dynein maintenance, the sequence homology search described in Chapter 3 suggests that it may also have epigenetic and gene regulatory roles (Table 3.3) beyond the function discovered by Jodoin, Sitaram et al. (Jodoin, Sitaram et al. 2013). For example, INTS12's PHD has similarity to histone-lysine N-methyltransferase 2A, a known epigenetic modifier. The studies described in this chapter concentrated on trying to elucidate potential genome-wide regulatory properties of INTS12 as well as their possible relationship to lung function.

5.1.1 Systematic INTS12 function discovery – aims and objectives

As mentioned in Chapter 1, one way to hypothesise about gene function is by measuring global gene expression following controlled and experimental perturbation of the gene. The observed transciptomic signature can then be used to test whether cellular homeostasis has been altered and what pathways may have been responsible for this alteration. Phenotypic assays may then be chosen based on these findings. The aim of this Chapter is to use the hypothesis-free GSEA method (Subramanian et al. 2005; see section 1.7.4) in combination with INTS12 depletion to accomplish the task of gene function discovery. Technical and biological validation of the data is also presented. A more candidate-driven approach was also relied on by comparing the acute versus longer term transcriptomic responses due to INTS12 knockdown with a goal of identifying genes important in lung biology. Based on the transcriptomic data, functional phenotypic assays were chosen and used

to quantitatively measure the relevant phenotype in INTS12 depleted cells. Moreover as part of NGS-based RNAseq, novel INTS12 mRNA variants have been identified and tested by PCR and Sanger sequencing for their validation.

5.2 Materials and Methods

5.2.1 RNAseq

INTS12 silencing in HBECs was performed according to the 48h and 120h knockdown protocols (see section 2.2.2.3). For main RNAseq and functional analyses the effects of INTS12 depletion were assessed 120h after initiation of interference. To compare the acute and chronic transcriptomic responses to knockdown, RNAseq profiling was also performed 48h after the initiation of interference (see section 2.5.1). There were four experimental conditions: un-transfected cells, cells transfected with scrambled D-siRNA control, and cells transfected with D-siRNAs A and C. Each experimental condition was performed in three independent biological replicates. As mentioned before, the 48h and 120h experiments were performed upon two different donors, keeping the donor the same within each time point. The rationale behind that is having confidence in a list of genes that are reproducibly detected as differentially expressed after 48h as well as 120h. The genes that appear dysregulated in this manner can be called “core subset regulome of INTS12”, as their regulation is observed in different donors and different times.

5.2.2 RNAseq and Pathway Data Analysis

The quality of raw FastQ files (100 base pairs) was assessed on FastQC. Tuxedo analysis pipeline was used for RNAseq analysis (Trapnell et al. 2012; see section 2.8.1.1): (i) TopHat’s (v2.0.1254) Bowtie2 read alignment was performed upon hg19 build, (ii) Cufflinks (v2.2.1) transcriptome assembly was performed on individual sample basis and merged by Cuffmerge (v2.2.1) using reference-based assembly, (iii) Cuffdiff (v2.2.1) differential gene expression was performed using Cuffmerge-predicted annotation. Loci with Benjamin-Hochberg

corrected P value (Benjamin and Hochberg, 1995) below 0.05 were considered significant. Transcriptomic comparisons were performed comparing scrambled D-siRNA to each anti-INTS12 D-siRNA and comparing un-transfected cells with scrambled D-siRNA transfected cells in order to account for off-target and mere transfection effects respectively.

5.2.2.2 General methodology in the identification of reproducibly and INTS12-specifically perturbed genes and pathways

In order to perform pathway analyses, FPKM expression values were obtained for each gene per individual RNAseq sample using Cuffnorm (v2.2.1). Loci containing multiple amalgamated genes were separated into individual genes and had assigned the equivalent expression values, while genes occurring multiple times on the dataset had their expression values summated using in-house written python script (see section 2.8.3.1).

GSEA approach (Subramaniana et al. 2005) using 4722 curated gene sets including 1320 canonical pathway definitions from the Molecular Signatures Database (Kanehisa and Goto, 2000; Croft et al. 2014; Glaab et al. 2010; accessed Dec 2014) was used, comparing scrambled D-siRNA to each INTS12 D-siRNA and comparing un-transfected cells with scrambled D-siRNA transfected cells. To provide internal replication and account for off-target effects (Jackson and Linsley, 2010), GSEA analyses were performed separately following treatment with either D-siRNA A or C, comparing scrambled D-siRNA treated cells to INTS12 depleted cells. Additionally, un-transfected cells were compared with scrambled D-siRNA treated cells to account for pathways that may be altered following treatment with non-specific D-siRNA as artefacts of the experimental exposure rather than being causally related to the gene knockdown. Pathways with Benjamin-Hochberg corrected P value below 0.05 were considered significant (Benjamini and Hochberg, 1995). Gene sets reproducibly perturbed by both D-siRNAs (FDR<0.05) but not affected by scrambled D-siRNA treatment (FDR>0.05) were shortlisted

and finally top dysregulated pathways were identified relying on normalized enrichment score sorting.

Results of the pathway analysis were displayed in a Cleveland's plot using ggplot2 R package (version 2.10, http://ggplot2.org/) while pathway heatmaps were drawn using heatplus R package (Ploner, 2015). Boxplots were drawn using build-in R function. Pearson's correlations of gene expression were calculated using hmisc R package and drawn using ggplot2.

5.2.2.2.1 Identification of INTS12 depletion deregulated subset of genes

Comparison of acute and chronic transcriptomic responses to INTS12 knockdown aimed at identifying core subset of genes significantly differentially expressed in 48h and 120h time points respectively. The rational of the analysis was similar to pathway analysis, i.e. genes were shortlisted if there were reproducibly dysregulated in both INTS12 D-siRNAs but not in scrambled D-siRNA. Genes that were dysregulated in both INTS12 D-siRNAs in a given direction while in the opposite direction in the scrambled D-siRNA sample were also included. Core subset of genes was identified by determining the common genes between the 48h and 120h significant gene lists.

5.2.3 qPCR

Pre-developed qPCR primers (Table 3 of Appendix) were used for SYBR® Green qPCR assays as described in Chapter 2 (see section 2.3.5.4). Technical validation of RNAseq findings was performed using at least three cDNA replicates derived from total RNA used for 120h RNAseq experiment (hence using D195307). Biological validation of target genes was performed upon different donor cells (D7F3206) with at least three biological cDNA replicates. In accordance with the chosen housekeeper (see section 3.6.1.3), gene expression was GAPDH normalized and analysed using the ΔΔCt method (see section 2.3.5.3).

5.2.4 Functional assays

Rates of protein synthesis were measured as described before (see section 2.9.1). Cell proliferation was assessed by comparing total cell

counts at the beginning and at end of the knockdown experiment as described before (see section 2.9.2).

5.2.5 mRNA splice variant assembly and validation

In order to elucidate the genetic architecture of *INTS12* locus, Cuffmerge generated novel gene transfer format (GTF) annotation files (containing genomic coordinates of the splice variants) from 48h and 120h RNAseq datasets were compared to Ensembl GTF annotation of hg19 build by using Cuffcompare (v2.2.1). The average FPKM isoform expressions in the basal un-transfected HBECs was obtained from the 48h Cuffdiff dataset and used to quantify the relative abundances of each of the individual isoforms. Splicing graphs depicting novel and known splice transcripts were generated using SpliceGrapher (v0.2.457). The sequences of novel INTS12 variants were retrieved using predicted genome annotation (GTF files) and reference human genome sequence (hg19 build) via gtf_to_fasta feature of Cuff package. The general command used was:

```
gtf_to_fasta annotation.gtf genome.fa out_file
```

Forward and reverse primers were designed (Table 5.1; see section 2.3.5.4) in order to verify the existence of these three novel INTS12 variants. Two unique and novel splice junctions specific for individual isoform were leveraged and primers were designed to span these splice junctions. Among the assembled INTS12 variants, only these variants had two unique splice junctions that would enable their specific PCR amplification. Forward primer was common for all three isoforms while reverse primers were isoforms specific. 40 cycles PCR reaction was run upon the two different 48h (D7F3206) and 120h (D195307) donors' basal HBECs cDNA (see section 2.3.5.1), amplicons were electrophoresed, cut and extracted from a gel, and sequenced by Sanger sequencing (see section 2.3.6.1). Thus the identity of the novel variants was verified by comparing the actual band size to the predicted band size and by comparing RNAseq-derived sequence from Sanger sequencing-derived sequence. As part of the novel variant analysis it was necessary to access a specific genomic region given specific coordinates. This was

achieved with samtools using the following command, where the `genome.fa` is the hg19 fasta sequence of human genome:

```
samtools faidx genome.fa 4:[5'coord]-[3'coord]
```

Novel INTS12 isoform target ID	Oligo	Sequence
TCONS_57,56,54	Forward	5'-GTGGATGTCTTGACTTCTGT-3'
TCONS_57	Reverse	5'-GAACGGTGTCCCTAAGG-3'
TCONS_56	Reverse	5'-GAGATTGCCAGGCGTTTGCAATG-3'
TCONS_54	Reverse	5'-CGGAACGGTGTCCCTAAG-3'

Table 5.1: Sequences of the designed forward and reverse primers used to amplify the novel INTS12 splice variants. All three assays share the same forward primer and thus specificity is conferred by the reverse primer. Target ID correspond to the isoforms seen elsewhere in this thesis.

5.3 Results

5.3.1 Quality control of sequencing data

Before proceeding with RNAseq-derived identification of dysregulated pathways and genes it is important to ensure that the raw sequencing is of good quality. Therefore, the 48h and 120h RNAseq raw sequencing library was used to evaluate the median Q-score throughout the sequenced read body (see section 2.8.1 for explanation of the meaning of Q-scores). As there were four experimental conditions, each performed in three biological replicates there was a total of 12 samples in the 48h and 120h datasets. Since the reads were sequenced in sense and antisense directions (i.e. paired-end sequencing) there were two FastQ files per sample. Thus there was a total 48 files for the quality assessment and the representative results are shown in the Figure 6.1 below.

As it can be seen all median base calls in all the samples had quality score above 28 and this is true for both the 120h and 48h datasets (Figure 6.1). Therefore, the probability of error throughout the read is less than 0.2%. As the chance of sequencing error is low, the raw reads

were not trimmed from their 3'-end, were the quality of sequencing tends to decrease, and thus were directly aligned to the genome. Indeed, when the libraries were Bowtie2 aligned, the alignment rate was >70% for all the samples (Table 5.2) and the proportion of mappable reads which aligned to multiple locations was less 3% throughout the datasets. A survey of best practices for RNAseq data analysis has shown that in an "ideal experiment" we should expect between 70 and 90% of reads to map onto the human genome (Conesa et al. 2016). Thus the gapped and un-gapped BAM alignment file was suitable for Cufflinks and Cuffmerge transcriptome assemblies, Cuffnorm FPKM absolute gene expression quantification, and Cuffdiff differential gene expression analyses.

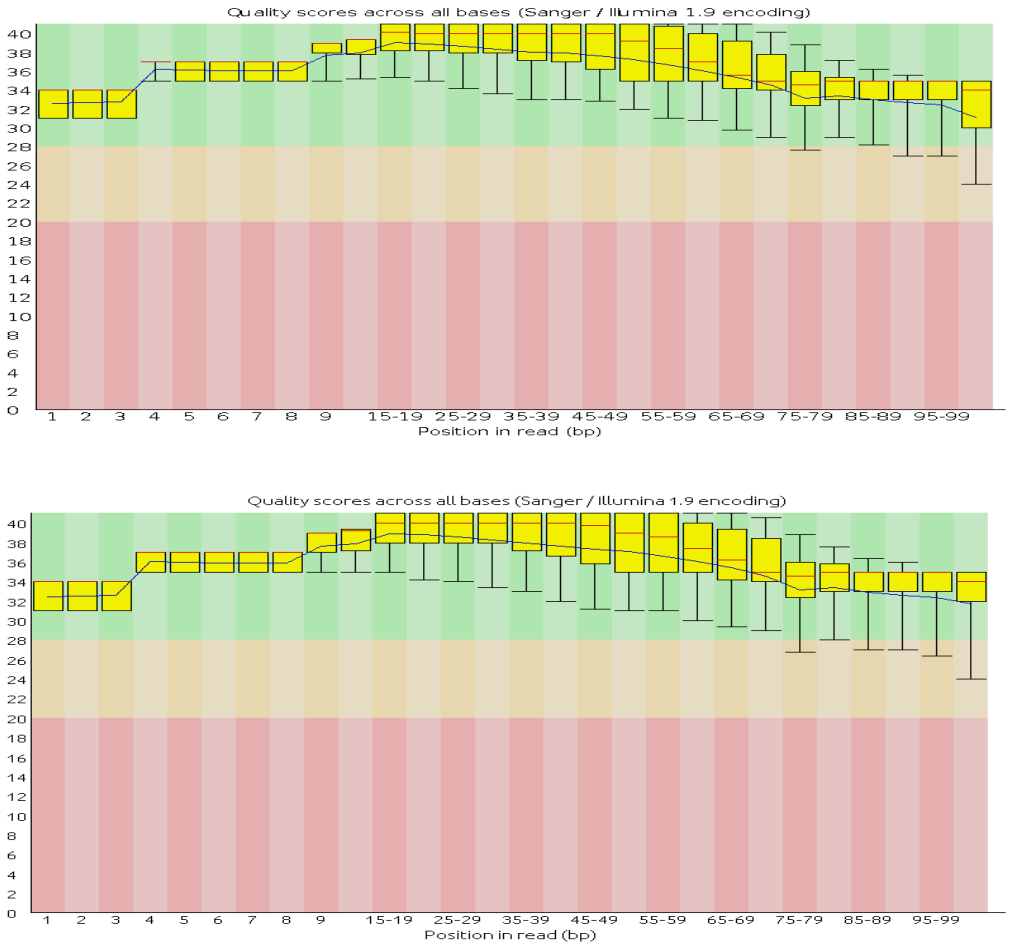


Figure 5.1: Representative Q-scores of RNAseq libraries. All FastQ files had median scores above 28 throughout the read length and in both sense and anti-sense directions. The panel represents the 120h dataset, while the bottom panel represents the 48h dataset.

RNAseq dataset	Replicate	Samples	Percent of mappable reads (concordant pair alignment rate)
120h dataset	Replicate 1	Un-transfected	83.1%
		Scrambled D-siRNA	83.6%
		D-siRNA A	82.7%
		D-siRNA C	85.7%
	Replicate 2	Un-transfected	86.0%
		Scrambled D-siRNA	83.6%
		D-siRNA A	83.4%
		D-siRNA C	84.9%
	Replicate 3	Un-transfected	83.7%
		Scrambled D-siRNA	86.6%
		D-siRNA A	82.7%
		D-siRNA C	84.9%
48h dataset	Replicate 1	Un-transfected	81.6%
		Scrambled D-siRNA	85.2%
		D-siRNA A	83.2%
		D-siRNA C	82.0%
	Replicate 2	Un-transfected	82.1%
		Scrambled D-siRNA	83.6%
		D-siRNA A	82.2%
		D-siRNA C	80.5%
	Replicate 3	Un-transfected	77.6%
		Scrambled D-siRNA	75.4%
		D-siRNA A	77.1%
		D-siRNA C	71.7%

Table 5.2: Alignment rates in RNAseq samples. The slightly lower rates of replicate 3 in 48h dataset may have ccured because of RNA degradation, but are still suitable for analysis according to the survey of RNAseq sample alignments (Conesa et al. 2016).

5.3.2 Differential transcriptome analysis reveals regulation of a core subset of genes relevant to airway biology

In order to identify a core subset of genes that are significantly regulated by INTS12 the acute versus longer term transcriptomic responses due to depletion were compared. RNAseq profiling was performed 48h and 120h after RNA interference (RNAi). After 48h the levels of knockdown were 74±1% and 78±2%, whilst after 120h, 89±1% and 80±2% for D-siRNAs A and C respectively (FDR<0.05; Figure 5.2). After accounting for off-target and transfection effects there were 67 and 1939 differentially expressed genes by INTS12 knockdown at 48h and 120h time points respectively (FDR<0.05; Figure 5.3 and Figure 5.4). These include differentially expressed genes by D-siRNA A and D-siRNA C in given direction but in opposite direction by scrambled D-siRNA. Thus, sustained knockdown resulted in a differential expression of ~30 times more genes than what was observed in acute response to knockdown (Figure 5.5). For those genes showing altered levels at both time points,

the magnitude of change was greater at 120h post initiation of RNAi (Figure 5.6) for all except one (Figure 5.7). Crucially the direction of differential expression for this set of genes is the same in the independent D-siRNAs treatments and at both time points (Figure 5.7, Table 5.2). Greater number of differentially expressed genes at 120h (Figure 5.2) relative to 48h (Figure 5.3) can be attributed to sustained knockdown having more pronounced effect on INTS12 regulome than in acute response to gene knockdown. It important to note that the number of differentially expressed genes only in scrambled D-siRNA is lower at 120h than at 48h suggesting that these are non-specific effects of transfection reagent.

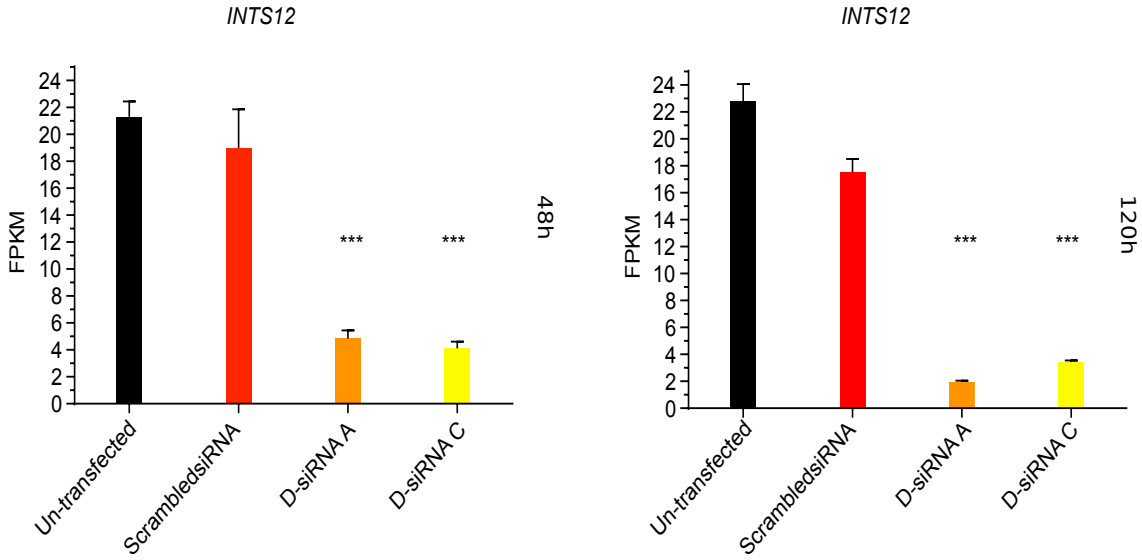


Figure 5.2: INTS12 knockdown at 48h and 120h post RNAi. Cuffdiff statistical tests were performed comparing to scrambled D-siRNA control and were FDR corrected for multiple comparisons: ***FDR<0.001.

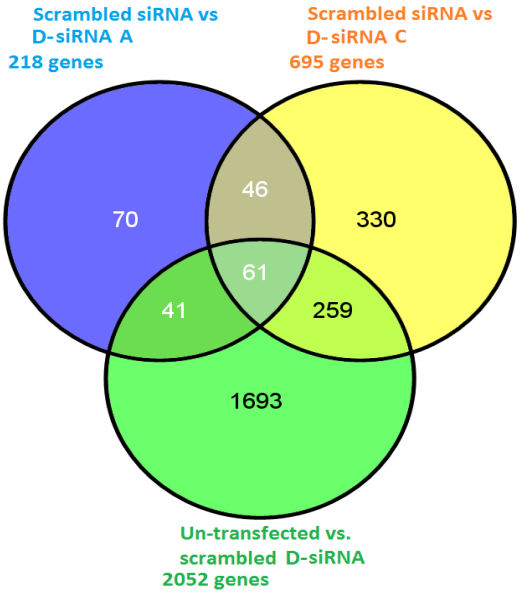


Figure 5.3: Venn diagrams of significantly deregulated genes at 48h. 46 reproducibly deregulated genes plus 21 genes deregulated in all three comparisons but in opposite direction in INTS12 knockdown conditions when compared to un-transfected vs. scrambled D-siRNA analysis were shortlisted from 48h dataset (total 67).

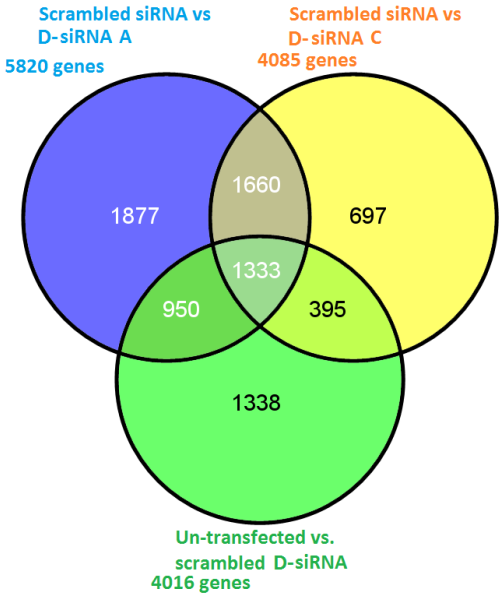


Figure 5.4: Venn diagrams of significantly deregulated genes at 120h. 1660 reproducibly deregulated genes plus 279 genes deregulated in all three comparisons but in opposite direction in INTS12 knockdown conditions when compared to un-transfected vs. scrambled D-siRNA analysis were shortlisted from 120h dataset (total 1939).

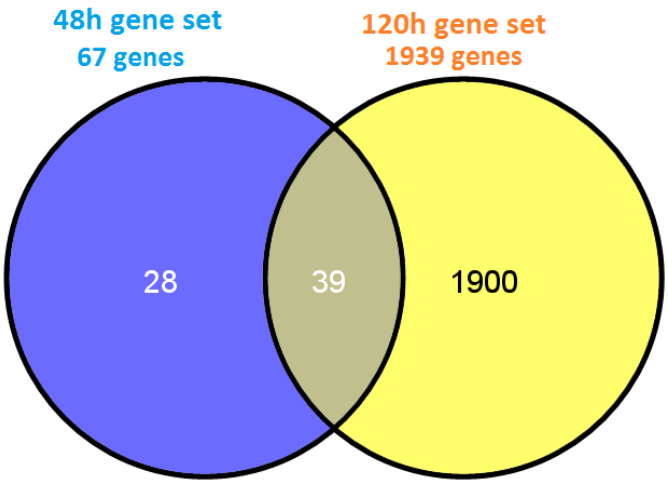


Figure 5.5: Comparison of 48h and 120h transcriptomic responses to INTS12 knockdown reveals dysregulation of key genes of importance for pulmonary physiology. Sustained depletion resulted in greater fold changes in respective expression in the same direction in both D-siRNAs treatments. The two gene sets contain 39 common genes of relevance to airway biology, i.e. shown to be important and/or critical to lung function and health based on the survey of literature data. D-siRNA A differential expression data shown.

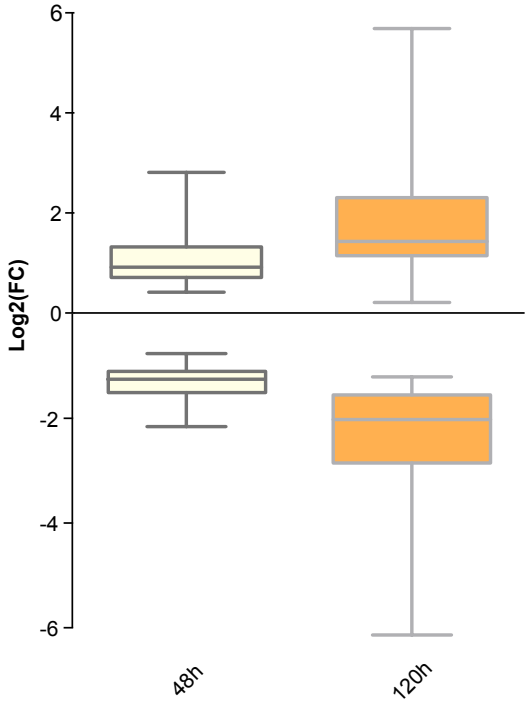


Figure 5.6: Box plot of log_2 fold changes of 39 genes significantly deregulated at 48h and 120h using D-siRNA A. Sustained depletion resulted in greater fold changes of gene expression.

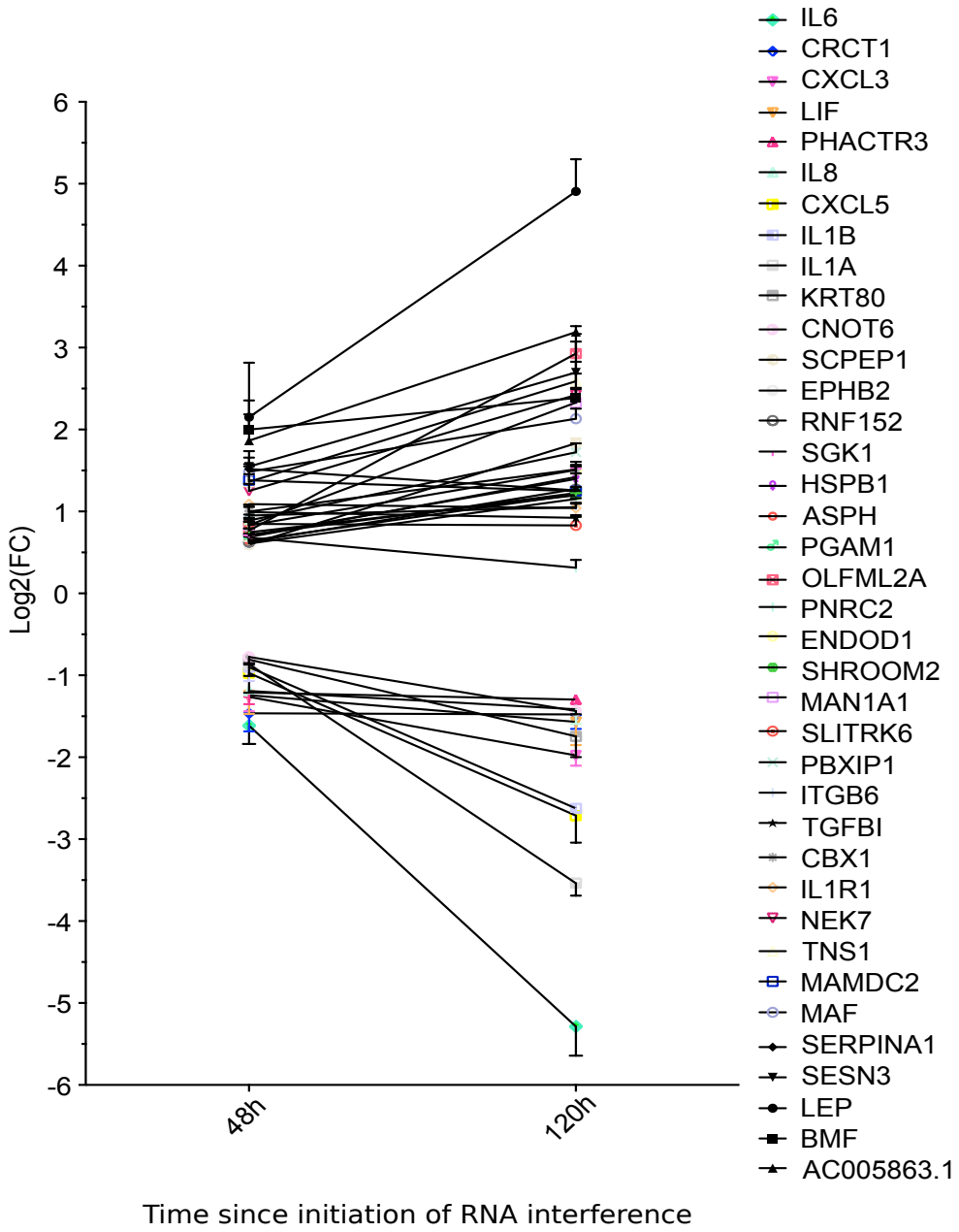


Figure 5.7: Log_2 fold changes of 39 genes significantly deregulated at 48h and 120h using D-siRNA A. Genes have greater effect sizes in 120h response for all except one.

Genes showing altered expression include a number of genes known to play important roles in lung disease such as α1-antitrypsin (SERPINA1) (Laurell and Eriksson 2013), transforming growth factor β 1 (*TGFβI*) (Makinde et al. 2007), interleukin 1 receptor 1 (*IL1R1*) (Frank et al. 2008) and *IL6*, *IL8*, *IL1B*, *IL1A* (Grutters et al. 2003, Heinzmann et al. 2004, Xie et al. 2009, Falfan-Valencia et al. 2012). *IL6* had the greatest reduction in expression. The gene with the greatest fold induction was Leptin (*LEP*) which was shown to be upregulated and secreted from HBECs infected

with respiratory syncytial virus (Qin et al. 2015) (Table 5.3). Interestingly, several polymorphisms in linkage with *LEP* are associated with lung function (van den Borst et al. 2011). LEP blood concentration was also shown to negatively correlate with lung function (Eising et al. 2013). Crucially, LEP upregulation has been validated in an additional donor HBECs depleted of INTS12 (Figure 5.8). These findings give support to the hypothesis that the altered expression of INTS12 in population studies is driving the genetic association signal for lung function.

48h and 120h consensus genes	MEAN FOLD CHANGES ± SEM			
	48h		120h	
	Scrambled vs D-siRNA A	Scrambled vs D-siRNA C	Scrambled vs D-siRNA A	Scrambled vs D-siRNA C
LEP	4.92 ± 2.12	19.14 ± 11.33	36.72 ± 14.81	32.06 ± 14.13
AC005863.1	3.83 ± 0.88	4.70 ± 1.29	9.15 ± 0.45	5.36 ± 0.81
OLFML2A	1.71 ± 0.13	2.54 ± 0.55	7.69 ± 0.81	2.63 ± 0.39
SESN3	2.94 ± 0.23	1.95 ± 0.07	6.54 ± 0.62	2.49 ± 0.47
TNS1	2.61 ± 0.33	5.27 ± 1.71	6.05 ± 0.40	5.97 ± 0.60
NEK7	2.39 ± 0.15	2.23 ± 0.23	5.38 ± 0.33	3.78 ± 0.18
MAN1A1	1.79 ± 0.09	2.01 ± 0.17	5.09 ± 0.59	2.58 ± 0.15
MAF	2.81 ± 0.17	4.87 ± 0.69	4.41 ± 0.36	5.37 ± 0.54
BMF	4.26 ± 1.12	3.76 ± 0.86	7.06 ± 3.86	7.93 ± 5.21
SCPEP1	1.51 ± 0.04	1.51 ± 0.11	3.56 ± 0.15	1.21 ± 0.06
PBXIP1	1.86 ± 0.09	2.07 ± 0.14	3.32 ± 0.24	2.22 ± 0.04
CBX1	2.00 ± 0.08	2.20 ± 0.08	2.86 ± 0.09	3.24 ± 0.15
ENDOD1	1.78 ± 0.05	1.77 ± 0.10	2.85 ± 0.20	2.93 ± 0.26
SGK1	1.61 ± 0.02	1.53 ± 0.08	2.69 ± 0.21	1.85 ± 0.12
HSPB1	1.63 ± 0.10	1.43 ± 0.11	2.67 ± 0.31	1.43 ± 0.02
RNF152	1.55 ± 0.07	1.81 ± 0.15	2.42 ± 0.05	1.89 ± 0.02
SERPINA1	2.93 ± 0.47	2.89 ± 0.56	2.44 ± 0.34	2.77 ± 0.36
PGAM1	1.65 ± 0.09	1.69 ± 0.16	2.34 ± 0.01	1.94 ± 0.05
ASPH	1.56 ± 0.07	1.61 ± 0.13	2.31 ± 0.02	2.26 ± 0.05
MAMDC2	2.61 ± 0.13	3.42 ± 0.59	2.48 ± 0.48	5.12 ± 0.59
SHROOM2	1.69 ± 0.15	1.95 ± 0.18	2.29 ± 0.04	1.72 ± 0.03
EPHB2	1.54 ± 0.13	2.07 ± 0.07	2.22 ± 0.01	2.54 ± 0.13
ITGB6	1.95 ± 0.17	2.74 ± 0.53	2.08 ± 0.06	3.86 ± 0.31
IL1R1	2.13 ± 0.06	1.95 ± 0.03	2.06 ± 0.09	1.60 ± 0.17
TGFBI	2.00 ± 0.13	2.75 ± 0.20	1.89 ± 0.04	5.62 ± 0.66
SLITRK6	1.80 ± 0.10	2.45 ± 0.55	1.79 ± 0.13	2.57 ± 0.37
PNRC2	1.60 ± 0.04	1.51 ± 0.05	1.25 ± 0.08	1.31 ± 0.12
PHACTR3	0.44 ± 0.04	0.58 ± 0.15	0.41 ± 0.01	0.51 ± 0.09
IL8	0.44 ± 0.01	0.50 ± 0.11	0.38 ± 0.04	0.18 ± 0.01
CRCT1	0.37 ± 0.06	0.58 ± 0.10	0.36 ± 0.04	0.56 ± 0.05
CNOT6	0.59 ± 0.03	0.62 ± 0.01	0.37 ± 0.01	0.53 ± 0.01
LIF	0.43 ± 0.06	0.41 ± 0.05	0.35 ± 0.07	0.33 ± 0.03
KRT80	0.57 ± 0.02	0.36 ± 0.03	0.31 ± 0.05	0.32 ± 0.05
CXCL3	0.42 ± 0.05	0.34 ± 0.06	0.26 ± 0.02	0.21 ± 0.03
IL1B	0.54 ± 0.07	0.38 ± 0.04	0.16 ± 0.00	0.31 ± 0.00
CXCL5	0.52 ± 0.09	0.45 ± 0.08	0.16 ± 0.03	0.33 ± 0.04
IL1A	0.56 ± 0.06	0.65 ± 0.08	0.09 ± 0.01	0.47 ± 0.02
IL6	0.33 ± 0.05	0.48 ± 0.06	0.03 ± 0.01	0.19 ± 0.05

Table 5.3: Deregulation of a core subset of genes due to INTS12 knockdown. The table is showing the mean fold changes of consensus significantly differentially expressed genes after 48h and 120h since the D-siRNA A and C transfections.

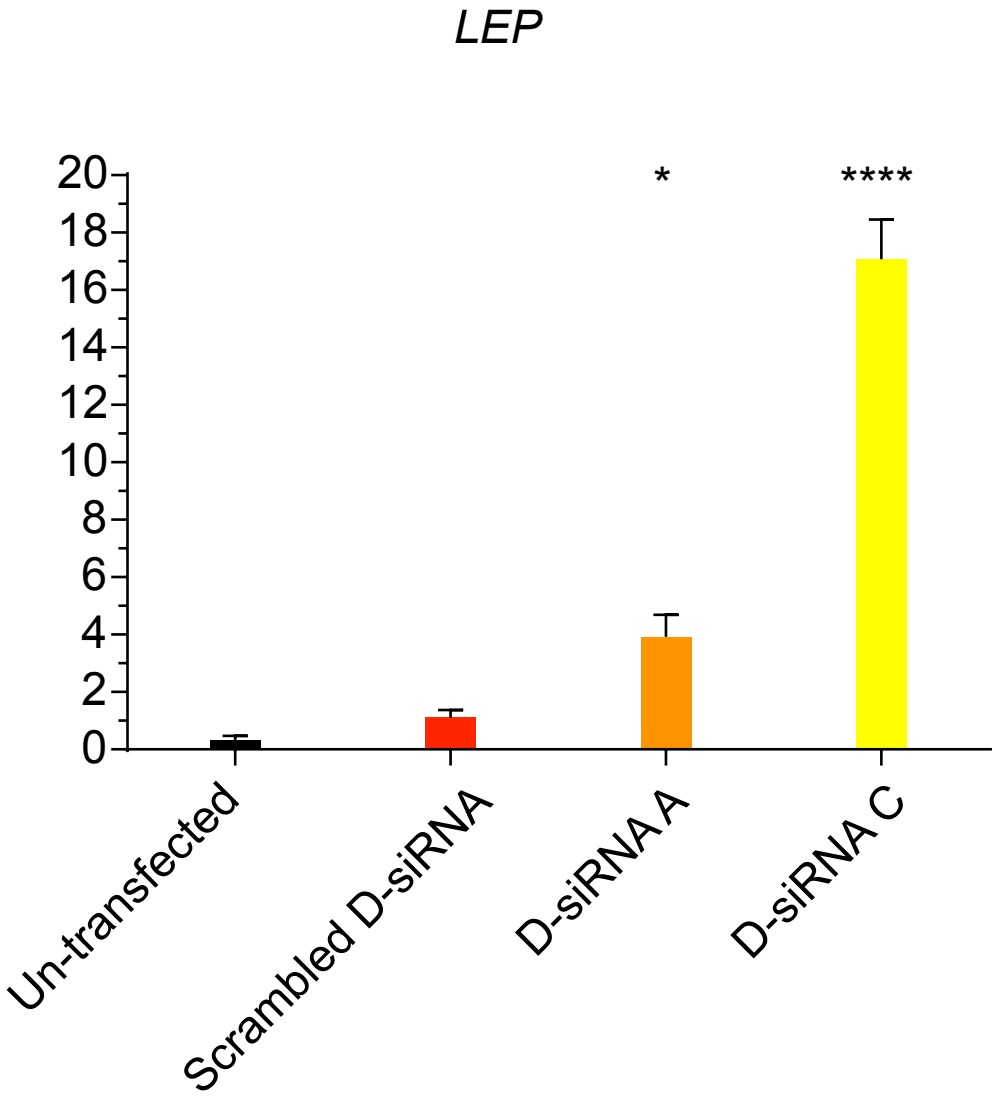


Figure 5.8: qPCR expression profiling of LEP expression in additional donor cells. LEP is significantly upregulated in validation donor HBECs depleted of INTS12. Statistical tests were performed comparing to scrambled D-siRNA control: *P<0.05, ****P<0.0001. Individual ΔΔCt gene expressions are GAPDH normalized and relative to the mean of the scrambled D-siRNA condition. No significant difference was observed between un-transfected and scrambled D-siRNA transfected cells.

5.3.3 Differential pathway analysis identifies dysregulation of protein synthesis and collagen formation pathways following INTS12 knockdown

Transcriptomic profiling of cells sustainably depleted of INTS12 (i.e. 120h dataset) revealed a set of 1660 genes which were altered in expression following treatment with both D-siRNA A and C (Figure 5.4). 279 genes deregulated in all three comparisons but in opposite direction in INTS12 knockdown conditions when compared to un-transfected vs. scrambled D-siRNA analysis were also shortlisted yielding a total of 1939 genes (Figure 5.5).

To interpret this large number of gene expression changes, GSEA approach was used as described in section 5.2.2. Using this method three pathways were upregulated and eight pathways were downregulated (Figures 5.9, 5.10). Collagen formation and extracellular matrix organization pathways were the top two upregulated pathways (defined by normalized enrichment scores) in D-siRNA A and D-siRNA C analyses (Figures 5.9, 5.10, 3.12). The top two downregulated pathways were cytosolic tRNA aminoacylation and PERK regulated gene expression, which is a sub-pathway of the unfolded protein response (Figure 5.9, 5.10, 5.11). Significant downregulation of other protein metabolism related pathways, including unfolded protein response, activation of genes by transcription factor 4 (*ATF4*) (Figure 5.10) and glycine, serine and threonine metabolism pathways (Figure 5.10) was observed. *ATF4* expression was reduced by 70±5% and 45±2% in D-siRNA A and C transfected cells when compared to scrambled D-siRNA transfected cells respectively (FDR<0.05). As it is has been reported that *ATF4* knockout cells have impaired expression of genes involved in resistance to oxidative stress and amino acid transport (Harding et al. 2003), this suggests an impact on integrated stress response (Marciniak et al. 2006) caused by INTS12 knockdown. Importantly, dysregulation of the above pathways was not observed when comparing un-transfected cells with scrambled D-siRNA cells.

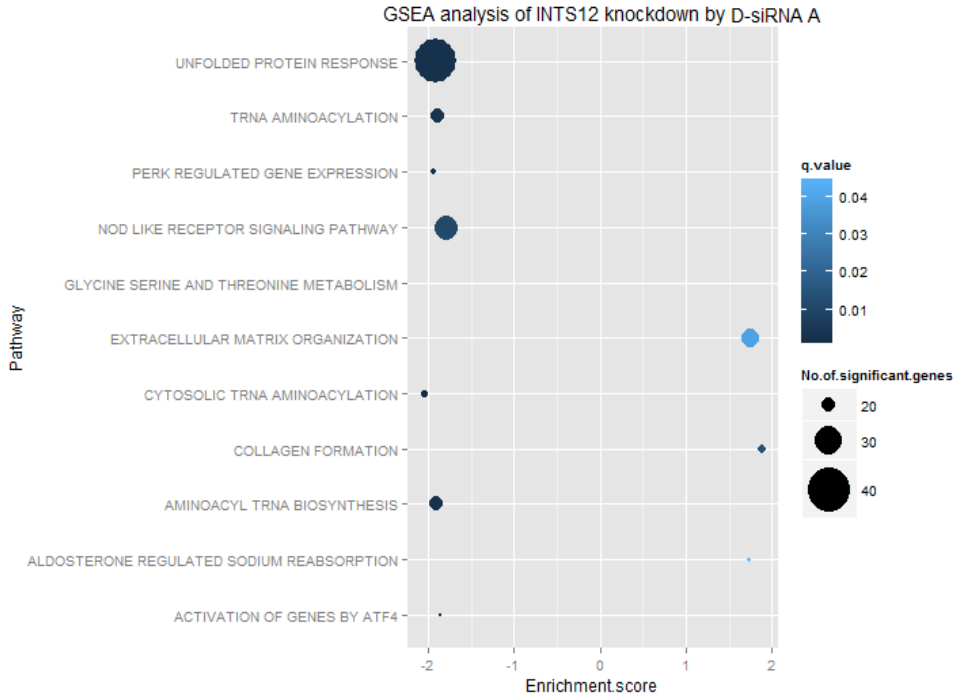


Figure 5.9: Cleveland's plot showing the GSEA results of D-siRNA A analysis. The shade of colour indicates statistical significance of enrichment after multiple testing correction. The size of dot reflects the number of statistically significant differentially expressed gene. The location of the dot on the X-axis reflects the enrichment score.

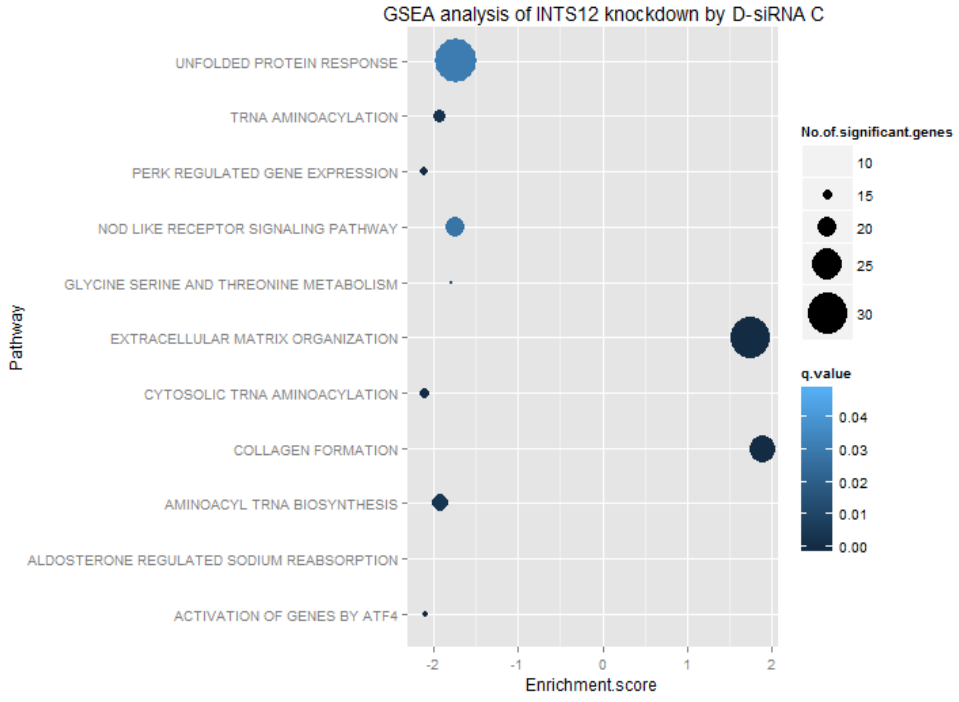
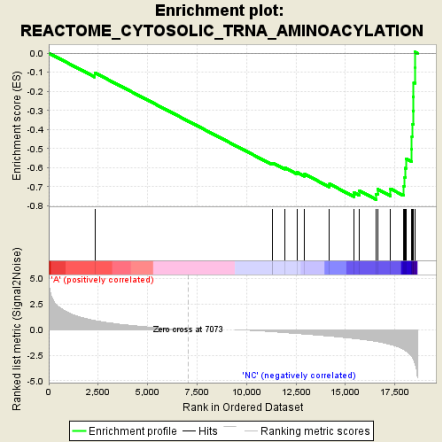
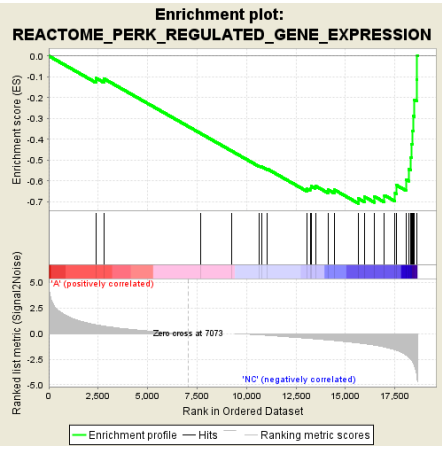


Figure 5.10: Cleveland's plot showing the GSEA results of D-siRNA C analysis.

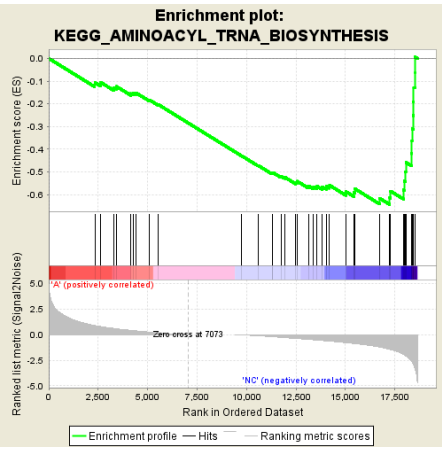
D-siRNA A analysis



FDR ≈ 0.0004; NSE ≈ -2.05

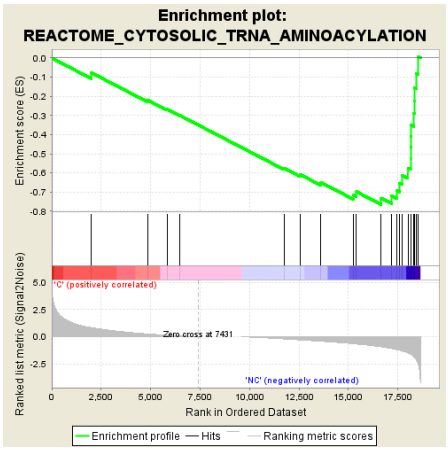


FDR ≈ 0.002; NSE≈ -1.95

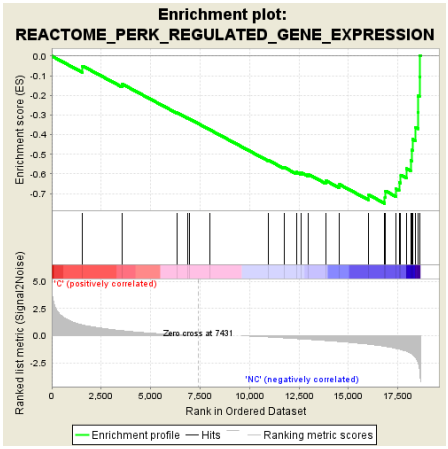


FDR ≈ 0.003; NSE ≈ -1.90

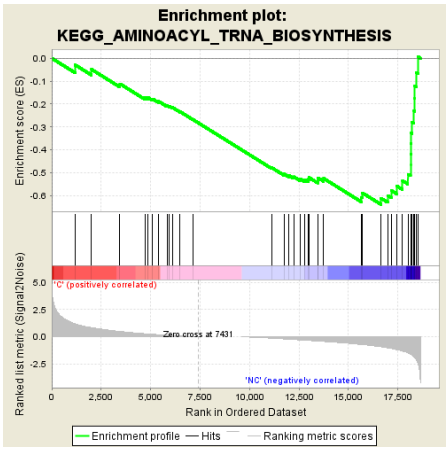
D-siRNA C analysis



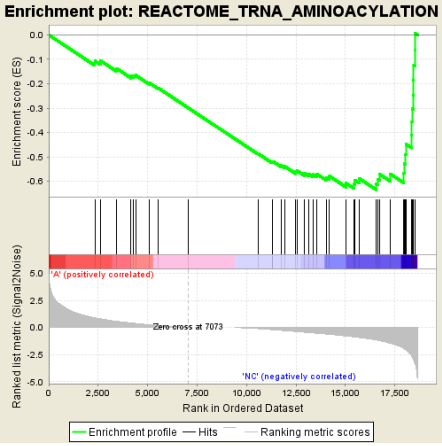
FDR ≈ 0.00009; NSE≈ -2.10



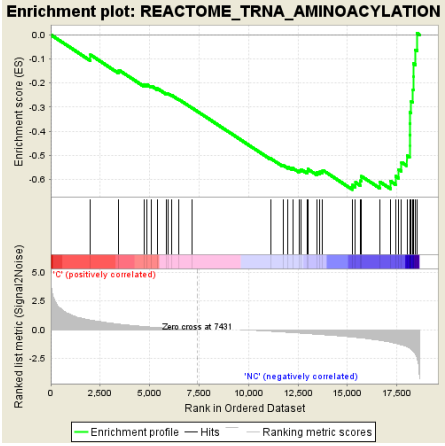
FDR ≈ 0.00006; NSE≈ -2.12



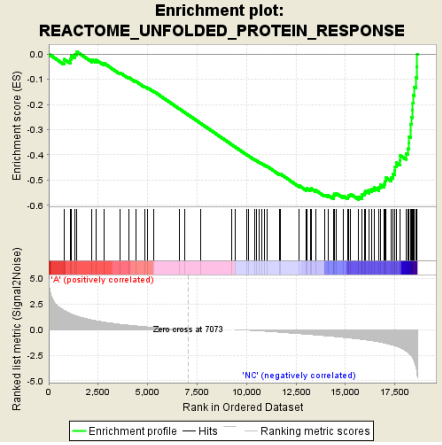
FDR ≈ 0.004; NSE≈ -1.92



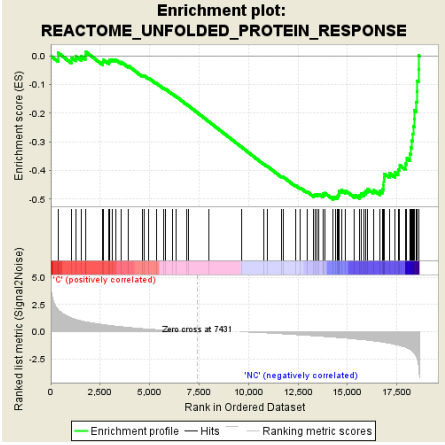
FDR ≈ 0.004; NSE≈ -1.89



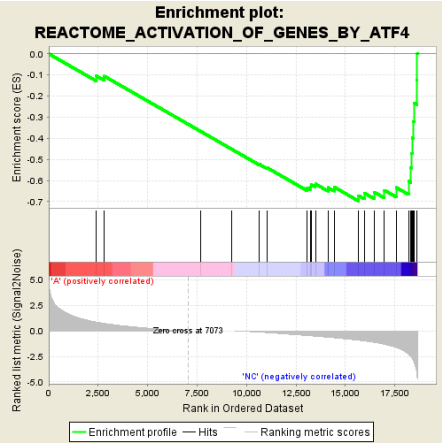
FDR ≈ 0.003; NSE ≈ -1.94



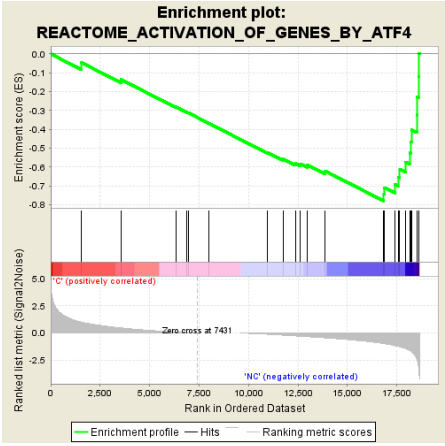
FDR ≈ 0.003; NSE ≈ -1.91



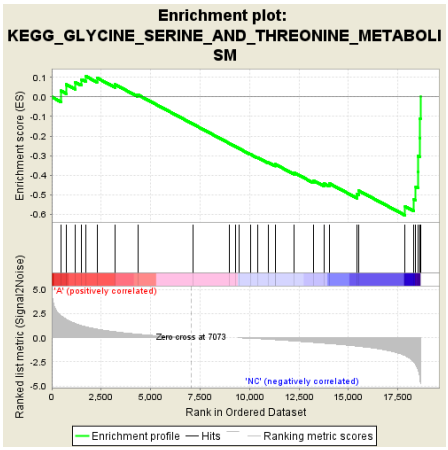
FDR ≈ 0.03; NSE ≈ -1.72



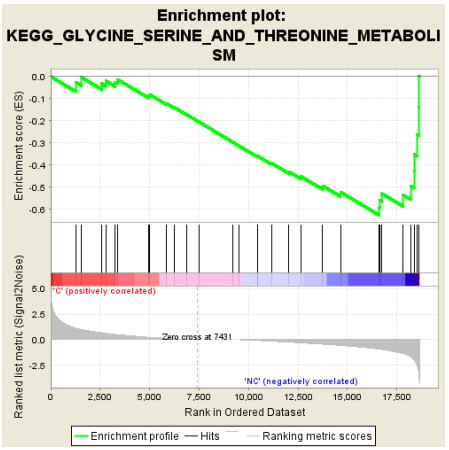
FDR ≈ 0.006; NSE ≈ -1.86



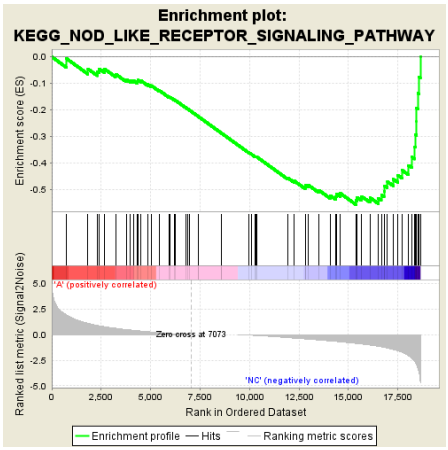
FDR ≈ 0.0001; NSE ≈ -2.10



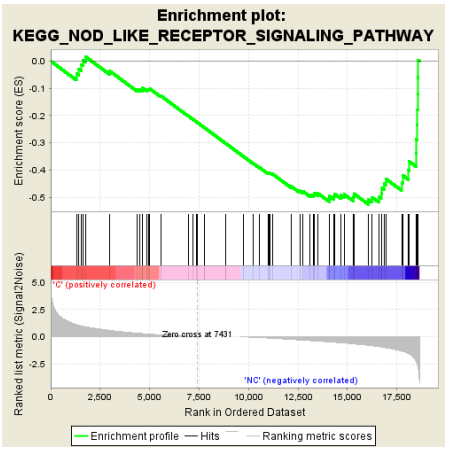
FDR ≈ 0.03; NSE ≈ -1.72



FDR ≈ 0.02; NSE ≈ -1.79



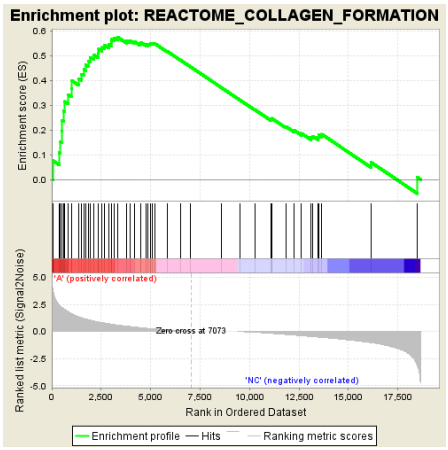
FDR ≈ 0.01; NSE ≈ -1.80



FDR ≈ 0.03; NSE ≈ -1.74

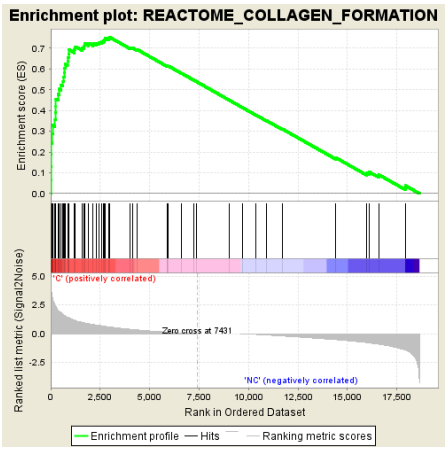
Figure 5.11: Enrichment plots of reproducibly downregulated pathways (FDR<0.05) in D-siRNA A and C analyses showing the distribution of enrichment scores per gene set throughout the signal-to-noise ranked gene list with indicated statistical significance and normalized enrichment score of their respective downregulations. The FDR and normalized enrichment score (NSE) values were rounded up to one and three significant figure respectively.

D-siRNA A analysis

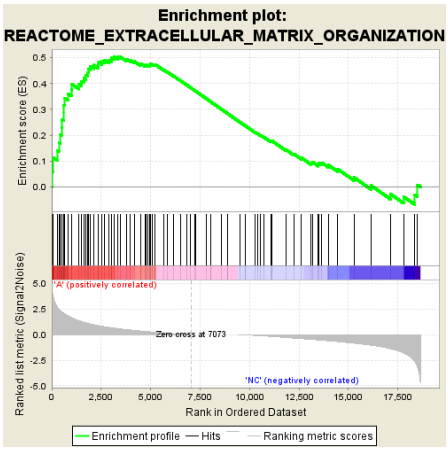


FDR ≈ 0.01; NSE ≈ 1.87

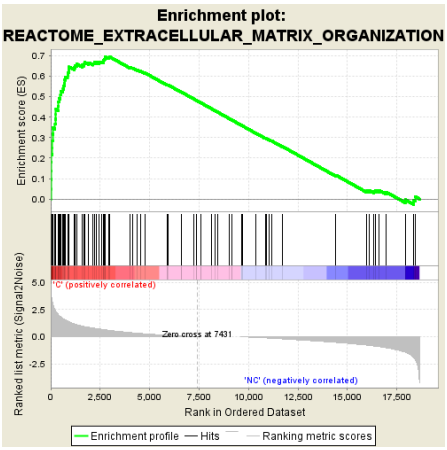
D-siRNA C analysis



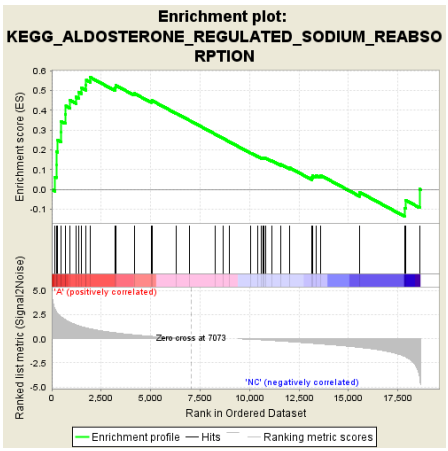
FDR < 0.00001; NSE ≈ 2.42



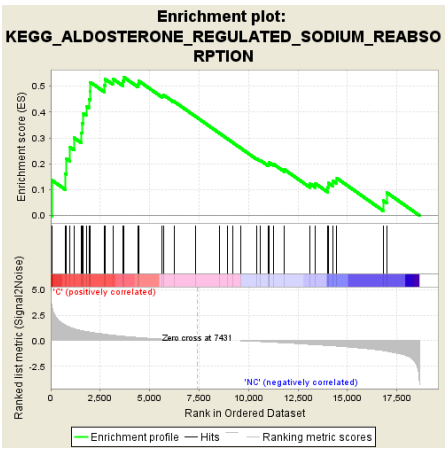
FDR ≈ 0.03; NSE ≈ 1.78



FDR < 0.00001; NSE ≈ 2.40



FDR ≈ 0.04; NSE ≈ 1.74



FDR ≈ 0.05; NSE ≈ 1.64

Figure 5.12: Enrichment plots of reproducibly upregulated pathways (FDR<0.05) in D-siRNA A and C analyses.

5.3.3.1 Consistency and robustness of identified pathways dysregulation

The main rationale for the presented analysis was to identify pathways genuinely dysregulated due to INTS12 depletion. Thus, as mentioned above, pathways with evidence of perturbation in two independent INTS12 targeting D-siRNAs and not dysregulated in scrambled D-siRNA are of primary interest. Equally importantly, it is crucial to test whether the observed effects are consistent in the different treatments (i.e. whether genes appear to have similar molecular signature in the D-siRNA A and D-siRNA C) and how robust are the downregulation and upregulation responses.

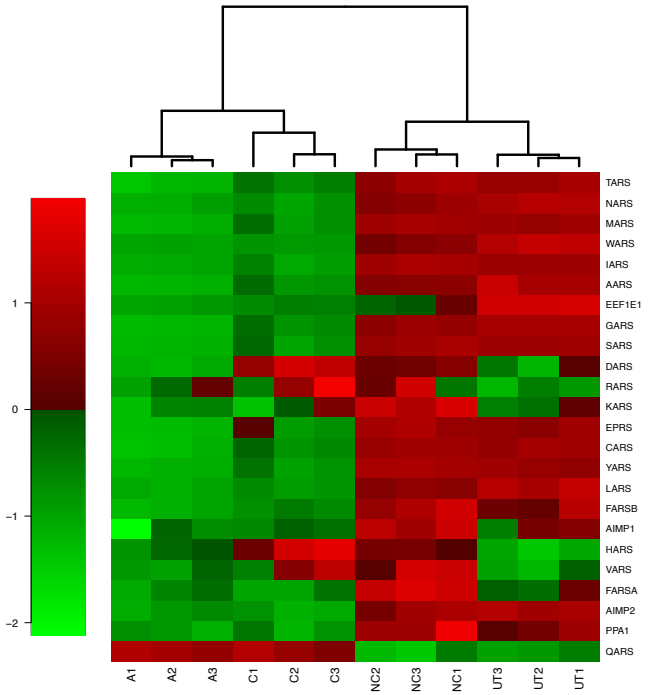
As far as the latter point is concerned, it is possible to say that among the identified dysregulation of 11 pathways, downregulated pathways tended to show more robust effects than upregulated pathways due to lesser inter-experimental variability (Figure 5.13 and 5.14), greater magnitude of effect (Figure 5.15) and larger number of significantly dysregulated pathways (Figures 5.9 and 5.10). Overall, the molecular signatures of upregulated pathways appear more variable and less reproducible than downregulated pathways. Moreover, the genes belonging to the downregulated pathways appear clearly more localized at the bottom of the signal-to-noise ranked list (see section 2.8.3) while genes belonging to the upregulated pathways appear more distributed throughout the ranked list (Figure 5.11 versus Figure 5.12).

The leading edge in the GSEA analysis is a core subset of pathway genes present before the enrichment plot curve starts descending for the upregulated pathways (Figure 5.12) or after enrichment plot curve starts ascending for downregulated pathways (Figure 5.11). Thus these genes are considered to be mostly contributing to the enrichment score and hence are predominantly responsible for pathway dysregulation (Subramanian et al. 2005). By comparing the leading edges in the D-siRNA A and D-siRNA C it is also possible to determine the consistency and robustness of a particular pathway perturbation.

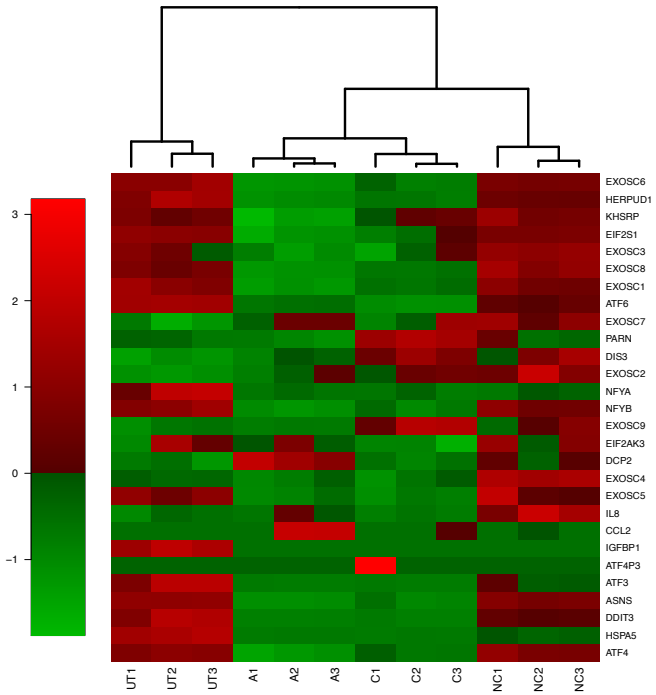
Through examination of these leading edges, it appeared that for the downregulated pathways with the exception of NOD-like receptor signalling, the genes mostly contributing to pathway downregulation in

D-siRNA A were very well represented in the D-siRNA C (Table 5.4). On the other hand, genes contributing to the enrichment of upregulated pathways in D-siRNA A analysis were notably less represented in the leading edge of D-siRNA C (Table 5.4). For example, in the cytosolic tRNA aminoacetylation pathway 93% of genes contributing to pathway downregulation in D-siRNA A also contributed to pathway downregulation in D-siRNA C and vice versa. As far as collagen formation is concerned 81% of genes that contributed to pathway upregulation in D-siRNA A also contributed to pathway upregulation in D-siRNA C, while only 57% of genes that contributed to pathway upregulation in D-siRNA C were represented in the leading edge of D-siRNA A.

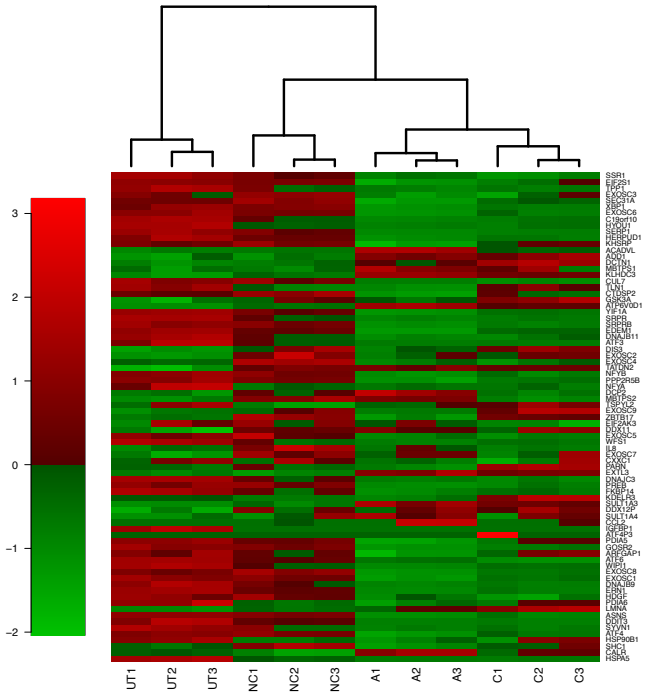
Cytosolic tRNA aminoacylation (REACTOME)



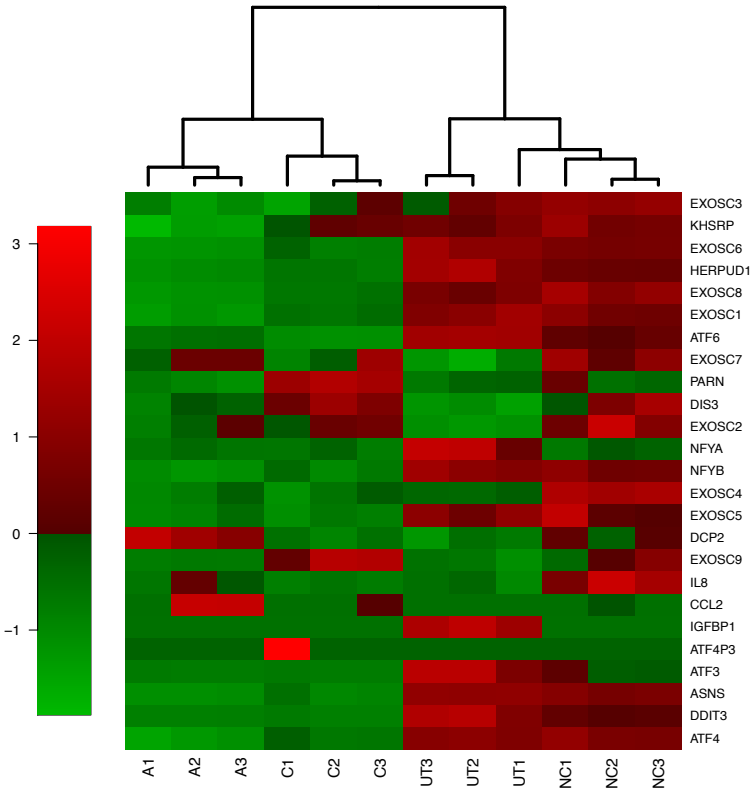
PERK regulated gene expression (REACTOME)



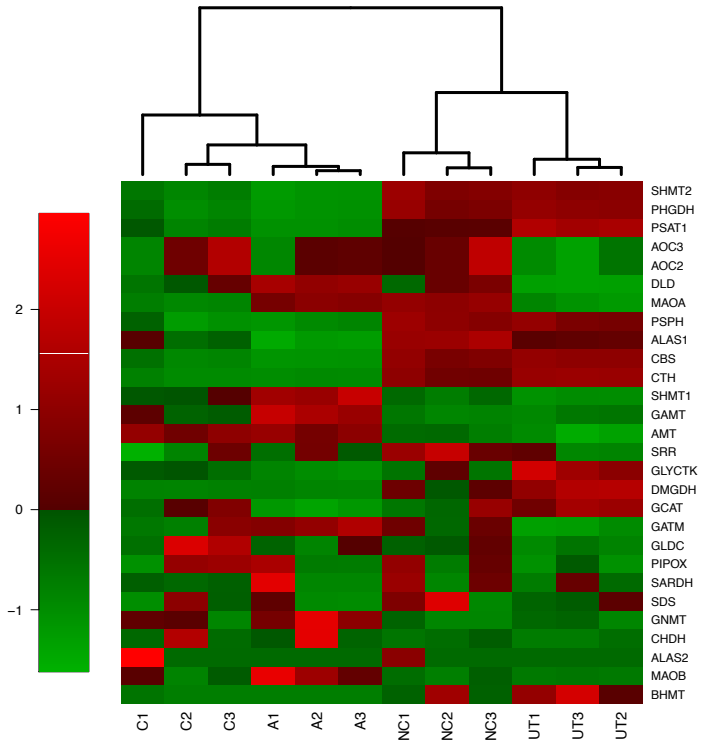
Unfolded protein response (REACTOME)



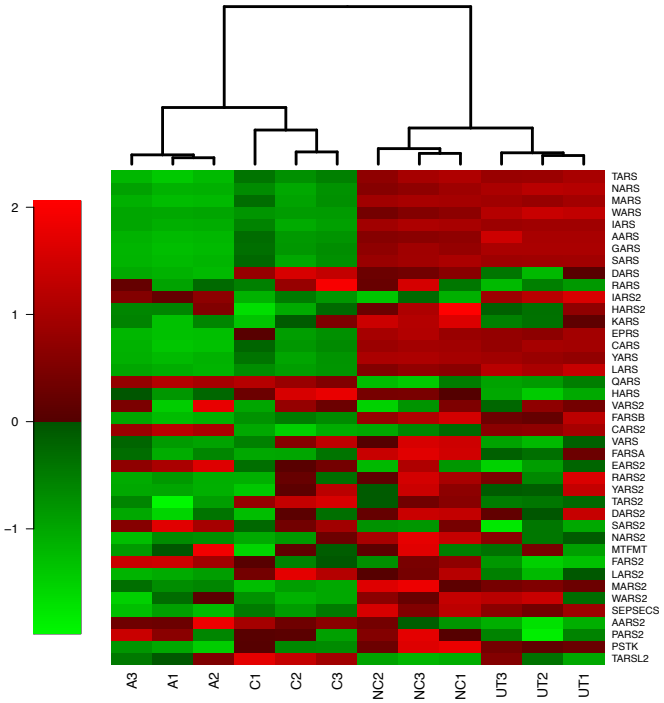
Activation of genes by ATF4 (REACTOME)



Glycine, serine and threonine metabolism (KEGG)



Aminoacyl tRNA biosynthesis (KEGG)



NOD like receptor signalling (KEGG)

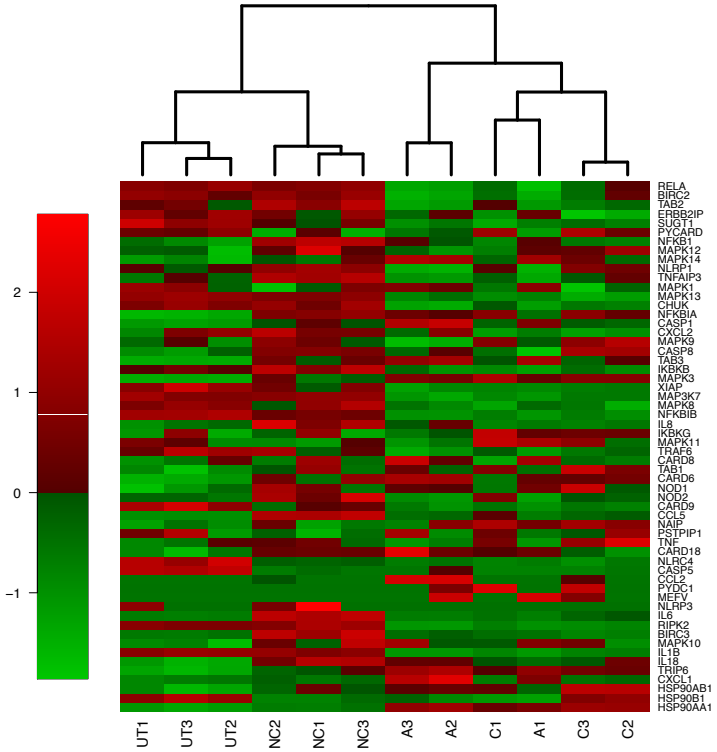
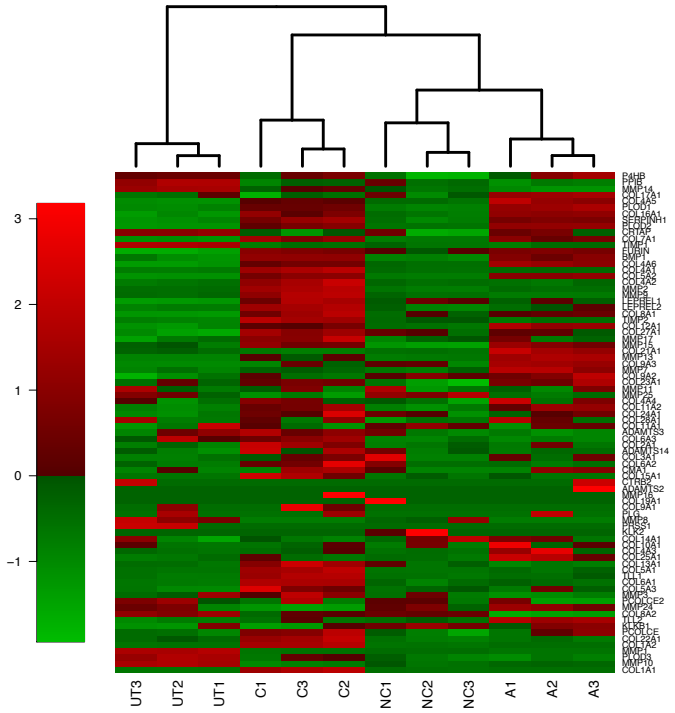
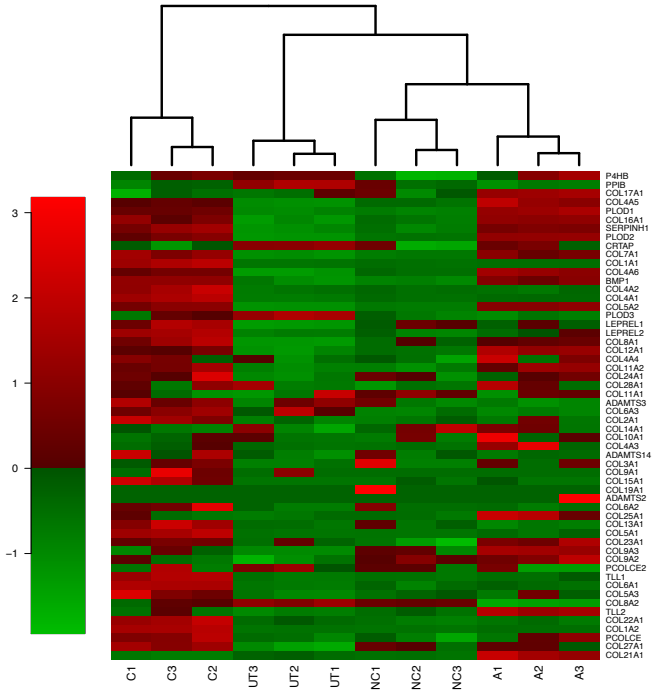


Figure 5.13: Heatmaps of genes belonging to downregulated pathways. Samples were clustered by unsupervised hierarchical clustering and resulted in clustering of three biological replicate samples of each of the four conditions: un-transfected cells (UT), cells transfected with scrambled D-siRNA control (NC), cells transfected with anti-INTS12 D-siRNA A (A) and cells transfected with anti-INTS12 D-siRNA C (C). Green and red colours on the Z-scale indicate lower and higher expression respectively.

Extracellular matrix organization (REACTOME)



Collagen formation (REACTOME)



Aldosterone regulated sodium re-absorption (KEGG)

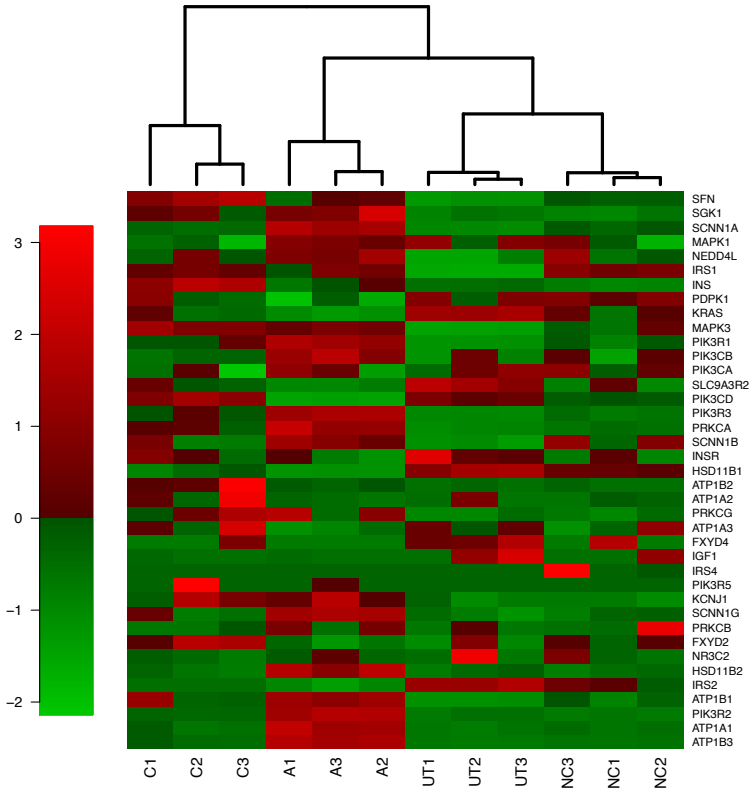


Figure 5.14: Heatmaps of genes belonging to upregulated pathways. Samples were clustered by unsupervised hierarchical clustering where green and red colours on the Z-scale indicate lower and higher expression respectively.

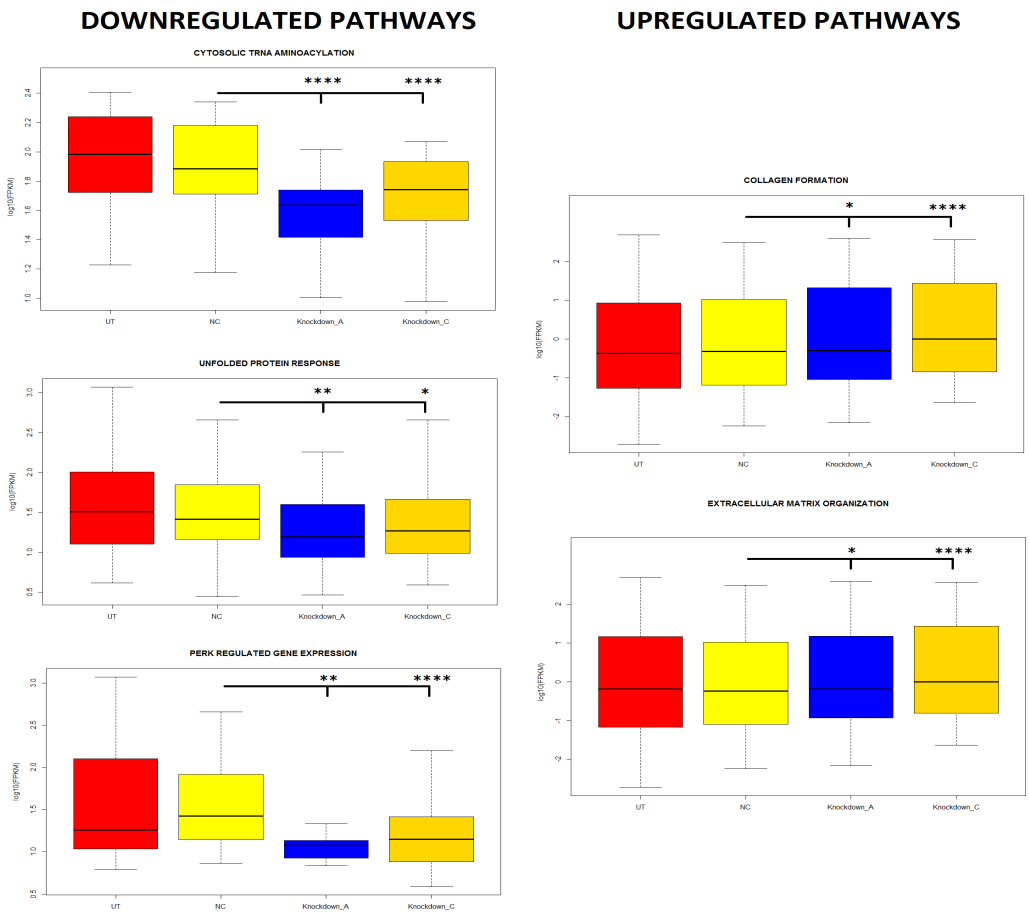


Figure 5.15: Box plots representing log_{10} of fragment per kilobase per million reads (FPKM) expression values of genes belonging to the dysregulated pathways due to INTS12 knockdown. The downregulation effects are more robust than upregulation effects, due to lesser inter-experimental variability, greater magnitude of effect and larger number of significantly dysregulated pathways. Cytosolic tRNA aminoacetylation and PERK regulated gene expression, both important in protein metabolism and integrated stress response were suppressed in both D-siRNA treatments which was not observed in scrambled D-siRNA treated cells. Stars indicate the significance of pathway dysregulation in GSEA.

Pathway	Percent of D-siRNA A leading genes represented in D-siRNA C edge	Percent of D-siRNA C leading genes represented in D-siRNA A edge
UPREGULATED PATHWAYS		
Collagen formation (REACTOME)	81%	57%
Extracellular matrix organization (REACTOME)	86%	54%
Aldosteron regulated sodium reabsorption (KEGG)	73%	65%
DOWNREGULATED PATHWAYS		
Activation of genes by ATF4 (REACTOME)	92%	92%
Cytosolic tRNA aminoacylation (REACTOME)	92%	92%
Aminoacyl tRNA biosynthesis (KEGG)	100%	94%
NOD-like receptor signalling pathway (KEGG)	71%	83%

Table 5.4: Representation of D-siRNA A leading edge genes in the leading edge of D-siRNA C and vice versa. Shown results for representative pathways.

5.3.3.2 Technical and biological validation of identified pathway dysregulation

In order to validate the RNAseq data, four genes from the top two downregulated pathways (cytosolic tRNA aminoacetylation and PERK regulated gene expression; see section 5.3.2) were selected: methionyl-tRNA synthetase (*MARS*) and glycyl-tRNA synthetase (*GARS*) genes from the tRNA synthetases pathway and *ATF4* and asparagine synthetase (*ASNS*) genes from the PERK pathway. qPCR technical validation of RNAseq findings was performed using three biological cDNA replicates derived from total RNA used in sequencing thus were upon the same RNA samples. Differences in gene expression between each experimental condition derived from RNAseq data were compared to the differences obtained from the qPCR data. Analysis revealed Pearson correlation of $\log_2$ of differences in gene expression derived from RNAseq and qPCR estimates to be 0.99 ($P<0.0001$; Figure 5.16).

In addition, biological validation was tested by repeating the INTS12 knockdown experiment using different donor cells, in at least three biological replicates. This set of experiments replicated the downregulation of these genes at a statistically significant level (except for *MARS* which appears downregulated when comparing its expression in D-siRNA A relative to scrambled control but did not reach statistical significance) albeit with a different effect size ($P<0.05$; Figure 5.18). The differences in the magnitude of changes observed can be attributed to the different efficiencies of INTS12 knockdown in the discovery and validation donors: INTS12 was suppressed by 72% and 86% in the validation donor for D-siRNA A and C respectively versus 93% and 85% in the discovery donor for D-siRNA A and C respectively (Figure 5.17). In it notable that in validation donor not only all the target genes appeared downregulated but also the magnitude of change corresponded to the level of INTS12 knockdown, i.e. there is a greater magnitude of change observed in D-siRNA C where INTS12 silencing was greater than in D-siRNA A. Interestingly, in the validation donor experiment, INTS12 expression was significantly upregulated in scrambled D-siRNA treatment relative to un-transfected control suggesting donor specific responses of INTS12 to non-specific knockdown. This effect was mirrored by MARS and GARS in a significant manner, implying that these genes are upregulated as a result of treatment to scrambled D-siRNA just as it was observed for INTS12. However, it is possible to say that the data for validation donor supports the observations seen in discovery donor as target genes are downregulated in INTS12 knockdown conditions relative to scrambled D-siRNA, making their upregulation in scrambled D-siRNA irrelevant for validation purposes.

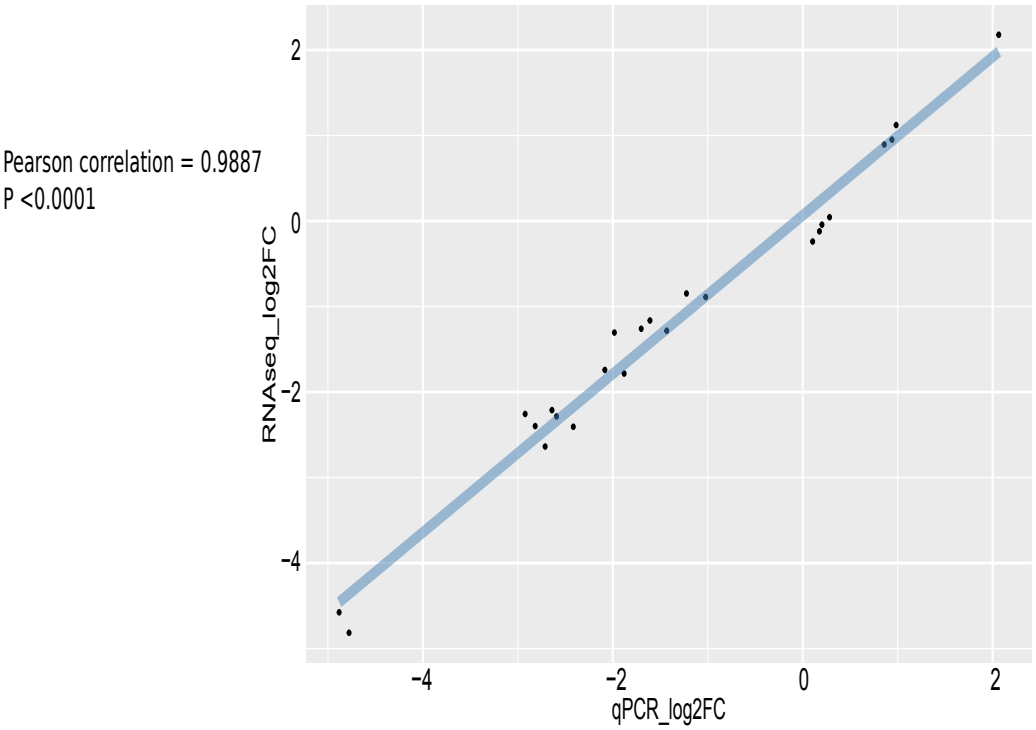


Figure 5.16: Technical validation of RNAseq findings by qPCR. Differences in gene expression derived from RNAseq strongly and significantly correlate with differences in gene expression derived from qPCR. Validation assays were performed on the same samples that were used for RNAseq study.

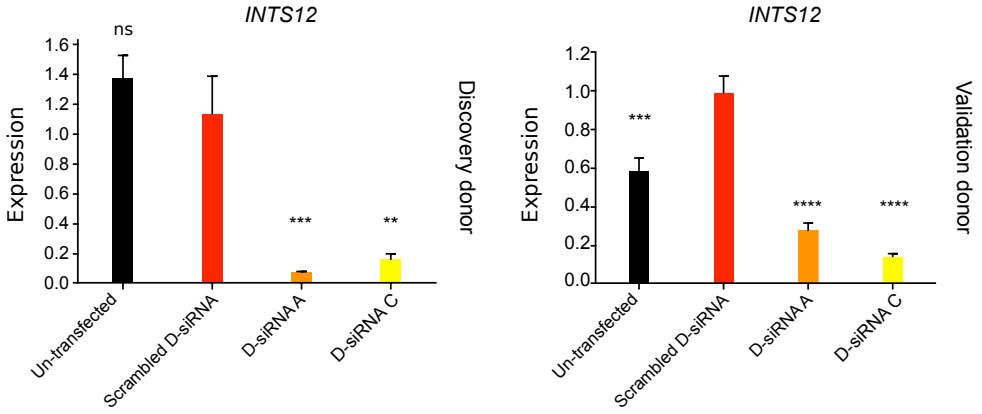


Figure 5.17: INTS12 expression in the discovery and validation donors. In the discovery donor INTS12 knockdown was greater in D-siRNA A than in D-siRNA C. On the other hand, in the validation donor INTS12 knockdown was greater in D-siRNA C than in D-siRNA A. Statistical tests were performed comparing to scrambled D-siRNA control: *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001. Individual ΔΔCt gene expressions are GAPDH normalized and relative to the mean of the scrambled D-siRNA condition.

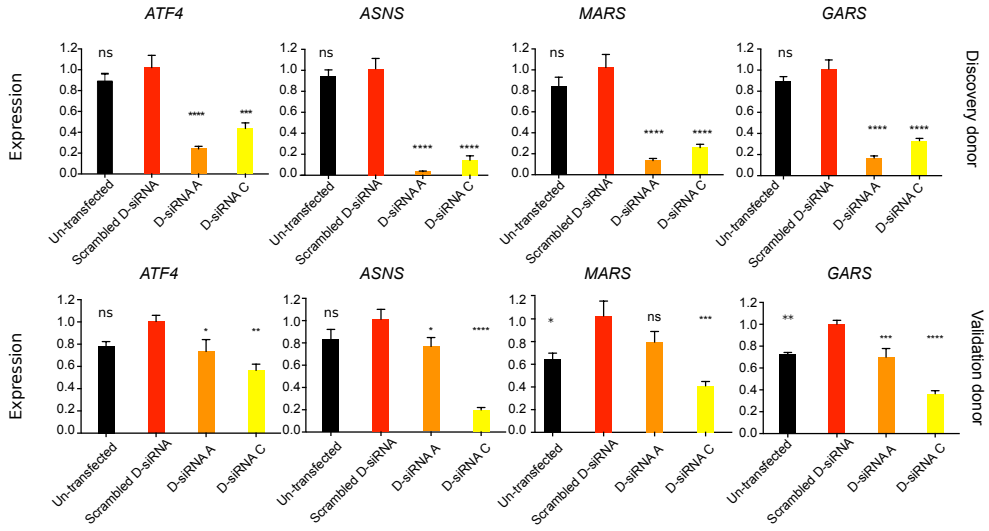


Figure 5.18: Biological validation of downregulation of genes belonging to cytosolic tRNA aminoacylation and PERK pathways in additional donor cells. ASNS and ATF4 expressions, representing PERK pathway, were qPCR assayed and had significantly reduced expression in discovery donor cells, as well as validation donor cells, with the exception of MARS in D-siRNA A condition. Statistical tests were performed comparing to scrambled D-siRNA control: *P<0.05, **P<0.01, ***P<0.001, ****P<0.0001. Individual ΔΔCt gene expressions are GAPDH normalized and relative to the mean of the scrambled D-siRNA condition.

5.3.4 *INTS12* regulates translation and proliferation

The uncovered molecular signature following INTS12 depletion allowed for novel hypothesis generation according to the scheme outlined in Chapter 1 The key question was whether the discovered disruption in the dysregulated pathway has an effect on the relevant cell phenotype? Because INTS12 knockdown induced downregulation of several key pathways involved in protein metabolism and translational control, including the tRNA synthetases aminoacylation, PERK regulated gene expression, unfolded protein synthesis, amino acids metabolism and ATF4 activated gene expression; it was hypothesised that this manipulation would affect protein synthesis.

To test this hypothesis a radiolabelled amino acid incorporation assay was used (Wong et al. 2010). As predicted, sustained 120h INTS12 silencing repressed protein synthesis by 23±3% and 47±3% in D-siRNA A and C respectively (P<0.05; Figure 5.19). Since cell division requires doubling of protein content prior to separation, it was also hypothesised

that INTS12 depletion would affect the cells capacity to proliferate. Cell counts revealed 25±13% and 48±4% decrease of proliferation in cells treated with D-siRNAs A and C respectively ($P<0.05$; Figure 5.20). It is possible to say that the observed reduction in protein synthesis can be causally attributed to INTS12 knockdown because (1) it is observed following specific manipulation of gene level (2) it is seen in two independent D-siRNAs targeting this genes, (3) it is not seen in scrambled D-siRNA control. However, it still remains to be elucidated whether this causal relationship between INTS12 and protein synthesis is direct or indirect, and Chapter 6 will shed light on this issue. Although it could be the case that INTS12 reduction activates apoptotic or cytotoxic processes which then effect protein synthesis, these phenotypic outcomes have not been measured. It remains speculative whether they are contributory and this still does not invalidate the theory emphasizing causal contribution of INTS12 to protein production. From a survey of RNAseq data generated in this thesis it appears that caspase 3 and caspase 7, the key effector molecules in the apoptotic pathway, are not dysregulated due to INTS12 depletion, making the role of apoptotic pathway very questionable. Thus, these data identify a novel and significant unrecognised role for INTS12 in regulation of cellular protein synthesis and proliferation.

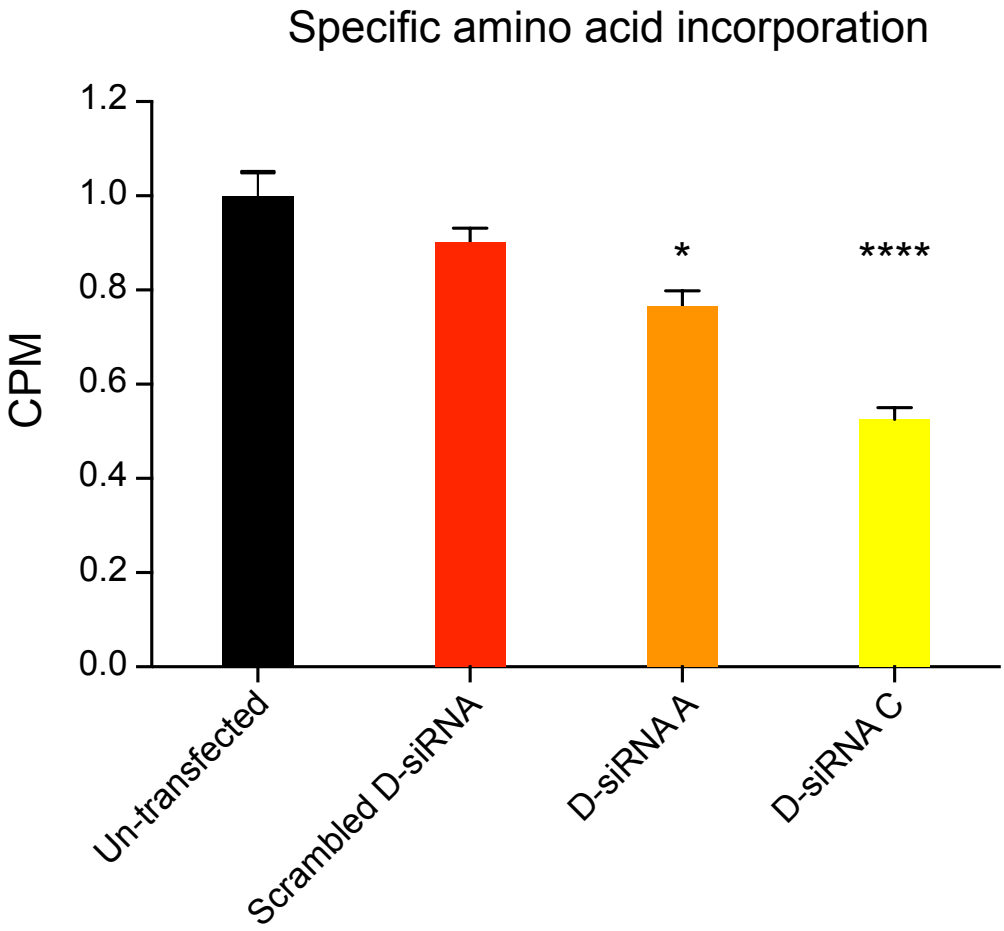


Figure 5.19: Amino acid incorporation measured by counts per methionine (CPM) in 120h since the start of RNAi radiolabelling experiment. Statistical tests were performed comparing to scrambled D-siRNA control: *P<0.05, ***P<0.001. Individual CPM values are normalized to the amount of total protein and are shown as relative to the mean of the un-transfected condition. No significant difference was observed between un-transfected and scrambled D-siRNA transfected cells.

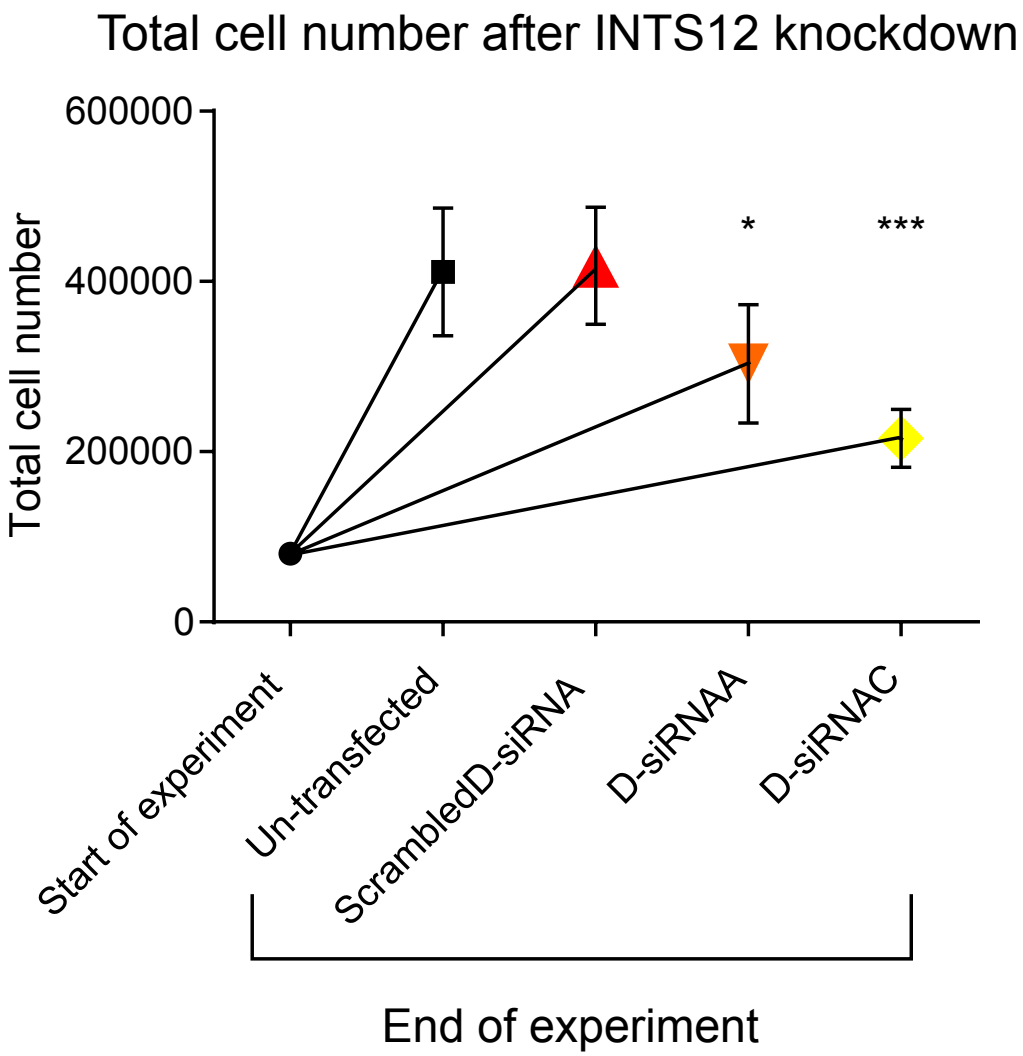


Figure 5.20: HBEC counts at the beginning and at the end of 120h INTS12 knockdown experiment. Statistical tests were performed comparing to scrambled D-siRNA control: *P<0.05, ***P<0.001. No significant difference was observed between un-transfected and scrambled D-siRNA transfected cells.

5.3.5 Using RNAseq, end-point PCR and Sanger sequencing to decipher the genetic architecture of INTS12 locus

In this thesis, the RNAseq datasets were used not only to understand the transcriptomic dynamics following INTS12 depletion, but also to uncover splicing at the *INTS12* locus. Figure 5.21 compares transcriptome assemblies for *INTS12* locus leveraging the 48h and 120h RNAseq datasets. 5 novel variants were discovered based on the 48h dataset while 2 new variants were discovered based on the 120h dataset. Since in these experiments INTS12 was targeted by D-siRNA knockdown and silencing efficiency was improved at day 5, there were more reads available for the assembly at day 2 and hence the quality of this assembly was better due to higher sequencing depth at the *INTS12* locus. Nevertheless, importantly, novel variants detected at day 5 are represented at day 2 as well (starred in Figure 5.21).

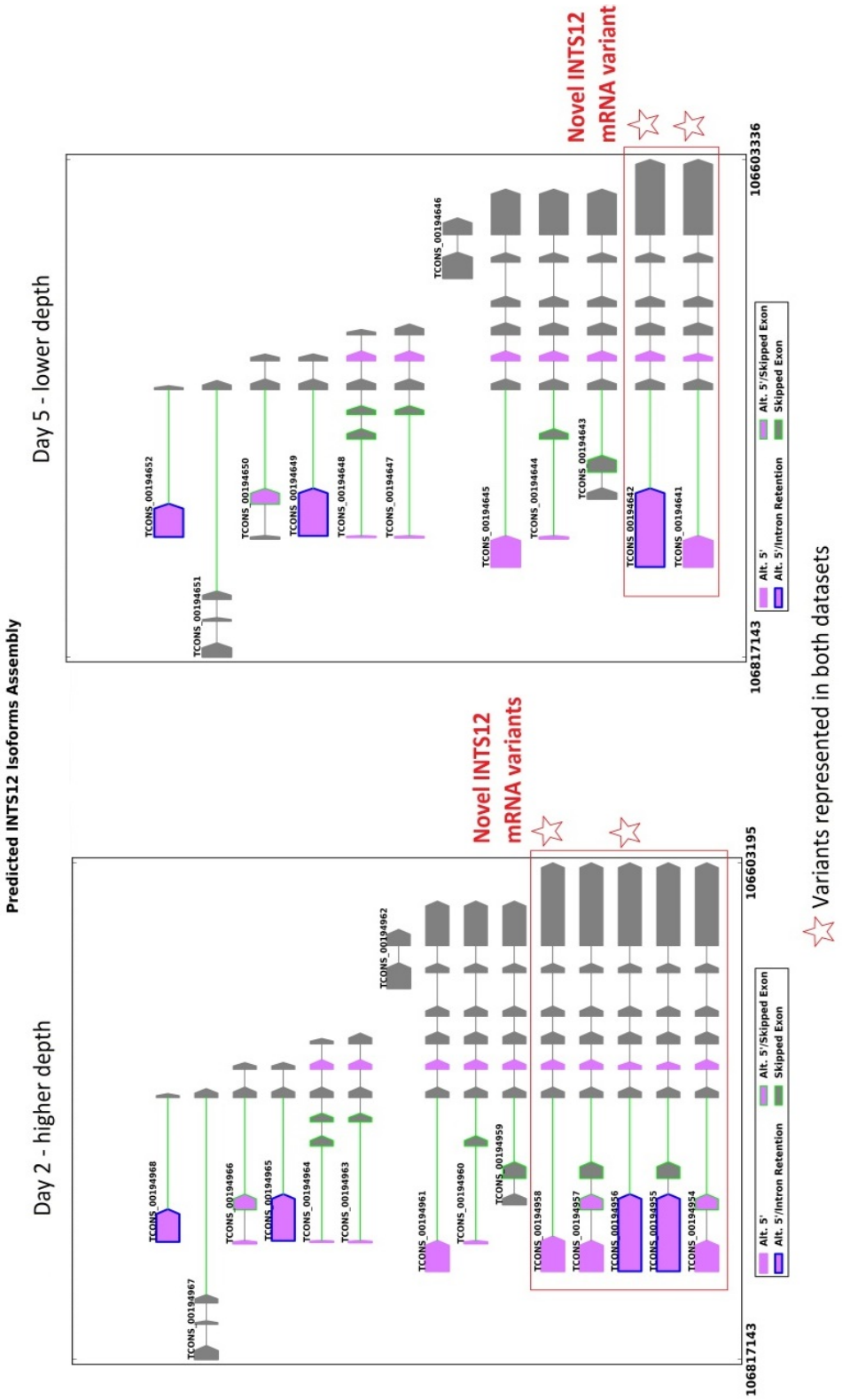
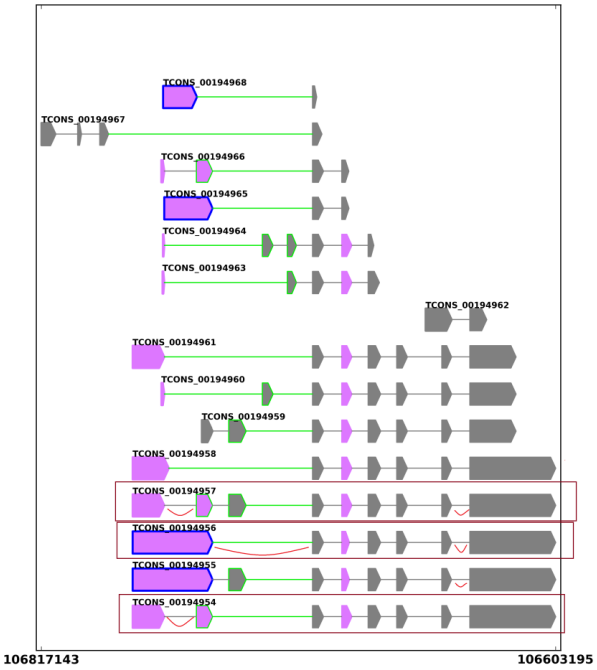


Figure 5.21: Comparison of INTS12 mRNA isoform assemblies at 48h (day 2, left side) and at 120h (day 5, right side). Despite the differences in sequencing depth between these two datasets, the 2 novel variants discovered at 120h are represented in the 48h dataset in which 5 novel variants were uncovered (indicated by stars).

There is an expected and acceptable discrepancy at the 5' end of the first exon (starts from right hand side on Figure 5.12) as indicated by slightly different coordinate on the X-axis. These are due to random coordinate estimation at the end of exons with no reads spanning the splice junction (Trapnell et al. 2010). However, the variants do agree in terms of internal exons as splice junction data is available for them. Because 48h data is more reliable in terms of transcriptome assembly for this gene due to higher depth of sequencing, it was taken forward for obtaining isoform sequences that were used as templates in primer design.

Figure 5.22 shows the new isoforms chosen for targeting by PCR (in red boxes) and their two unique novel splice junctions indicated. As it can be seen, these isoforms have a common forward primer and are differentiated by the reverse primer. These isoforms are among the least common INTS12 variants in basal HBECs (<1% out of the total pool of INTS12 isoforms; Figure 5.22). The predicted sequences of target isoforms are shown in Figure 5.23 while predicted amplicon sizes are shown in Table 5.5.



Transcript ID	Percent of total pool (%)
TCONS_00194968	1.66
TCONS_00194967	0.00
TCONS_00194966	0.31
TCONS_00194965	3.36
TCONS_00194964	1.57
TCONS_00194963	0.21
TCONS_00194962	0.73
TCONS_00194961	7.09
TCONS_00194960	83.82
TCONS_00194959	0.40
TCONS_00194958	0.01
TCONS_00194957	0.00
TCONS_00194956	0.71
TCONS_00194955	0.00
TCONS_00194954	0.13

Figure 5.22: Novel INTS12 splice variants chosen for PCR and sequencing validation (in red boxes) are shown with indicated splice sites to which forward and reverse primers align. Relative abundances of each isoform are shown in table next to the figure and are rounded-up to two decimal places.

TCONS_00194954

CGTACTATGTGCATGTATTACATTGGCAATGTAGGAAAAGTGAGCAATGTGAGAAAAAATATTGTATTAAGCACCAGGAGAGTTCACAATTTATTGGAAAA
AGATATACCCTTAAATAATTGAAATGAGGTGATAAAAAAGGTAAAGATATAATATGAAGTATGATAAATCAATTAGGAAAATCCATTGATATGAGAAATCAA
TAATGAAAAAATAATTTTGAAAAAGATGAAATAATTGGGAGGGAGAGGGAGAGTCAGTTGAAATGGGGCTAGCACAGAAAATGAACACCCAGAATACGG
GGAGGAAAATAGAAAACTATGTTGGGTCGTCCCTGCCGCCACATAAGGCACCTCATTCATCAGGATCAACTCTTGGGGCTTTGCTAGAAGATGAACTAT
TTCTTTTTCCTTAAATTCACAGATGCTTTCAATCTTCTGAAGATCTCATTGAGAACCAGTCATTCTAATCATTGTTTTCACACACTATGGAAATCATTAGAAT
AAAAATGTTGAAGATTGATTTAAAATGAAAGTTTCCAAGTTTTGTTATATAATAATTTAGCATTTTAAGGTAAGAAACAATAGAAATTTGATTATGAAGACTTT
TATTAAATTACAGTGTATTACAGATTATATCATAATAATAAGCCTTTCATCTTTAGGCTAATATGATACAAAAACCTACTTGGCCACATTACTTCTTGAGTTT
CTTTTGGGCAGCTTTCTTCTTGACCATCTGTAATCGCTTCATAGCATTGAGCTGTGATTCTTGTGAAGTTGGGCCTTTAAGGGATGCTGAGGGAGAGCTG
CTGGATTCTGAAGTAGTTTTGCTGGTAGTACTTCCACTAGGTCCTGATGTTCCACTATTTCCATTCCCACTTAGTTGGCTGCTGCTTCCTGGAACTAAACT
ACTTGGACTAGGAAGACCTACTTTGCTGACATTGTCACAACTAACTGAGCGACTAAGGCCAGTTTTACCCAAGGTTAGAGGTGGAGGTGGTTTTAAAGG
TACAGTGGGCGTAGTGCTGTTATTGGAACCTATTTTGGAACCTATTCCACCTTTGGATGATGTTGCCAGACCAGTCAAACCCACAGGTTTCTGGTTAGCT
GACGAAGTAGCAGGTTTCCCAGTATTGTTTTGTGTTGTTGAACTCAATTTTGCTGTTGAAGGACCAGCAGAGGAAGTTTTGGCTGCAAAAGCTGCCCATC
CAGTTAAGCCACTAGTTACTGACGAGGAAACGCTGGCACTAGAAGAATTTCCTGAAATAACT**GTGGATGT**CTTGACTTCTGTTCTCTTAAACGCTAGAAA
AGTTGTCTCTTGTTTCAGTTTAGTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGGGCTGGTTTCTGCGGTGGT
TTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACATCGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTT
TATGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGATGCCACCATCATTTGCCTACAAACAACGCAGGCCAATCC
CATCTCCATGGCAAAATCATCAGCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACAGTAATGGGAGATGACTGT
GTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACTCCTTCAGTGATGTCTGATTTCATTTTATCAGCAGGTCTCTTTTCAGCTTCCTTCTTTACCT
TTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTGGGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTG
GGTGGCTCCACATCCTTTTGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTTAGCTTTCAGCAGAATCTTTAC
TCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATGGGATCAAGTTCCAAGTTCACAGTAGCAGCCATTGCAAACGCCTGGCAATCTCAAAAGAAGTC
CAAAGTTGATCATATGCACGTGCCCCAGATGAAGGAAACAGAAACTGTTCACATAGAGTGCAGTATTATCGTTGCCATTTCATAATGGCTCTAAAAAACT
TTGTTCTTCATGACTTCGCAAGCTCTTGCAACCCCTTCCAATAAACATTATTTTAAAGATGCTGCCTTTTTCATTCCTCCACTGAGTTCTTGGTTCTGGATA
ACACTGATAC**CTTAGGGACAC**CGTTCCGCCCCGCCCTGCCGATCCGTCTGTTCCCGGTGGTCCCTTCGGAAACGGTTCCCGCACTGGCCGGCTCCGA
AAGCAGGAAAACAAAGGTTCCCACAGTAGGGGCGGGGGAAACGTTTGGCAGTGCGACAGTAGGAAGTGACGTTACTTCCCTTTTCCGGTCCGCCGG
ATTATGAATGACGGCCGGCGCGAGTATTTTCCACATAAGGTGGCTGTCGTTTTTCTCCTGGCGTCTGTGGAGGCGAGTGGTCTGCGGGCAGCAGCTCC
CAGAGGCAGCCTTGGAATTCCAGCTCGGACTGGGCGGGAAGGCGCAGGCGGCCCAGGTCGCCGACACGCTCACGCACCCTCCCTGCCTGGCCGCC
CCTCTGCGACCAGGTAAAGAGGGCGCTCGGGCCGCCGGCTTCTCAGCCTCCGCGACCCCCTAAGAACTGGTCTTTTCTTCGGGGGTCTGCAGGGCTG
AGGATGCG

TCONS_00194956

CGTACTATGTGCATGTATTACATTGGCAATGTAGGAAAAGTGAGCAATGTGAGAAAAAATATTGTATTAAGCACCAGGAGAGTTCACAATTTATTGGAAAA
AGATATACCCTTAAATAATTGAAATGAGGTGATAAAAAAGGTAAAGATATAATATGAAGTATGATAAATCAATTAGGAAAATCCATTGATATGAGAAATCAA
TAATGAAAAAATAATTTTGAAAAAGATGAAATAATTGGGAGGGAGAGGGAGAGTCAGTTGAAATGGGGCTAGCACAGAAAATGAACACCCAGAATACGG
GGAGGAAAATAGAAAACTATGTTGGGTCGTCCCTGCCGCCACATAAGGCACCTCATTCATCAGGATCAACTCTTGGGGCTTTGCTAGAAGATGAACTAT
TTCTTTTTCCTTAAATTCACAGATGCTTTCAATCTTCTGAAGATCTCATTGAGAACCAGTCATTCTAATCATTGTTTTCACACACTATGGAAATCATTAGAAT
AAAAATGTTGAAGATTGATTTAAAATGAAAGTTTCCAAGTTTTGTTATATAATAATTTAGCATTTTAAGGTAAGAAACAATAGAAATTTGATTATGAAGACTTT
TATTAAATTACAGTGTATTACAGATTATATCATAATAATAAGCCTTTCATCTTTAGGCTAATATGATACAAAAACCTACTTGGCCACATTACTTCTTGAGTTT
CTTTTGGGCAGCTTTCTTCTTGACCATCTGTAATCGCTTCATAGCATTGAGCTGTGATTCTTGTGAAGTTGGGCCTTTAAGGGATGCTGAGGGAGAGCTG
CTGGATTCTGAAGTAGTTTTGCTGGTAGTACTTCCACTAGGTCCTGATGTTCCACTATTTCCATTCCCACTTAGTTGGCTGCTGCTTCCTGGAACTAAACT
ACTTGGACTAGGAAGACCTACTTTGCTGACATTGTCACAACTAACTGAGCGACTAAGGCCAGTTTTACCCAAGGTTAGAGGTGGAGGTGGTTTTAAAGG
TACAGTGGGCGTAGTGCTGTTATTGGAACCTATTTTGGAACCTATTCCACCTTTGGATGATGTTGCCAGACCAGTCAAACCCACAGGTTTCTGGTTAGCT
GACGAAGTAGCAGGTTTCCCAGTATTGTTTTGTGTTGTTGAACTCAATTTTGCTGTTGAAGGACCAGCAGAGGAAGTTTTGGCTGCAAAAGCTGCCCATC
CAGTTAAGCCACTAGTTACTGACGAGGAAACGCTGGCACTAGAAGAATTTCCTGAAATAACT**GTGGATGT**CTTGACTTCTGTTCTCTTAAACGCTAGAAA
AGTTGTCTCTTGTTTCAGTTTAGTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGGGCTGGTTTCTGCGGTGGT
TTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACATCGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTT
TATGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGATGCCACCATCATTTGCCTACAAACAACGCAGGCCAATCC
CATCTCCATGGCAAAATCATCAGCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACAGTAATGGGAGATGACTGT
GTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACTCCTTCAGTGATGTCTGATTTCATTTTATCAGCAGGTCTCTTTTCAGCTTCCTTCTTTACCT
TTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTGGGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTG

GGTGGCTCCACATCCTTTTGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTTAGCTTTTCAGCAGAATCTTTAC
TCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATGGGATCAAGTTCCAAGTTCACAGTAGCAGCCATTGCAAACGCCTTAGCACTTCAGTAGAAAGT
GTGCCACCCCCACCTCCATCTTAAGTCACCTCTCAAGTTTCAGTTTTTCCTGCACATTCCCTCAACAAAACTAGAAGTTTCACTGTTACATAGCCCAAAAA
CATTCTGTACTTATTTGTAATATTAATCACACTGTTAATTATGTAAATAATCTTTCTTTCTCATTAGAATGTAAGTTCCATGAAGGCAGGCAGTGGATTGGT
CTTGCCTACTGTTGTAACCCCTGTCTGGCAGAGTTCCTGGCAATCTCAAAAGAAGTCCAAAGTTGATCATATGCACGTGCCCCAGATGAAGGAAACAGA
AACTGTTCACATAGAGTGCAGTATTATCGTTGCCATTTCATAATGGCTCTAAAAAACTTTGTTCTTCATGACTTCGCAAGCTCTTGCAACCCCTTCCAATA
AACATTATTTTAAAGATGCTGCCTTTTTCATTCCTCCACTGAGTTCTTGGTTCTGGATAACACTGATA**ACCTTAGGGACAC**CGTTCCGCCCCGCCCTGCCG
ATCCGTCTGTTCCCGGTGGTCCCTTCGGAAACGGTTCCCGCACTGGCCGGCTCCGAAAGCAGGAAAACAAAGGTTCCCACAGTAGGGGCGGGGGAAA
CGTTTGGCAGTGCGACAGTAGGAAGTGACGTTACTTCCCTTTTTCCGGTCCGCCGGATTATGAATGACGGCCGGCGCGAGTATTTTCCACATAAGGTG
GCTGTCGTTTTTCTCCTGGCGTCTGTGGAGGCGAGTGGTCTGCGGGCAGCAGCTCCCAGAGGCAGCCTTGGAATTCCAGCTCGGACTGGGCGGGAAG
GCGCAGGCGGCCCAGGTCGCCGACACGCTCACGCACCCTCCCTGCCTGGCCGCGCCTCTGCGACCAGGTAAAGAGGGCGCTCGGGCCGCCGGCTT
CTCAGCCTCCGCGACCCCCTAAGAACTGGTCTTTTCTTCGGGGGTCTGCAGGGCTGAGGATGCG

Figure 5.23: Predicted sequences of the novel INTS12 mRNA variants. Alternating colours indicate different exons while bold indicates the forward and reverse primer sites.

Isoform ID	Amplicon size
TCONS_00194957	1330 bp
TCONS_00194956	796 bp
TCONS_00194954	1072 bp

Table 5.5: The predicted amplicon sizes corresponding to the novel INTS12 mRNA variants

5.3.5.1 End-point PCR validation

Figure 5.24 shows the image of the electrophoresis gel for the amplified novel INTS12 mRNA products. Water and RT- samples had no DNA amplicons detected. TCONS_00194954 was detected in all RT+ samples and the bands are at expected position relative to the ladder (aligned above 1000 below 1100). TCONS_00194956 was detected in all RT+ samples and bands are relatively close to the expected position of ~800bps (aligned above 800 below 900). TCONS_00194957 was detected in two RT+ samples out of the three (both donors) and bands are relatively close to the expected position of 1330bps (aligned above 1000 below 1200). Nevertheless, TCONS_00194956 and TCONS_00194957 variants, in contrast to TCONS_00194954, are a little off their expected position. Interestingly, bands intensities correspond to the percent abundance of total pool found: TCONS_00194956 had the highest abundance followed by TCONS_00194954, followed by TCONS_00194957 (Figure 5.22).

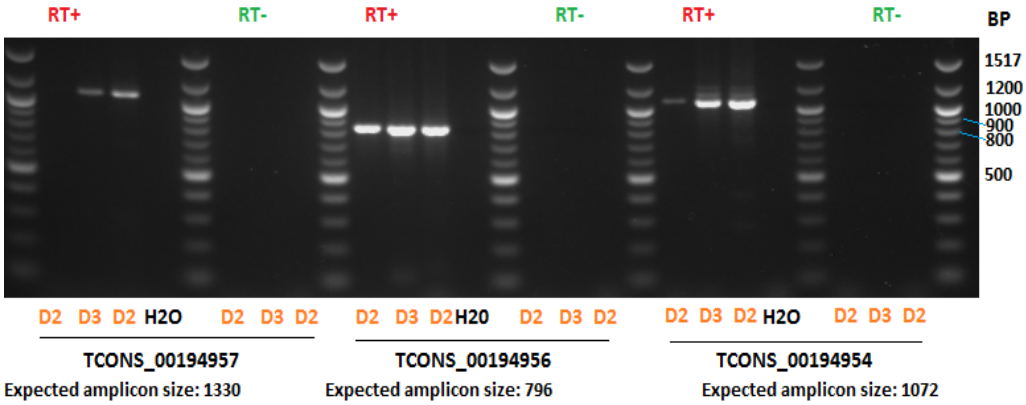


Figure 5.24: Gel electrophoresis of novel INTS12 mRNA variants following their PCR amplification.

5.3.5.2 Sequence verification of the amplicons

The sequence of novel variants was determined by Sanger reaction and was compared to the predicted sequence obtained from the RNAseq analysis. This is particularly important in the light of slight discrepancy between the expected and actual band sizes observed for TCONS_00194956 and TCONS_00194957 variants.

5.3.5.2.1 TCONS_00194954 analysis

There was 96% identity between these two sequences. The difference is mostly due to sequencing error at the 3' end of Sanger sequence as perfect alignment is observed within the sequences. The lack of the alignment at the 5' of the sequences is because the sequence of the primer is not retrieved in Sanger sequencing (Figure 5.25). Therefore, it is possible to say that there is an agreement between the RNAseq predicted and Sanger sequencing determined sequence of TCONS_00194954 variant. Thus not only was this variant validated by PCR by aligning at the expected position after gel electrophoresis (Figure 5.24) but verified on a sequence level as well.

```
T54_RNA_seq     GTGGATGTCTTGACTTCTGTTCTCTTAAACGCTAGAAAAGTTGTCTCTTGTTTCAGTTTA
T54_PCR         ---------------------------------CACAAGGTTGTCTCTTGTTT-AGTTTA
                                                 . **.************** ******

T54_RNA_seq     GTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGG
T54_PCR         GTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGG
                ************************************************************

T54_RNA_seq     GCTGGTTTCTGCGGTGGTTTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACAT
T54_PCR         GCTGGTTTCTGCGGTGGTTTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACAT
                ************************************************************

T54_RNA_seq     CGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTTTA
T54_PCR         CGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTTTA
                ************************************************************

T54_RNA_seq     TGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGAT
T54_PCR         TGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGAT
                ************************************************************

T54_RNA_seq     GCCACCATCATTTGCCTACAAACAACGCAGGCCAATCCCATCTCCATGGCAAAATCATCA
T54_PCR         GCCACCATCATTTGCCTACAAACAACGCAGGCCAATCCCATCTCCATGGCAAAATCATCA
                ************************************************************

T54_RNA_seq     GCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACA
T54_PCR         GCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACA
                ************************************************************

T54_RNA_seq     GTAATGGGAGATGACTGTGTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACT
T54_PCR         GTAATGGGAGATGACTGTGTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACT
                ************************************************************

T54_RNA_seq     CCTTCAGTGATGTCTGATTTCATTTTATCAGCAGGTCTCTTTTCAGCTTCCTTCTTTACC
T54_PCR         CCTTCAGTGATGTCTGATTTCATTTTATCAGCAGGTCTCTTTTCAGCTTCCTTCTTTACC
                ************************************************************

T54_RNA_seq     TTTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTG
T54_PCR         TTTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTG
                ************************************************************

T54_RNA_seq     GGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTGGGTGGCTCCACATCCTTT
T54_PCR         GGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTGGGTGGCTCCACATCCTTT
                ************************************************************

T54_RNA_seq     TGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTT
T54_PCR         TGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTT
                ************************************************************

T54_RNA_seq     AGCTTTTCAGCAGAATCTTTACTCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATG
T54_PCR         AGCTTTTCAGCAGAATCTTTACTCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATG
                ************************************************************

T54_RNA_seq     GGATCAAGTTCCAAGTTCACAGTAGCAGCCATTGCAAACGCCTGGCAATCTCAAAAGAAG
T54_PCR         GGATCAAGTTCCAAGTTCACAGTAGCAGCCATTGCAAACGCCTGGCAATCTCAAAAGAAG
                ************************************************************

T54_RNA_seq     TCCAAAGTTGATCATATGCACGTG-CCCCAGATGAAGGAAACAGAAACTGTTCACATAGA
T54_PCR         TCCAAAGTTGATCATATGCACGTGCCCCCCAGATGAAGGAAACAGAAACTGTTCACATAGA
                ************************ ************************************

T54_RNA_seq     GTGCAGTATTATCGTTGCCATTTCATAATGGCTCT-AAAAAACTTTGTTCTTCATGACTT
T54_PCR         GTGCAATATTATCGCTGCCATTTCATAATGGCTCTAAAAAAACTTTATCATTCATGACTT
                *****.********.******************** **********.*. **********

T54_RNA_seq     CGCAAGCTCTTGCAACCCCTTCCAATAAACATTATTTTAAAGATGCTGCCTTTTTCATTC
T54_PCR         CCCAAGCTCTTGCAACCCCTTCCAATAAACAT--ATTTAAAGATGCTGAATATTCATTCC
                * ****************************   *************  * **.  *.*

T54_RNA_seq     CTCCACTGAGTTCTTGGTTCTGGATAAC-----ACTGATACCTTAGGGACACCG----TT
T54_PCR         TTCCTACTGAGTCCTGATTCTGGATTACGAGTGATAAGCAAAAAAGAAACACTATATCTG
                .***  . .. **.**.******** **     *. ...*    **..****..    *

T54_RNA_seq     CCG
T54_PCR         TCA
                 .*.
```

Figure 5.25: Multiple sequence alignment of RNAseq predicted and Sanger sequencing sequenced TCONS_00194954 variant

5.3.5.2.2 TCONS_00194956 analysis

There appeared to be sequence discrepancy inside the alignment which meant that the sequence of amplified DNA is different than that of isoform sequence found via RNAseq (Figure 5.26). Importantly, the length of the Sanger sequenced isoform is larger than the one identified via RNAseq approach and this goes hand in hand with the observation that the DNA band appears slightly larger than expected on the agarose gel. Interestingly, the 'insert sequence' observed in amplified DNA falls precisely between exons 4 and 5 (Figure 5.27). Therefore, Sanger sequencing data provided evidence for either misplacing of exon boundaries or a missed exon between exons 4 and 5.

```
T56_RNA_seq     GTGGATGTCTTGACTTCTGTTCTCTTAAACGCTAGAAAAGTTGTCTCTTGTTTCAGTTTA
T56_F_PCR       ---------------------------------CAACAAGTTGTCTCTTGTTT-AGTTTA
                                                 .* ***************** ******

T56_RNA_seq     GTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGG
T56_F_PCR       GTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGG
                ************************************************************

T56_RNA_seq     GCTGGTTTCTGCGGTGGTTTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACAT
T56_F_PCR       GCTGGTTTCTGCGGTGGTTTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACAT
                ************************************************************

T56_RNA_seq     CGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTTTA
T56_F_PCR       CGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTTTA
                ************************************************************

T56_RNA_seq     TGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGAT
T56_F_PCR       TGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGAT
                ************************************************************

T56_RNA_seq     GCCACCATCATTTGCCTACAAACAACGCAGGCCAATCCCATCTCCATGGCAAAATCATCA
T56_F_PCR       GCCACCATCATTTGCCTACAAACAACGCAGGCCAATCCCATCTCCATGGCAAAATCATCA
                ************************************************************

T56_RNA_seq     GCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACA
T56_F_PCR       GCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACA
                ************************************************************

T56_RNA_seq     GTAATGGGAGATGACTGTGTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACT
T56_F_PCR       GTAATGGGAGATGACTGTGTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACT
                ************************************************************

T56_RNA_seq     CCTTCAGTGATGTCTGATTTCAT-------------------------------------C
T56_F_PCR       CCTTCAGTGATGTCTGATTTCATTTTATCAGCAGGTCTCTTTTCAGCTTCCTTCTTTACC
                ***********************                                     *

T56_RNA_seq     TTTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTG
T56_F_PCR       TTTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTG
                ************************************************************

T56_RNA_seq     GGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTGGGTGGCTCCACATCCTTT
T56_F_PCR       GGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTGGGTGGCTCCACATCCTTT
                ************************************************************

T56_RNA_seq     TGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTT
T56_F_PCR       TGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTT
                ************************************************************

T56_RNA_seq     AGCTTTTCAGCAGAATCTTTACTCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATG
T56_F_PCR       AGCTTTTCAGCAGAATCTTTACTCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATG
                ************************************************************

T56_RNA_seq     GGATCAAGTTCCAAGTTCACAGTAGCAGCCATTGCAAACGCCT-GGCAATCTC-
T56_F_PCR       GGATCAAGTTCCAAGTTCACAGTAGCAGCCATTGCAAACGCCTGGGCAATCTCA
                ****************************************** *********
```

Figure 5.26: Multiple sequence alignment of RNAseq predicted and Sanger sequenced sequence of TCONS_00194956 revealed an insert sequence highlighted in red.

CGTACTATGTGCATGTATTACATTGGCAATGTAGGAAAAGTGAGCAATGTGAGAAAAAATATTGTATTAAGCACCAGGAGAGTTCACAATTTATTGGAAAA
AGATATACCCTTAAATAATTGAAATGAGGTGATAAAAAAGGTAAAGATATAATATGAAGTATGATAAATCAATTAGGAAAATCCATTGATATGAGAAATCAA
TAATGAAAAAATAATTTTGAAAAAGATGAAATAATTGGGAGGGAGAGGGAGAGTCAGTTGAAATGGGGCTAGCACAGAAAATGAACACCCAGAATACGG
GGAGGAAAATAGAAAACTATGTTGGGTCGTCCCTGCCGCCACATAAGGCACCTCATTCATCAGGATCAACTCTTGGGGCTTTGCTAGAAGATGAACTAT
TTCTTTTTCCTTAAATTCACAGATGCTTTCAATCTTCTGAAGATCTCATTGAGAACCAGTCATTCTAATCATTGTTTTCACACACTATGGAAATCATTAGAAT
AAAAATGTTGAAGATTGATTTAAAATGAAAGTTTCCAAGTTTTGTTATATAATATTTAGCATTTTAAGGTAAGAAACAATAGAAATTTGATTATGAAGACTTT
TATTAAATTACAGTGTATTACAGATTATATCATAATAATAAGCCTTTCATCTTTAGGCTAATATGATACAAAAACCTACTTGGCCACATTACTTCTTGAGTTT
CTTTTGGGCAGCTTTCTTCTTGACCATCTGTAATCGCTTCATAGCATTGAGCTGTGATTCTTGTGAAGTTGGGCCTTTAAGGGATGCTGAGGGAGAGCTG
CTGGATTCTGAAGTAGTTTTGCTGGTAGTACTTCCACTAGGTCCTGATGTTCCACTATTTCCATTCCCACTTAGTTGGCTGCTGCTTCCTGGAACTAAACT
ACTTGGACTAGGAAGACCTACTTTGCTGACATTGTCACAACTAACTGAGCGACTAAGGCCAGTTTTACCCAAGGTTAGAGGTGGAGGTGGTTTTAAAGG
TACAGTGGGCGTAGTGCTGTTATTGGAACCTATTTTGGAACCTATTCCACCTTTGGATGATGTTGCCAGACCAGTCAAACCCACAGGTTTCTGGTTAGCT
GACGAAGTAGCAGGTTTCCCAGTATTGTTTTGTGTTGTTGAACTCAATTTTGCTGTTGAAGGACCAGCAGAGGAAGTTTTGGCTGCAAAAGCTGCCCATC
CAGTTAAGCCACTAGTTACTGACGAGGAAACGCTGGCACTAGAAGAATTTCCTGAAATAACT**GTGGATGT****CTTGACTTCTGT**TCTCTTAAACGCTAGAAA
AGTTGTCTCTTGTTTCAGTTTAGTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGGGCTGGTTTCTGCGGTGGT
TTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACATCGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTT
TATGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGATGCCACCATCATTTGCCTACAAACAACGCAGGCCAATCC
CATCTCCATGGCAAAATCATCAGCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACAGTAATGGGAGATGACTGT
GTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACTCCTTCAGTGATGTCTGATTTCAT..CTTTTCAGTTGTGAGGACCTTGCCATTATTATTACC
AGAAGGAAGACTGGATGATATTTTGGGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTGGGTGGCTCCACATCCTTTTGAGATGGACGGTAA
CTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTAGCTTTTCAGCAGAATCTTTACTCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAA
AATGGGATCAAGTTCCAAGTTCACAGTAGCAGC**CATTGCAAACGC****CTGGCAATCTC**AAAAGAAGTCCAAAGTTGATCATATGCACGTGCCCCAGATGAA
GGAAACAGAAACTGTTCACATAGAGTGCAGTATTATCGTTGCCATTTCATAATGGCTCTAAAAAACTTTGTTCTTCATGACTTCGCAAGCTCTTGCAACCC
CTTCCAATAAACATTATTTTAAAGATGCTGCCTTTTTCATTCCTCCACTGAGTTCTTGGTTCTGGATAACACTGATACCTTAGGGACACCTGGAATCGCAG
GAGAAAAGGCTCCGATTCCAAACCAATCAAGTGGCTGAAAGCTGACAGCAAACACTTAGAAAAGGAATGTCAATCCTTTGATCAAAGCATCACACATCG
AAGTCCTAGAATGGGGGATGGTCCTTGACAGGGGAGCCAAAAGATGAGAGACAATGAGGGAGCTACGTAGAGAAGTCACTAAAACCCTCCGACAAAGT
CCGAGGCAGGATTCCAAGTCCCTTGTCCCTAGACCTAGAAATATGCTCGGTCCGCATGTCCCGGCCCGCAACTCCCTCGGAGGAAGCACAGTCCTCAC
TGAAAGGGGAGAGCAGTCAGTCTAGGGCACGCAACCGCCCGTAAATTATGATTCCCCTTGAGGAGCGAGGAAACTAATACCCCACGTAGCTACTCTCC
CACCTCGCTCCTGGCGCGGGGGTCTCGAGCCTCCAGGAGGCCAGGAGCAGAGTCGCTCAGCATAACTCACCGTTCCGCCCCGCCCTGCCGATCCGT
CTGTTCCCGGTGGTCCCTTCGGAAACGGTTCCCGCACTGGCCGGCTCCGAAAGCAGGAAAACAAAGGTTCCCACAGTAGGGGCGGGGGAAACGTTTG
GCAGTGCGACAGTAGGAAGTGACGTTACTTCCCTTTTTCCGGTCCGCCGGATTATGAATGACGGCCGGGCGCGAGTATTTTCCACATAAGGTGGCTGTC
GTTTTTCTCCTGGCGTCTGTGGAGGCGAGTGGTCTGCGGGCAGCAGCTCCCAGAGGCAGCCTTGGAATTCCAGCTCGGACTGGGCGGGAAGGCGCA
GGCGGCCCAGGTCGCCGACACGCTCACGCACCCTCCCTGCCTGGCCGCGCCTCTGCGACCAGGTAAAGAGGGCGCTCGGGCCGCCGGCTTCTCAG
CCTCCGCGACCCCCTAAGAACTGGTCTTTTCTTCGGGGGTCTGCAGGGCTGAGGATGCG

Figure 5.27: RNAseq predicted sequence of TCONS_00194956 with indicated location (...) of the missing bit of the sequence.5.3.5.2.2.1 Refinement of TCONS_00194956 structure

In order to examine the origin of additional sequence observed by Sanger sequencing but not observed by RNAseq, the genomic sequence between the coordinates 4:106614643-106616709 representing the region between the exons 4 and 5 was obtained. Multiple sequence alignment of exon 4.1 and region 4:106614643-106616709 was performed and revealed a missing 36 base sequence that is adjacent to exon 5. Thus transcriptome assembler misplaced the splice junction and the true coordinates of TCONS_00194956's exon 5 is 36 bases away from the junction inferred from RNAseq (Figure 5.28 and Table 5.6). The final isoform size is 832bp and therefore considering this information it is possible to say that DNA band aligned in its expected position relative to the DNA ladder (Figure 5.24).

CGTACTATGTGCATGTATTACATTGGCAATGTAGGAAAAGTGAGCAATGTGAGAAAAAATATTGTATTAAGCACCAGGAGAGTTCACAATTTATTGGAAAA
AGATATACCCTTAAATAATTGAAATGAGGTGATAAAAAAGGTAAAGATATAATATGAAGTATGATAAATCAATTAGGAAAATCCATTGATATGAGAAATCAA
TAATGAAAAAATAATTTTGAAAAAGATGAAATAATTGGGAGGGAGAGGGAGAGTCAGTTGAAATGGGGCTAGCACAGAAAATGAACACCCAGAATACGG
GGAGGAAAATAGAAAACTATGTTGGGTCGTCCCTGCCGCCACATAAGGCACCTCATTCATCAGGATCAACTCTTGGGGCTTTGCTAGAAGATGAACTAT
TTCTTTTTCCTTAAATTCACAGATGCTTTCAATCTTCTGAAGATCTCATTGAGAACCAGTCATTCTAATCATTGTTTTCACACACTATGGAAATCATTAGAAT
AAAAATGTTGAAGATTGATTTAAAATGAAAGTTTCCAAGTTTTGTTATATAATATTTAGCATTTTAAGGTAAGAAACAATAGAAATTTGATTATGAAGACTTT
TATTAAATTACAGTGTATTACAGATTATATCATAATAATAAGCCTTTCATCTTTAGGCTAATATGATACAAAAACCTACTTGGCCACATTACTTCTTGAGTTT
CTTTTGGGCAGCTTTCTTCTTGACCATCTGTAATCGCTTCATAGCATTGAGCTGTGATTCTTGTGAAGTTGGGCCTTTAAGGGATGCTGAGGGAGAGCTG
CTGGATTCTGAAGTAGTTTTGCTGGTAGTACTTCCACTAGGTCCTGATGTTCCACTATTTCCATTCCCACTTAGTTGGCTGCTGCTTCCTGGAACTAAACT
ACTTGGACTAGGAAGACCTACTTTGCTGACATTGTCACAACTAACTGAGCGACTAAGGCCAGTTTTACCCAAGGTTAGAGGTGGAGGTGGTTTTAAAGG
TACAGTGGGCGTAGTGCTGTTATTGGAACCTATTTTGGAACCTATTCCACCTTTGGATGATGTTGCCAGACCAGTCAAACCCACAGGTTTCTGGTTAGCT
GACGAAGTAGCAGGTTTCCCAGTATTGTTTTGTGTTGTTGAACTCAATTTTGCTGTTGAAGGACCAGCAGAGGAAGTTTTGGCTGCAAAAGCTGCCCATC
CAGTTAAGCCACTAGTTACTGACGAGGAAACGCTGGCACTAGAAGAATTTCCTGAAATAACT**GTGGATGT**CTTGACTTCTGTTCTCTTAAACGCTAGAAA
AGTTGTCTCTTGTTTCAGTTTAGTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGGGCTGGTTTCTGCGGTGGT
TTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACATCGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTT
TATGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGATGCCACCATCATTTGCCTACAAACAACGCAGGCCAATCC
CATCTCCATGGCAAAATCATCAGCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACAGTAATGGGAGATGACTGT
GTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACTCCTTCAGTGATGTCTGATTTCATTTTATCAGCAGGTCTCTTTTCAGCTTCCTTCTTTACCT
TTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTGGGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTG
GGTGGCTCCACATCCTTTTGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTTAGCTTTTCAGCAGAATCTTTAC
TCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATGGGATCAAGTTCCAAGTTCACAGTAGCAGC**CATTGCAAACGC**CTGGCAATCTCAAAAGAAGT
CCAAAGTTGATCATATGCACGTGCCCCAGATGAAGGAAACAGAAACTGTTCACATAGAGTGCAGTATTATCGTTGCCATTTCATAATGGCTCTAAAAAAC
TTTGTTCTTCATGACTTCGCAAGCTCTTGCAACCCCTTCCAATAAACATTATTTTAAAGATGCTGCCTTTTTCATTCCTCCACTGAGTTCTTGGTTCTGGAT
AACACTGATACCTTAGGGACACCTGGAATCGCAGGAGAAAAGGCTCCGATTCCAAACCAATCAAGTGGCTGAAAGCTGACAGCAAACACTTAGAAAAGG
AATGTCAATCCTTTGATCAAAGCATCACACATCGAAGTCCTAGAATGGGGGATGGTCCTTGACAGGGGAGCCAAAAGATGAGAGACAATGAGGGAGCT
ACGTAGAGAAGTCACTAAAACCCTCCGACAAAGTCCGAGGCAGGATTCCAAGTCCCTTGTCCCTAGACCTAGAAATATGCTCGGTCCGCATGTCCCGG
CCCGCAACTCCCTCGGAGGAAGCACAGTCCTCACTGAAAGGGGAGAGCAGTCAGTCTAGGGCACGCAACCGCCCGTAAATTATGATTCCCCTTGAGGA
GCGAGGAAACTAATACCCCACGTAGCTACTCTCCCACCTCGCTCCTGGCGCGGGGGTCTCGAGCCTCCAGGAGGCCAGGAGCAGAGTCGCTCAGCAT
AACTCACCGTTCCGCCCCGCCCTGCCGATCCGTCTGTTCCCGGTGGTCCCTTCGGAAACGGTTCCCGCACTGGCCGGCTCCGAAAGCAGGAAAACAA
AGGTTCCCACAGTAGGGGCGGGGGAAACGTTTGGCAGTGCGACAGTAGGAAGTGACGTTACTTCCCTTTTTCCGGTCCGCCGGATTATGAATGACGGC
CGGCGCGAGTATTTTCCACATAAGGTGGCTGTCGTTTTTCTCCTGGCGTCTGTGGAGGCGAGTGGTCTGCGGGCAGCAGCTCCCAGAGGCAGCCTTG
GAATTCCAGCTCGGACTGGGCGGGAAGGCGCAGGCGGCCCAGGTCGCCGACACGCTCACGCACCCTCCCTGCCTGGCCGCGCCTCTGCGACCAGGT
AAAGAGGGCGCTCGGGCCGCCGGCTTCTCAGCCTCCGCGACCCCCTAAGAACTGGTCTTTTCTTCGGGGGTCTGCAGGGCTGAGGATGCG

Figure 5.28: Sanger sequencing corrected and validated sequence of TCONS_00194956.

EXON NUMBER	EXON COORDINATES		EXONS SIZE
1	106603195	106604474	1280
2	106607849	106607995	147
3	106613133	106613292	160
4	106614456	106614643	188
5	106616709 – 36 = 106616673	106616825	117+36 = 153
6	106621007	106621171	165
7	106629076	106630265	1190

Table 5.6: The final coordinates of TCONS_00194956 isoform with corrected coordinates of exon 5. Exon size calculated as (stop coordinate – start coordinate) + 1.

5.3.5.2.3 TCONS_00194957 analysis

As it was observed for TCONS_00194956 the identity of RNAseq and Sanger sequencing derived sequences were not the same due to sequence discrepancy within the sequence while there was a perfect alignment at each side of discrepancy region (Figure 5.29 and Table 5.7). From the examination of discrepancy region it appeared that the DNA amplicon had missing exon 7 because the sequences surrounding the discrepancy region perfectly corresponded to exon 6 and exon 8, i.e. the transcriptome assembler inserted an exon that do exist in this novel variant.

```
T57_RNA_seq     GTGGATGTCTTGACTTCTGTTCTCTTAAACGCTAGAAAAGTTGTCTCTTGTTTCAGTTTA
T57_F_PCR       ACAG----------------------------TAGAAAGTTGTCTCTTGTTT-AGTTTA
                ...*                            ..***************** ******

T57_RNA_seq     GTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGG
T57_F_PCR       GTTTCTGGTTTCTTAACCAATGGATCTTTGACAGCTGGAGTTACAGAAACAACTGCAGGG
                ************************************************************

T57_RNA_seq     GCTGGTTTCTGCGGTGGTTTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACAT
T57_F_PCR       GCTGGTTTCTGCGGTGGTTTCTGAGTTTTTTGAGCCATTCTTTTCATTTGTCTGGTACAT
                ************************************************************

T57_RNA_seq     CGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTTTA
T57_F_PCR       CGGGCACAATACCACACCAGGCGAGGGTCATTCGCTTCCTTGTCTGTCACCTGGGGTTTA
                ************************************************************

T57_RNA_seq     TGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGAT
T57_F_PCR       TGACAATCTCGGTGGTAGAGATTATGGCACTCCTGACATTCTACTAATTGATTGCCAGAT
                ************************************************************

T57_RNA_seq     GCCACCATCATTTGCCTACAAACAACGCAGGCCAATCCCATCTCCATGGCAAAATCATCA
T57_F_PCR       GCCACCATCATTTGCCTACAAACAACGCAGGCCAATCCCATCTCCATGGCAAAATCATCA
                ************************************************************

T57_RNA_seq     GCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACA
T57_F_PCR       GCACTGGTCTCCTCAAAACTGGAAAGGTCAGCCATAGGTAAATCCTTGCTACTTTGGACA
                ************************************************************

T57_RNA_seq     GTAATGGGAGATGACTGTGTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACT
T57_F_PCR       GTAATGGGAGATGACTGTGTTTCTGGTTTCTCCAATCTAGGTTTCTTTGGAATATCAACT
                ************************************************************

T57_RNA_seq     CCTTCAGTGATGTCTGATTTCATTTTATCAGCAGGTCTCTTTTCAGCTTCCTTCTTTACC
T57_F_PCR       CCTTCAGTGATGTCTGATTTCATTTTATCAGCAGGTCTCTTTTCAGCTTCCTTCTTTACC
                ************************************************************

T57_RNA_seq     TTTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTG
T57_F_PCR       TTTTCAGTTGTGAGGACCTTGCCATTATTATTACCAGAAGGAAGACTGGATGATATTTTG
                ************************************************************

T57_RNA_seq     GGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTGGGTGGCTCCACATCCTTT
T57_F_PCR       GGCTCTTGCTTAATGGAAATGTTTTTTGTGCTTGAAATTTTGGGTGGCTCCACATCCTTT
                ************************************************************

T57_RNA_seq     TGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTT
T57_F_PCR       TGAGATGGACGGTAACTGGAATCAATGCCCCGAGCCAAAGATTCATCAAGCAGTGCTTTT
                ************************************************************

T57_RNA_seq     AGCTTTTCAGCAGAATCTTTACTCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATG
T57_F_PCR       AGCTTTTCAGCAGAATCTTTACTCTTTGAATGCAAGAAACCTAGTGCTTTCAAAAAAATG
                ************************************************************

T57_RNA_seq     GGATCAAGTTCCAAGTTCACAGTAGCAGCCATTGCAAACGCCTTAGCACTTCAGTAGAAA
T57_F_PCR       GGATCAAGTTCCAAGTTCACAGTAGCAGCCATTGCAAACG--------------------
                ****************************************

T57_RNA_seq     GTGTGCCACCCCCACCTCCATCTTAAGTCACCTCTCAAGTTTCAGTTTTTCCTGCACATT
T57_F_PCR       ------------------------------------------------------------

T57_RNA_seq     CCCTCAACAAAACTAGAAGTTTCACTGTTACATAGCCCAAAAACATTCTGTACTTATTTG
T57_F_PCR       ------------------------------------------------------------

T57_RNA_seq     TAATATTAATCACACTGTTAATTATGTAAATAATCTTTCTTTCTCATTAGAATGTAAGTT
T57_F_PCR       ------------------------------------------------------------

T57_RNA_seq     CCATGAAGGCAGGCAGTGGATTGGTCTTGCCTACTGTTGTAACCCCTGTCTGGCAGAGTT
T57_F_PCR       ------------------------------------------------------------

T57_RNA_seq     CCTGGCAATCTCAAAAGAAGTCCAAAGTTGATCATATGCACGTGCCCCAGATGAAGGAAA
T57_F_PCR       CCTGGCAATCTCAAAAGAAGTCCAAAGTTGATCATATGCACGTGCCCCAGATGAAGGAAA
                ************************************************************

T57_RNA_seq     CAGAAA--CTGTTCACATAGAGTGCAGTATTATCGTTGCCATTTCATAATGGCTCTAAAA
T57_F_PCR       CAGAAAACTGTTTCACATAGAGTGCAGTATTATCGTTGCCATTTCATAATGGCTCTAAAA
                ******  .  ************************************************

T57_RNA_seq     AACTTTGTTCTTCATGACTTCGCAAGCTCTTGCAACCCCTT-CCAATAAACATT-ATTTT
T57_F_PCR       AACTTTGTTCTTCATGACTTCGCAAGCTCTTGCAACCCCTTCCCATTAAACATTAATTTT
```

```
                ************************************** *** ******** *****

T57_RNA_seq     AAAGATGCTGCCTTTTTCATTCCTCCACTGAGTTCTTGGTTCTGG-ATAACACTGATA-C
T57_F_PCR       AAAGATGCTGCCTTTTCATTCCTCCCACTGAGTTCTTGGTCTTGGAATAACACTGATACC
                ****************.  *.*..****************..*** ************ *

T57_RNA_seq     CTTAGGGACACCGTTC------
T57_F_PCR       CTTAAGGGGACCACCCGTTTCA
                ****.**. ***...*
```

Figure 5.29: The multiple sequence alignment of RNAseq predicted and Sanger sequenced sequence of TCONS_00194957 revealed a discrepancy between the two. This mis-inserted region appears to be an additional exon and is highlighted in red.

5.3.5.2.3.1 Refinement of TCONS_00194957 structure

Based on the above observation the coordinates of TCONS_00194957 can be refined and this implies the absence of exon 7. The updated sequence has a length of 1070bp. In context of this information the observed alignment of DNA amplicon relative to the ladder is as expected (Figure 5.24).

EXON NUMBER	EXON COORDINATES		EXONS SIZE
1	106603195	106604474	1280
2	106607849	106607995	147
3	106613133	106613292	160
4	106614456	106614643	188
5	106616673	106616825	153
6	106621007	106621171	165
8	106629076	106629319	244
9	106629795	106630265	471

Table 5.7: The final coordinates of TCONS_00194957 isoform with missing exon 7. Exon size calculated as (stop coordinate – start coordinate) + 1.

5.4 Discussion

In Chapter 4 data was presented showing a minor effect of INTS12 knockdown on snRNA processing in a HBEC model. Therefore, it is possible to say that allele carriers with low INTS12 expression, hence at risk of lowered lung function (Obeidat et al. 2013), are unlikely to have decreased levels of INTS12 to an extent resulting in snRNA misprocessing. Thus it may be difficult to argue that the potential effects of INTS12 on lung function occur exclusively via snRNA processing pathway. It is thus necessary to identify other biological pathways that may be underpinning the predisposition to low pulmonary function and/or COPD. Moreover, as mentioned in Chapters 3 and 4, the evolutionary conserved INTS12's PHD is disposable for *Drosophila*'s snRNA processing (Chen et al. 2013) strongly suggesting presence of other still unrealized activities.

RNAseq gene expression profiling following INTS12 depletion was used in the experiments described in this chapter, to provide insight on the regulatory properties of this gene. Therefore, hypothesis-free approach (Kheirallah et al. 2016) was relied on in order to generate new functional hypotheses about INTS12 function. Following knockdown with two D-siRNAs, marked downregulation of pathways critical in protein synthesis and forming part of the integrated stress response was observed. The top two downregulated pathways in the two independent D-siRNAs were the tRNA amino acetylation and PERK regulated gene expression. Although, from a respiratory perspective, interesting pathways appeared to be upregulated after INTS12 knockdown (e.g. collagen formation and extracellular matrix deposition) (Suki, 2005), these pathways show less robust effects than the downregulated pathways owning to effect sizes, degree of variance and reproducibility. Thus subsequent functional experiments were guided by the effects observed upon the downregulated pathways.

The RNAseq derived differences in gene expression correlated very strongly with qPCR derived differences in expression, technically validating the sequencing findings. Genes belonging to the key

downregulated pathways were also biologically validated in additional donor cells, again showing downregulation following INTS12 depletion. The weaker effect size in the validation donor can be attributed to the lesser efficiency of INTS12 knockdown. It is important to note that the key genes selected from PERK regulated gene expression and tRNA synthatases pathways have been convincingly show to have reduced expression as a result of INTS12 knockdown on the mRNA but not protein level. Future work may aim at showing this effect upon the proteins as well.

To further investigate the functional importance of these observations, additional experiments were undertaken which showed that suppression of INTS12 reduces protein synthesis. As doubling of protein content is necessary for cell division, a decrease in proliferative capacity was also seen. Intriguingly, the decrease of total cell numbers in D-siRNAs A and C conditions respectively at the end of the experiment, mirrored the observed reduction in protein translation. Thus the observed molecular signature impacted the relevant phenotypes, demonstrating INTS12 as a regulator of genes forming part of translational pathways. These novel data suggest as yet unrecognized function of INTS12 in regulating cellular translation in humans and possibly in other species as well.

It has been hypothesised that the mechanisms involved in the early human lung development may alter lung function and predispose to COPD later in life (Probert et al. 2015). Although a subset of lung function associated genes show evidence of differential expression between various stages of embryonic pulmonary tissue formation (Kheirallah et al. 2016; Miller et al. 2016), there is still an incomplete understanding of the molecular mechanisms behind normal respiratory system development and how the alterations therein contribute to disease pathophysiology. Given that there is no homologous *INTS12* in unsegmented *C. elegans* or unicellular *S. cerevisiae*, strong conservation and lethal effect of its knockout in *M. musculus* (Obeidat et al. 2013), this gene may have been important for the evolution of metazoan tissue differentiation and specialization.

It therefore seems plausible that INTS12 regulates lung development or repair. Thus genetically determined levels of INTS12 may in part account for the genetic association signal seen for lung function parameters via a developmental pathway. At this stage, it is not entirely clear how differences in the rates of protein synthesis may be responsible for the population variance in lung function or predisposition to COPD. It may be the case that individuals with slow collagen synthesis during lung development are vulnerable to low lung function later in life. Interestingly, in INTS12 depletion experiment, collagen formation and extracellular matrix organization pathways were upregulated which perhaps is activated as compensatory mechanism to cope with reduced collagen production. These conjectures require further investigation to be either falsified or ascertained.

INTS12 knockdown for 48h and 120h resulted in reproducible dysregulation of core subset of genes important in airway biology. Of particular interest is LEP which had 4.51 and 29.16-fold upregulation relative to control in D-siRNA A condition at 48h and 120h time points respectively. *LEP* genetically associates with the same lung function parameter as INTS12 (Wain et al. 2015, van den Borst et al. 2011) albeit weaker than what was reported for 4q24 locus. Crucially LEP levels negatively correlate with lung function (Eising et al. 2013). It might be the case that reduced levels of INTS12 in specific allele carriers are responsible for elevated expression of LEP which may in turn account for reduced lung function. Again, although these causal hypotheses provide biological understanding of the genetic association signal for pulmonary function, they still require further exploration.

Finally, three novel INTS12 mRNA variants were discovered using RNAseq datasets and these were doubly validated by end point PCR and Sanger sequencing. Out of these novel variants, one agreed with Sanger sequence and aligned at the expected position on the electrophoresis gel while the other two had some discrepancy between RNAseq and Sanger inferred internal structures. The misplacing of splice junctions for these two variants is not surprising considering the lower sequencing depth at the *INTS12* locus caused by targeted

knockdown. Relying on Sanger sequence the true structures of new mRNAs were refined.

6. Functional analysis of genome-wide INTS12 binding

6.1 Introduction

The work described in the previous chapters shows that INTS12 has a moderate role in snRNA processing (see Chapter 4) and that it predominantly regulates protein synthesis pathways (see Chapter 5) which is a previously unrecognized function for this gene. It is also possible to say that although INTS12's PHD motif domain is dispensable for snRNA processing in *Drosophila* this is likely to be true for other species. Therefore, the negative selection experienced by this particular part of the protein cannot be attributed to snRNA processing activity and requires an explanation.

In the light of data by Jodoin, Sitaram et al. it may be the case that this conservation is due to a role for INTS12 in the maintenance of perinuclear dynein (Jodoin, Sitaram et al. 2013) but equally likely it may be preserved due to a role in the positive regulation of protein synthesis pathways, because the lower the levels of INTS12 the lesser the rate of protein synthesis (see Chapter 5). Protein synthesis forms a key part of cellular translation and is necessary for life, i.e. without adequate protein synthesis the cell is not capable to sustain core processes. Therefore, the key question is how is it that INTS12 brings about its regulation of protein synthesis related pathways, which showed most prominent disruption in the pathway analyses? It is not possible to answer this question relying merely on the data presented in Chapter 5, as although the disruption in translational homeostasis was observed, it is unclear what the mechanism behind this observation is.

One possible scenario is that INTS12 levels dictate an output of some unknown process, which in turn triggers a cascade of reactions leading to the disruption of protein synthesis. Alternatively, INTS12 may be a direct regulator of genes belonging to protein synthesis pathways, thus downregulation of these pathways would be considered primary rather than secondary epiphenomena (Figure 6.1). These two schemes provide an explanation of molecular mechanisms leading to repression of protein synthesis following INTS12 depletion.

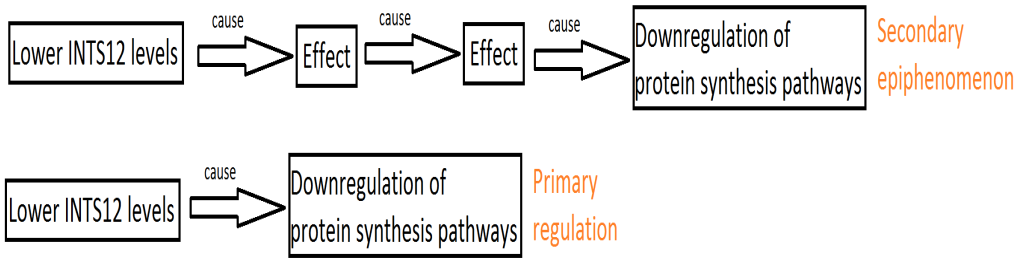


Figure 6.1: Is INTS12 a direct regulator of protein synthesis pathways, or is their dysregulation a secondary epiphenomenon?

It is clear from the literature that INTS12 is a nuclear protein (see section 1.8.5.1) and this has been confirmed by work presented in this thesis as well (Chapter 3). Initially it was thought that the nuclear localization of this protein is because of a role in snRNA processing which occurs in the nucleus (see section 1.8.4). In the first studies researchers failed to detect INTScom proteins near genes other than snRNAs (Baillat et al. 2005). However latter studies leveraging immunoprecipitation antibodies with improved antigen affinities, succeeded in detecting a subset of INTScom protein at the promoters of protein coding genes displaying POLII pausing which was increased following EGF stimulation (Gardini et al. 2014). This has been shown with INTS1 and INTS11 but it is unclear at this stage whether these two INTScom members are representative of all the INTScom proteins.

In *Drosophila* INTS12 has been detected by ChIP-PCR near the TSS of the gene coding for HSP70Aa and this binding was increased due to heat shock treatment (heat shock responsive genes are a classical model displaying POLII pausing in *Drosophila* and hence why HSP70Aa was chosen to test for INTS12 binding). Although these observations are promising for studying the genome-wide binding of INTS12, it cannot be a priori presumed that either INTS1 or INTS11 binding represents INTS12 binding nor that the human homolog of *Drosphila's* HSP70Aa would display INTS12 promoter localization. This is particularly important in the light of the moderate and minor roles for INTS12 in snRNA processing in fly and human cells respectively (see Chapter 4).

Moreover, if there is a plentiful amount of binding between INTS12 and the genome, this binding might underpin the differential gene expression

effects observed following INTS12 depletion. More generally it is not clear what would be the relationship between transcription and INTS12 binding properties. Last but not least, although INTS12's PHD domain is a putative histone H3 binding protein, based on the sequence similarity data, its co-localization with histone marks has not been explored before.

5.1.1 Aims and objectives

The aim of this Chapter is to functionally study the genome-wide binding properties of INTS12 in order to provide mechanistic insight into the effect of INTS12 knockdown on the identified target genes. As ChIPseq for INTS12 has not been performed before, the data from a pilot experiment is going to be presented first, as a proof of concept that INTS12 ChIPseq can yield enough depth to infer its binding sites. Then INTS12 binding characterization by investigating its interaction with fixed features (Marnetto et al. 2014) and cell-type-specific regulatory elements of the human genome (see section 1.5; Jiang and Pugh, 2009; Madrigal and Krajewski, 2012) will be shown. Finally, in order to understand the relationship between transcription in general and transcriptional dynamics following INTS12 depletion, the generated RNAseq (see Chapter 5) and ChIPseq datasets are going to be combined.

6.2 Materials and Methods

6.2.1 ChIPseq and ChIP-PCR

The detailed description of ChIPseq and ChIP-PCR experimental procedures is present in Chapter 2 (see sections 2.6.1 and 2.6.2 for ChIPseq and ChIP-PCR respectively).

6.2.2 ChIPseq data analysis

Reads were aligned to hg19 using BWA (Li and Durbin 2009) using default settings. Artefactual read duplicates were removed using samtools prior to further analyses. MACS INTS12 peak calling was run on each donor separately comparing ChIPseq samples to input control (Zhang et al. 2008). Calling was performed with a multiple comparisons corrected P value of less than 0.05 considered as significant. Generated

fragment pileup signal was normalized to library size. Fragment pileup was converted to wig files based on fold enrichment above input background for each donor. To compare peak metrics between two donor samples, overlapping intervals were grouped into active regions which are defined by the start coordinate of the most upstream binding interval and the end coordinate of the most downstream interval in both samples, i.e. it is the union of the overlapping intervals. In locations where only one sample has an interval, this interval defines the active region. ChIP signal at these active regions was compared between the two donor samples and correlation drawn and calculated by ggplot2 and rcmdr R packages respectively.

Intervals from representative donor were annotated, percentage of total INTS12 binding sites falling on the fixed annotated genomic features as well as enrichment over meta-gene body were determined using CEAS package (Shin et al. 2009). Enrichment over various gene classes, expressed and not expressed genes, and differentially and non-differentially expressed genes following INTS12 depletion was drawn using the ngs.plot package (Shen et al. 2014). Genes were defined as expressed or silenced based on un-transfected HBECs data in D195307. Genes were determined as differentially expressed or silent relying on the sustained knockdown RNAseq dataset. BETA was used to predict the regulatory function (Wang et al. 2013). Gene classes were retrieved using Ensembl's BioMart tool. HOMER and MEME were used for *de novo* identification of enriched DNA motif at INTS12 binding sites (Heinz et al. 2010, Machanick et al. 2011). TomTom was used to compare *de novo* identified motif to a set of currently known motifs (Gupta et al. 2007). Details about the used commands can be found in Chapter 2 (see section 2.8.2.1).

6.2.2.1 Epigenetic data from ENCODE

Airway epithelial cells specific epigenetic and CTCF ChIPseq datasets were obtained from ENCODE data repository (ENCBS417ENC; www.encodeproject.org) and analysed as INTS12 ChIPseq datasets with the only difference that broad region calling was used for the

epigenetic marks. Percent of overlap between INTS12 intervals and ENCODE intervals and its statistical significance was determined using regioneR R package (Gel et al. 2016) using random permutation test. Correlation of ChIPseq signals was performed using cistrome package (Liu et al. 2011).

6.2.2.2 An assessment of pilot ChIPseq experiment

For the purpose of determining the proportion of uniquely mapped and nonredundant reads the pilot sequencing library was aligned with the option of retaining only read tags that have only one hit. This was achieved with bowtie aligner (which is used internally by TopHat program) with the command shown below. In it, `-m 1` specifies that reads with only one hit on the genome are retained, `-S` specifies the output to be in SAM format, `-q` specifies the input to be in FASTQ format, `/path_to/` is the path to prebuilt bowtie indexes.

```
bowtie -m 1 -S -q path_to_bowtie_genome_index/hg19
INTS12_reads.fastq INTS12_reads.sam
```

6.3 Results

6.3.1 Pilot INTS12 ChIPseq

Before any large-scale production run of a ChIPseq experiment, it is desirable to conduct a pilot experiment (Ma and Wong, 2011). The purpose of pilot experiment is to provide valuable data for quality control prior to large-scale production run which may consume big amounts of material and sequencing machine time. This is particularly important for INTS12 for which a ChIPseq assay has never been run before. Previously it has been demonstrated that a polyclonal INTS12 antibody was specific for INTS12 by immunocytochemistry of bronchial epithelial cells in combination with knockdown (Chapter 3), as INTS12 nuclear signal completely disappeared in cells treated with D-siRNAs. According to ENCODE, an antibody is deemed appropriate for ChIPseq if either it is shown to be specific via a Western blot or by IF provided it is performed

in combination of gene depletion (Landt et al. 2012). Thus the tested antibody was used for the pilot INTS12 ChIPseq experiment.

For a pilot ChIPseq experiment to be considered successful (a) the sequencing depth ought to yield a total number of reads in the range of millions rather than thousands (b) percentage of uniquely mappable reads should achieve at least one-third of total reads, (c) percentage of nonredundant reads should be greater than 50% of the total mappable reads, and (d) upon visual inspection of the binding it is possible to observe distinctive punctuate or broad peaks (Ma and Wong 2011). Uniquely mappable reads are those reads that align to a single unique location in the genome while nonredundant reads are those that do not have exact sequence copy replicas. A high proportion of redundant reads indicate the artefactual source of the library as redundant reads in the ChIPseq dataset are more likely to be PCR artefact generated as part of NGS protocol (Ma and Wong 2011). In contrast, RNAseq read duplicates may come from increased levels of gene expression.

6.3.1.1 Evaluation of pilot INTS12 ChIPseq experiment

A total of 8,351,750 single-ended reads were generated in the pilot INTS12 ChIPseq experiment. Since the median quality score is above of 28, i.e. within the green area, for every base position the probability of error throughout the read is less than 0.2% (Figure 6.2). Therefore, unaltered raw reads were used for subsequent read mapping. The general consensus is to limit the manipulation of raw library data as much as possible (Williams et al. 2016). In order to quantify the proportion of unique reads in the library dataset, reads were aligned with the option of retaining only those that have single alignment location (see section 6.2.2.2).

Based on the reported statistics it appeared that 87% of mappable reads are unique (Figure 6.3, Table 5.1). In order to determine the nonredundant rate, alignment file was inputted to MACS (see section 2.8.2.1) and its output revealed that 90% of reads are nonredundant (Table 5.1). The pattern of binding was examined on the genome browser and revealed distinctive punctuate and broad binding near the

TSS of four key downregulated genes following INTS12 depletion (Figure 6.4). Thus, overall it is possible to say that pilot INTS12 ChIPseq experiment was successful according to all the criteria outlined in section 6.3.1. Therefore, full scale INTS12 ChIPseq experiment in two donor cells and input control was proceeded with.

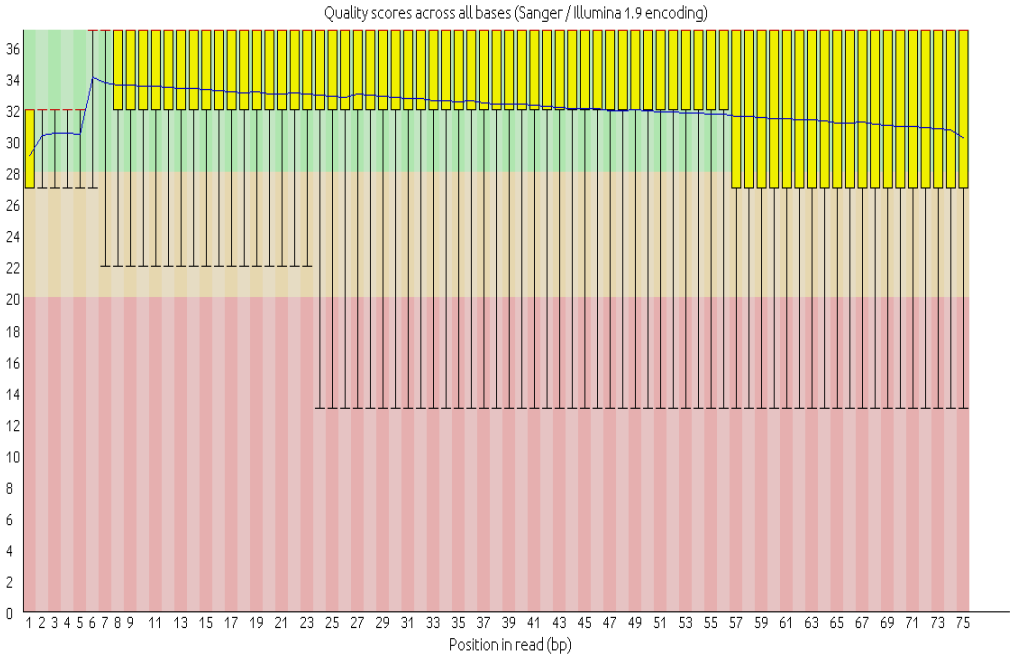
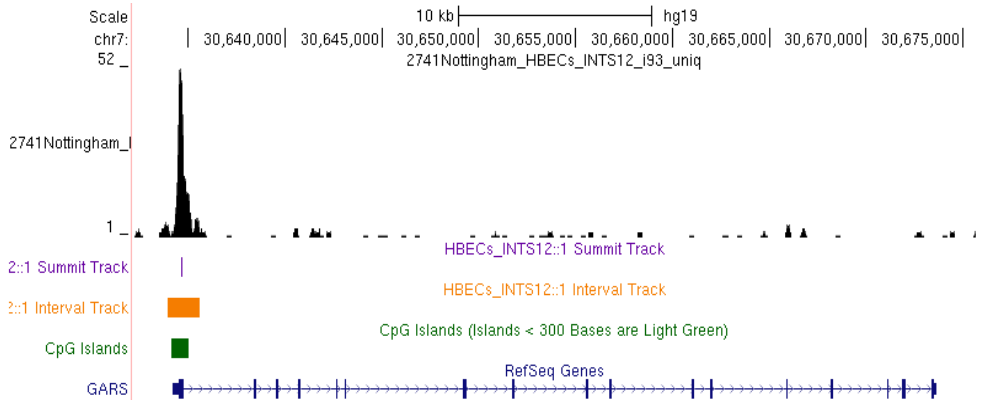


Figure 6.2: Quality scores of pilot INTS12 ChIPseq reads library.

```
# reads processed: 8351750
# reads with at least one reported alignment: 6028490 (72.18%)
# reads that failed to align: 1529442 (18.31%)
# reads with alignments suppressed due to -m: 793818 (9.50%)
```

Figure 6.3: Original bowtie-reported statistics about the ChIPseq alignment. Reads with alignments suppressed due to `-m` are those reads that would have aligned to more than one location otherwise.



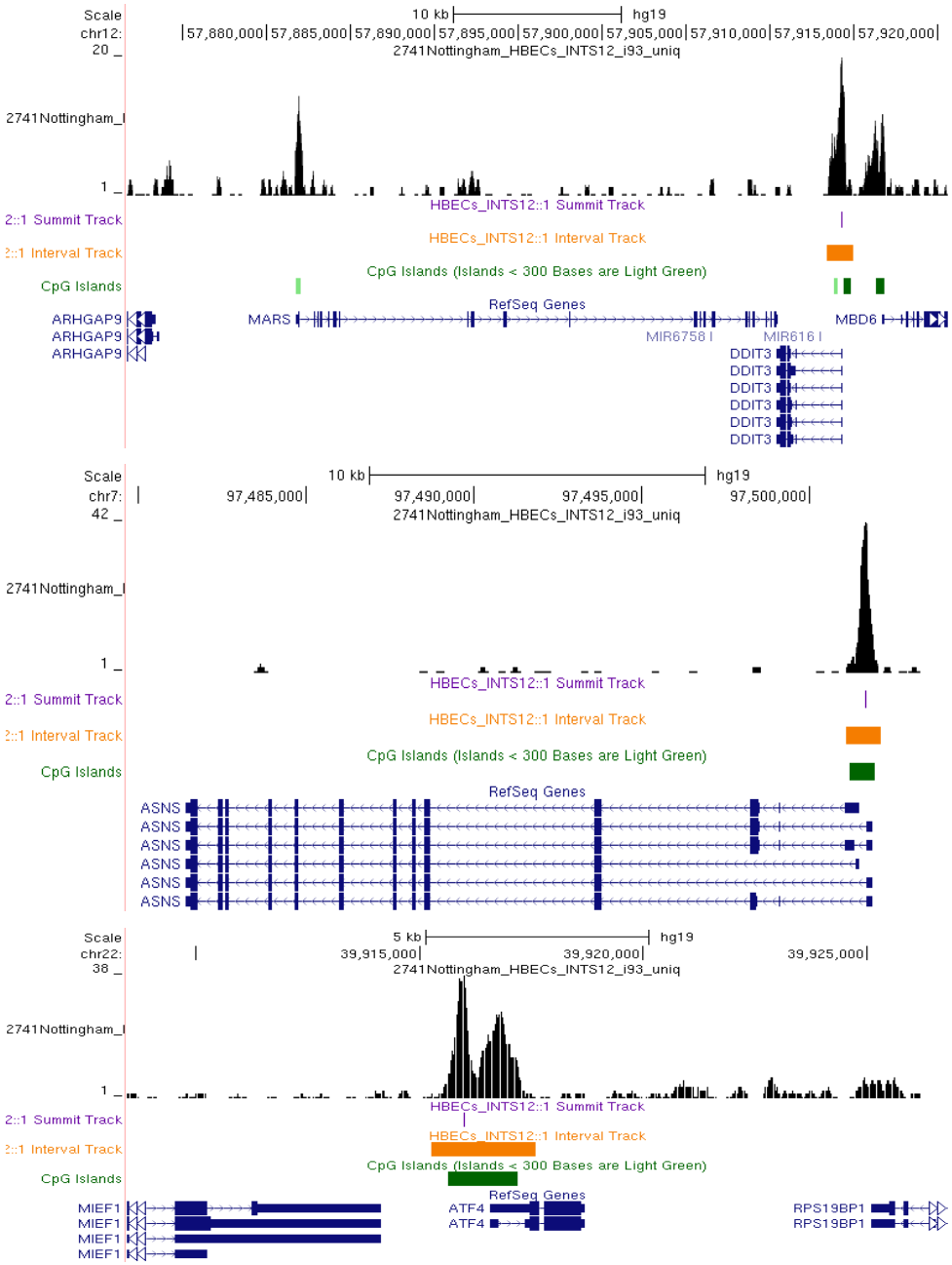


Figure 6.4: Genome browser views of INTS12 binding near differentially expressed genes *GARS*, *MARS*, *ASNS*, *ATF4* (shown in this order) reveals distinctive punctuate and broad pattern of binding.

Number of raw reads	Number of mappable reads and percent of total library	Number of unique reads and percent of total aligned	Number of nonredundant reads and percent of total aligned
8,351,750	6,028,490 (72%)	5,234,672 (87%)	5,427,524 (90%)

Table 6.1: Inferred relative proportion of uniquely mapped and nonredundant reads.

6.3.2 ChIPseq deep sequencing of INTS12

6.3.2.1 Pre and post alignment data quality control

Although the pilot INTS12 ChIPseq yielded 90% nonredundant and 87% uniquely mappable tags out of 8,351,750 raw reads, it is important to test these and other quality indices in the deep sequencing INTS12 ChIPseq as well. The ChIPseq procedure (see section 2.6.1) resulted in the generation of three raw FASTQ files: (1) FASTQ file for INTS12 ChIPseq from D195307 (2) FASTQ file for INTS12 ChIPseq from D7F3158 (3) FASTQ file for input control which is a sonicated, un-precipitated genomic DNA prepared from a pool of equal aliquots of the 2 donor samples. Negative sample is used for the determination of background signal noise. The quality of raw files was assessed and this revealed that reads had median quality scores above 28 across entire read length in all three files (Figure 6.5). As the probability of sequencing error is very low, just as was the case for the pilot library, untrimmed datasets were used for subsequent analyses.

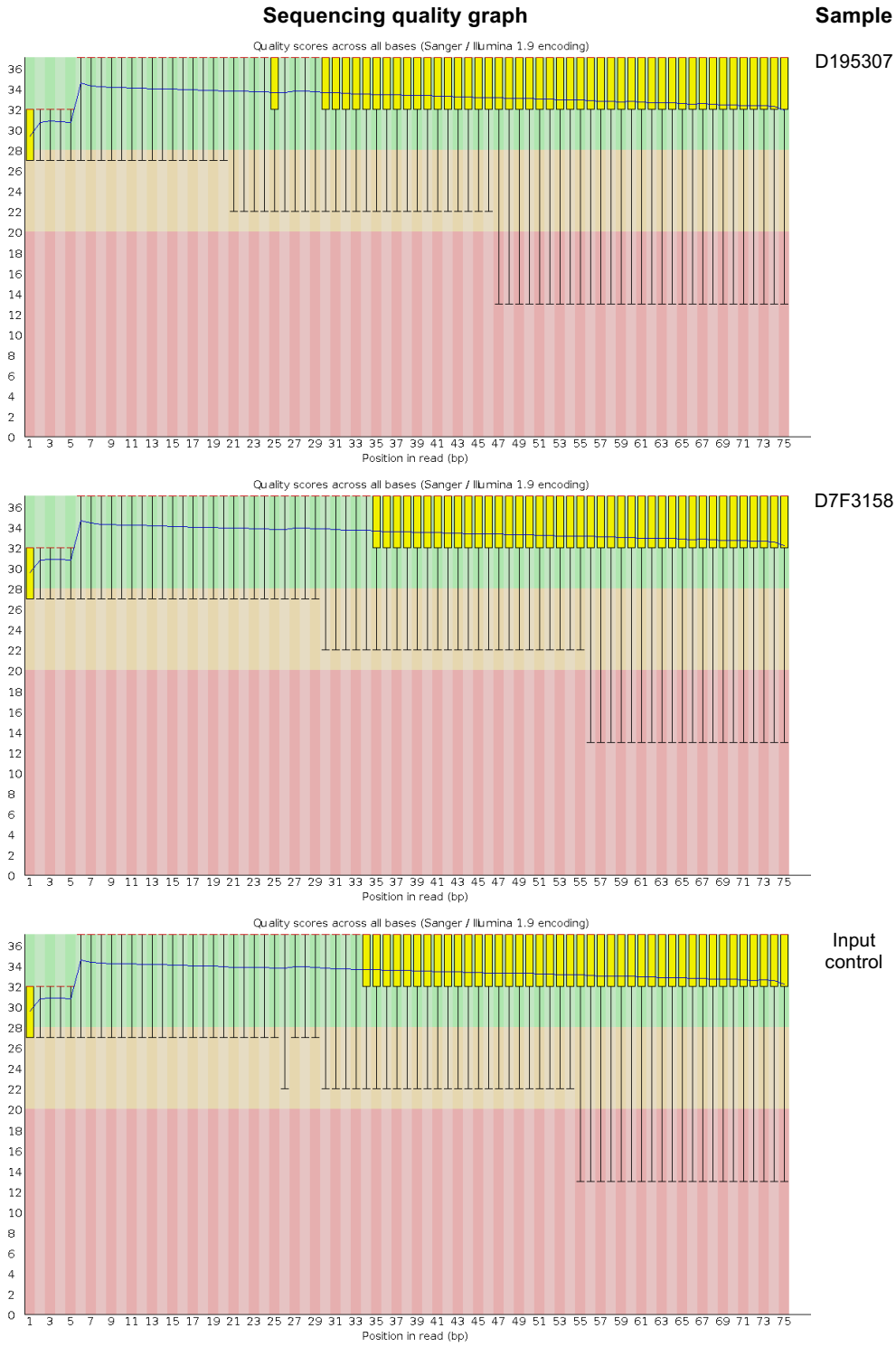


Figure 6.5: Sequencing quality graphs for INTS12 ChIPseq and input control samples

There was a total of 37,142,070; 47,776,470; 42,932,683 reads in the ChIPseq libraries of D195307, D7F3158 and input control samples respectively. Out of these, approximately 77-78% aligned to the genome and around 12% failed to align for all the samples (Table 6.2). Out of aligned reads 14.18%, 13.31%, and 5.46% were redundant duplicates

for D195307, D7F3158 and input respectively. The multi-hits rate was approximately 12-13% for all the samples. Based on these alignment statistics it is possible to say that ChIPseq samples are comparable for the downstream analyses. The similarity in multi-hits rate between input control and INTS12 ChIPseq samples suggest that reads aligning to multiple locations are not specific to INTS12 binding but rather are part of the background genome noise. However, there is a noticeable ~10% difference between the number of redundant reads of the two INTS12 ChIPseq samples and input control (Table 6.2).

As some differences in the library size was observed, tag counts were normalized to the library size in order to be able to compare signals between the samples. Overall, these quality indices are within the acceptable range outlined in section 6.3.1 rendering the samples suitable for downstream analyses. Importantly there is an equivalent depth between input and ChIPed samples, which is a crucial factor in ChIPseq experiment design (Sims et al. 2014).

Sample	Number of raw reads	Number of mappable reads and percent of total library	Number of unique reads and percent of total aligned	Number of nonredundant reads and percent of total aligned
D195307	37,142,070	29,085,938 (78.31%)	25,847,261 (88.87%)	23,820,417 (81.90%)
D7F3158	47,776,470	37,436,461 (78.36%)	33,086,311 (88.38%)	31,079,421 (83.02%)
Input control	42,932,683	33,205,103 (77.34%)	28,915,395 (87.08%)	30,860,044 (92.94%)

Table 6.2: Quality control indices of INTS12 ChIPseq samples and input control.

6.3.2.2 Coverage as additional quality control

Coverage depth is defined as the average number of times a nucleotide is represented in a collection of aligned sequence tags and is used to quantify how well the genome is interrogated in a particular sequencing library (Sims et al. 2014). For example, a mean x1 coverage is interpreted that on average one base is covered by at least one sequence tag for a given genomic interval.

The coverage of INTS12 and input control ChIPseq libraries were obtained with Qualimap (see section 2.8.2.1). In sample D195307, the coverage for autosomes varied between 0.62 and 1.08. Autosomes had coverage fluctuating close to 1. For chromosomes X, Y and mitochondrial, the coverage was 0.38, 0.24 and 73.07 respectively. In sample D7F3158, the coverage for autosomes varied between 0.81 and 1.38 and most of the autosomes had coverage well above 1. For chromosomes X, Y and mitochondrial the coverage was 0.5, 0.32 and 65.9 respectively. On the other hand, in input control sample the coverage for autosomes varied between 0.74 to 1.1 and most of the autosomes had coverage just below 1. For chromosomes X, Y and mitochondrial the coverage was 0.49, 0.37, and 526.61 respectively.
The coverage for D7F3158 appears slightly better than coverage for D195307. The ENCODE project's guidelines for ChIPseq experiments suggest that factor experiments should use at least 20 million reads in the mammalian cells (Landt et al. 2012). Thus given the facts that on average every nucleotide in the genome is interrogated with at least one tag, the pattern of coverage fluctuation is similar across the samples, and library size exceeding the ENCODE project recommendations it is possible to say that binding sites can be reliably inferred using this ChIPseq datasets (Landt et al. 2012, Ma and Wong 2011).

6.3.2.3 Characterization of INTS12 binding: peak calling and inter-donor reproducibility

INTS12 peak calling, i.e. identification of INTS12 binding sites, was performed using the second generation of MACS (Zhang et al. 2008). The BAM files used in peak calling had redundant tags removed (see section 2.8.2.1). ChIPed samples were compared to input control and peak calling was performed with a multiple comparisons corrected FDR less or equal to 0.05 considered significant. Larger dataset of the two submitted BAM files was scaled down towards the smaller dataset and the generated fragment pileup signal, i.e. ChIPseq signal, was normalized per million reads to account for differences in the size of the library as mentioned in section 6.3.2.1. The above analysis resulted in

identification of 70,772 and 51,377 INTS12 binding sites in donors D195307 and D7F3158 respectively (Figure 6.6).

In order to compare the biological reproducibility of identified binding sites, the respective hg19 coordinates of identified binding sites were tested for their intersection using on-line ChIPseek tool (Chen et al. 2014). From this analysis it appeared that 88% of D7F3158 peaks were also discovered in D195307. On the other hand 55% of D195307 peaks were also discovered in D7F3158. Therefore, it is possible to say D7F3158 peaks are largely a subset of D195307 peaks, although some of the binding observed in D7F3158 did not occur in D195307 (Figure 6.6).

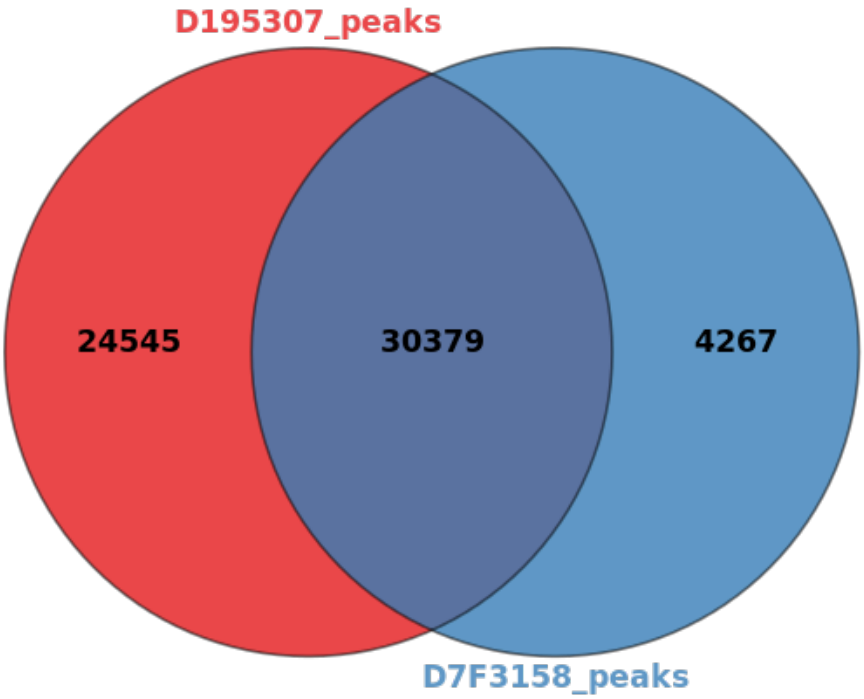


Figure 6.6: Genomic intersection of D195307 and D7F3158 binding sites reveals that 88% of D7F3158 peaks were also discovered in D195307 while 55% of D195307 peaks were also discovered in D7F3158.

It is important to note that although the total number of sequence tags in D7F3158 cells was greater than total number of sequence tags in D195307 cells, more peaks were identified in D195307 and their statistical significance was greater in D195307 in comparison to D7F3158 (Figure 6.7). This is clearly reflected in the strength of ChIPseq signal observed in the genome browser (Figure 6.8). An inter-donor

association test of ChIPseq signal was performed in active regions (see section 6.2.2). The use of active regions is necessary because the locations and lengths of binding sites are rarely exactly the same when comparing different samples. The association of the binding signal in these regions demonstrated a Pearson's correlation of 0.85 implying strong biological reproducibility (P<0.0001; Figure 6.9).

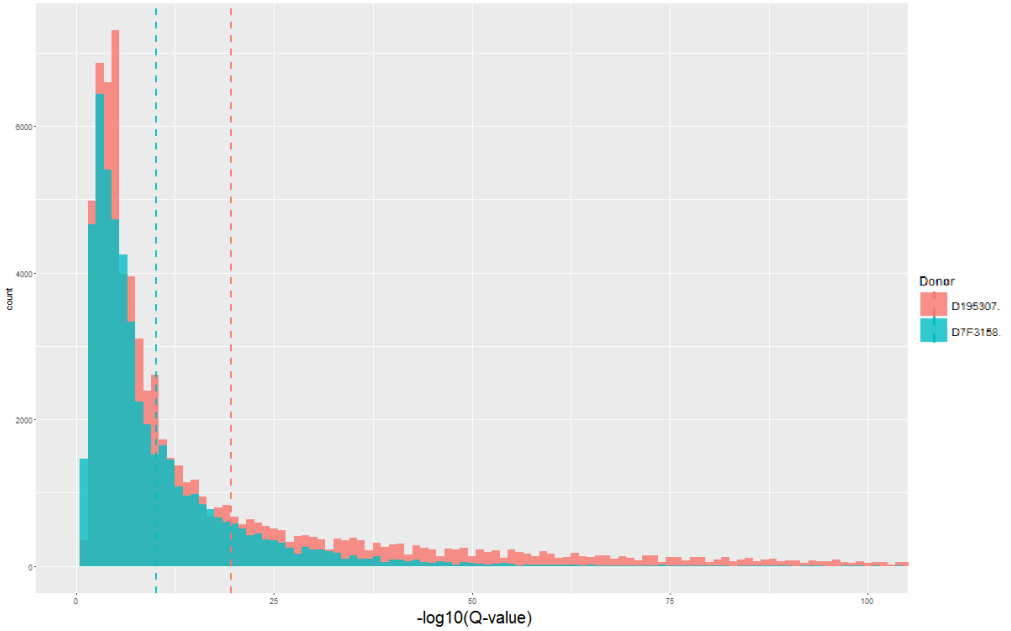


Figure 6.7: Comparison of the density of $-\log_{10}$(FDR) of enriched peak regions in D195307 and D7F3158 shown in red and pale blue respectively. The averages are denoted by dashed lines. Left side of the density bars show that overall greater number of sites in D195307 has as significant enrichment as in D7F3158. The average of statistical significance values is greater in D195307 (red line) than in D7F3158 (pale blue line) despite a greater number of sequence tags in D7F3158 library.

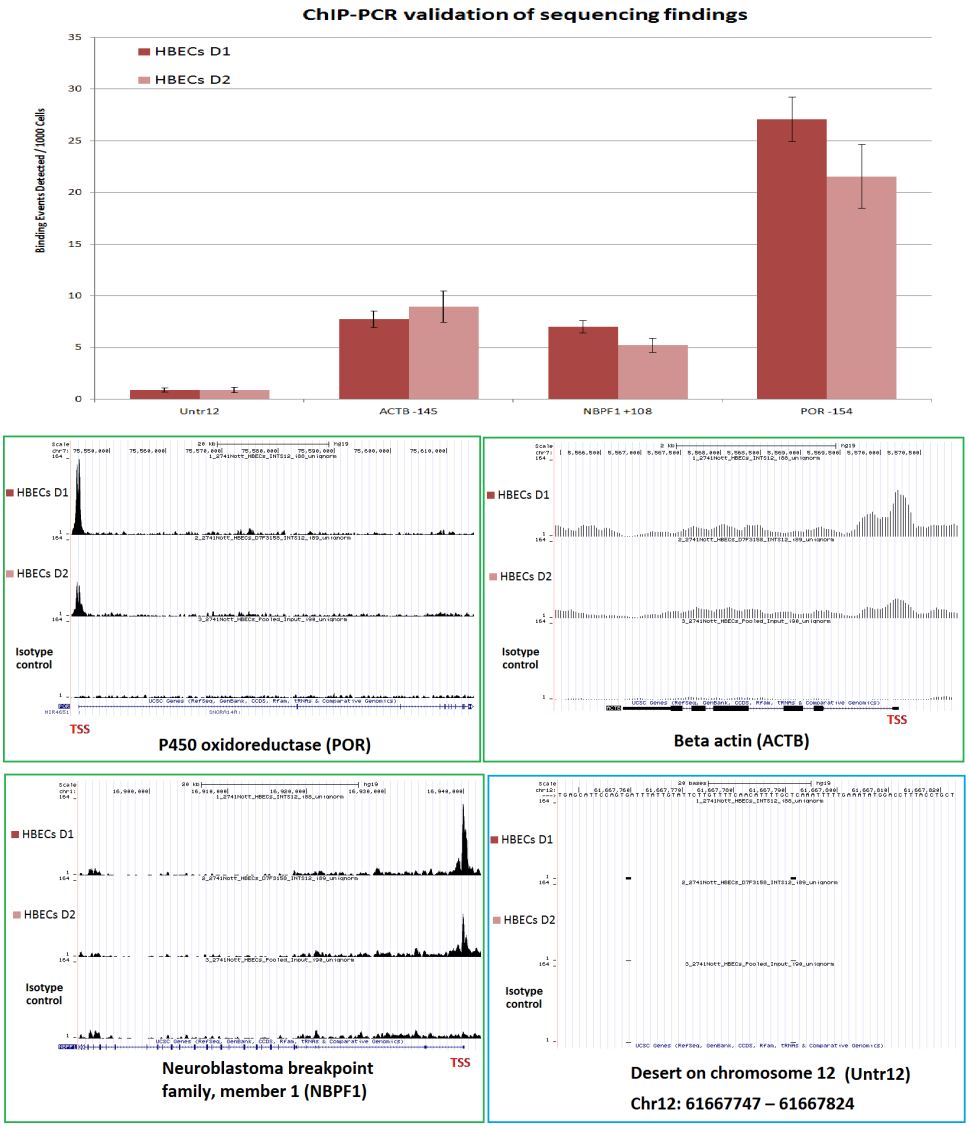


Figure 6.8: ChIP-PCR validation of ChIPseq findings. Three ChIPseq positive sites (POR, ACTB, NBPF1) shown in green boxes and one negative site (Untr12) shown in blue box were selected for ChIP-PCR testing to determine the number of binding events detected per thousand cells of donor D195307 (denoted D1) and donor D7F3158 (denoted D2). ChIP-PCR results corresponded well with ChIPseq data seen on the genome browser.

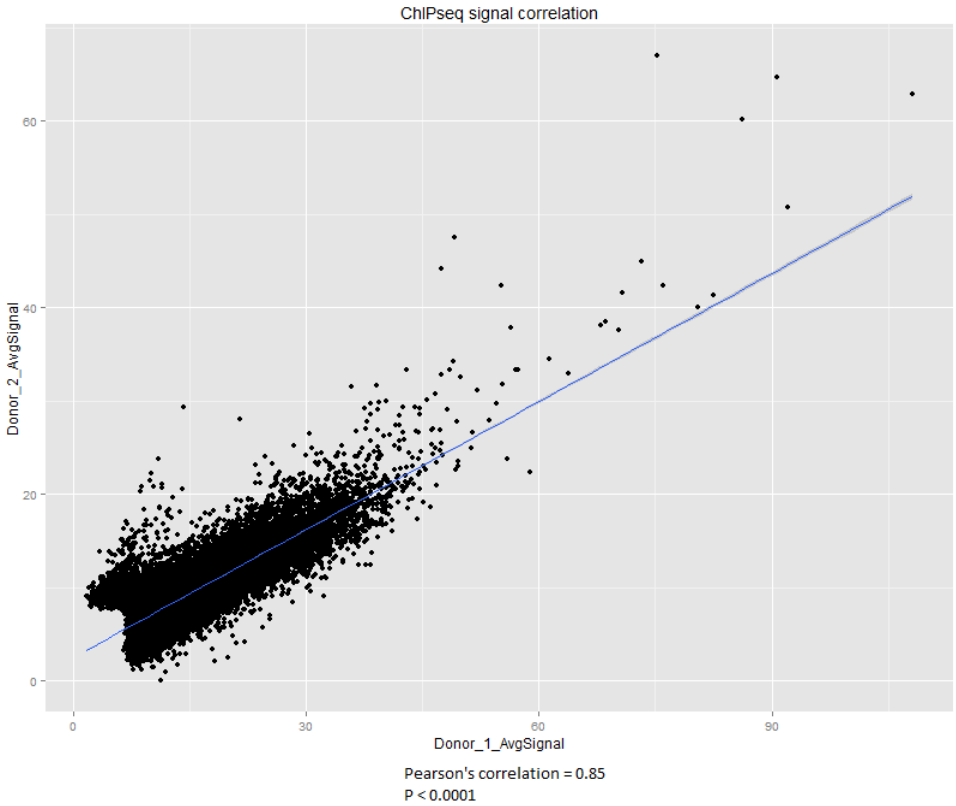


Figure 6.9: Biological reproducibility of genome-wide INTS12 binding ChIPseq signal. Correlation ChIPseq signals observed in D195307 (donor 1 on the figure) cells and D7F3158 (donor 2 on the figure) cells in active regions revealed a Pearson's correlation of 0.85 (P<0.0001).

6.3.2.4 ChIP-PCR validation of ChIPseq findings

In order to validate ChIPseq findings, three positive sites and one negative site were selected for technical validation by ChIP-PCR in each ChIP sample (see section 2.6.2). The number of binding events per thousand cells derived from ChIP-PCR corresponded well with the ChIPseq signal observed in genome browser validating our sequencing results (Figure 6.8).

6.3.2.5 Association of INTS12 binding sites with fixed elements of the genome

As INTS12's PHD motif domain shows homology with epigenetic regulators of gene expression (Table 3.3) and the fact that in *Drosophila* it is dispensable for snRNA processing (Chapter 4), the next aim was to investigate the interaction of identified INTS12 binding sites with fixed features (Marnetto et al. 2014), e.g. gene locations; and cell-type-

specific regulatory elements of the human genome (Jiang and Pugh, 2009; Madrigal and Krajewski, 2012), e.g. histone modification marks. The purpose of this analysis is to provide insights into its potential functions. For simplicity of data presentation, ChIPseq profile from D195307 was considered representative because of the greater number of binding sites detected (Figure 6.6), the improved quality of INTS12 binding observed in terms of strength of ChIP signal (Figure 6.8) as well as higher statistical significance of the enrichment (Figure 6.7).

A genome-wide analysis revealed that the three main fixed genomic features associated with INTS12 binding were intron, intergenic regions, and promoter (TSS±3000bp) which occupied 37%, 31% and 17% of the total binding sites respectively in donor D195307 (Figure 6.10). Thus contrary to the expectation, binding was not limited purely to the promoter regions. It is worth to point out that 75% of all promoter binding occurred proximally to TSS (deined as within TSS±1000bp). Indeed, a gene-centric analysis over a meta-gene body, i.e. collection of hg19 RefSeq genes, revealed INTS12 binding to be in close proximity to the TSS (Figure 6.11), mirroring the binding pattern of POLII in HeLa cells (Gardini et al. 2014). However, a widespread distribution of binding was observed in both donors (Figure 6.12).

The key question was whether this binding reflects gene-relevant regulatory roles or non-specific binding. A correlation analysis of INTS12 binding sites with number of annotated genes and with total nucleotide length of each chromosome was then performed in an attempt to answer this question. INTS12 binding in both donors correlated very well with the number of genes (Pearson correlations of 0.93 and 0.95 for donors D195307 and D7F3158 respectively; Figure 6.13). Correlations with chromosome length were notably weaker (Pearson's correlations of 0.73 and 0.63 for donors D195307 and D7F3158 respectively; Figure 6.13). Despite promoters being the least associated feature among the top three most enriched elements, because INTS12 correlated more with gene number than with chromosome length, it is possible to say that binding is more likely to reflect gene-relevant regulatory roles rather than mere random binding distribution between the genes.

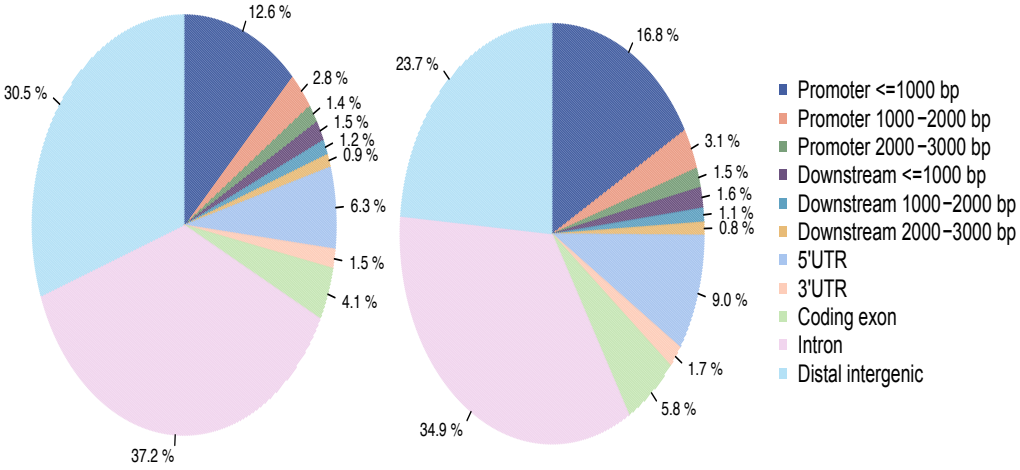


Figure 6.10: Percentage of INTS12 binding sites falling on the fixed annotated genomic features in the first (left) and second (right) donor.

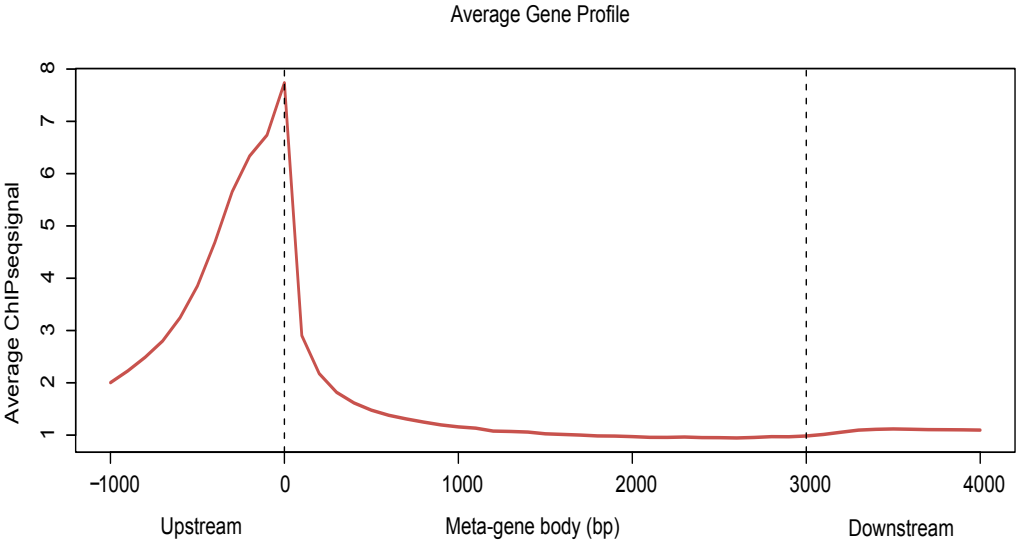


Figure 6.11: Gene-centric analysis of INTS12 binding in the first donor across the gene bodies of all the known human genes shows clear localization near the TSS.

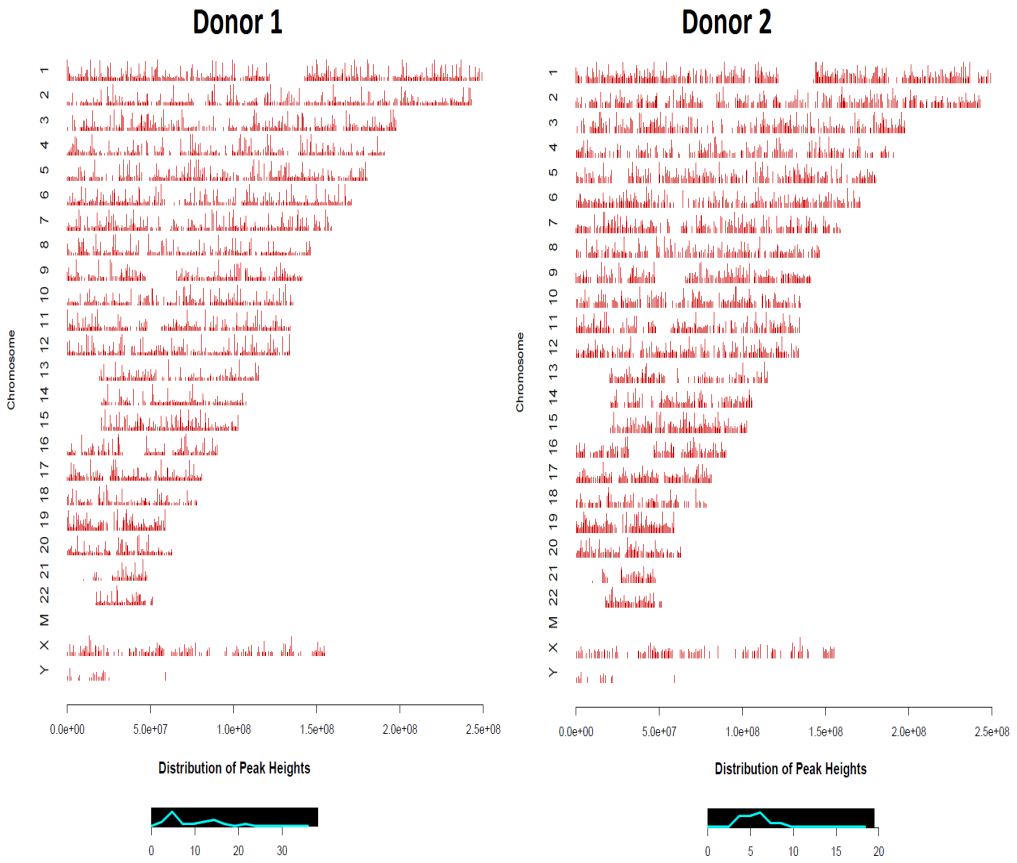


Figure 6.12: INTS12 ChIPseq regions and peaks over the human genome in first (D195307) and second (D7F3158) donor cells. Although there is a greater spread of INTS12 peak heights in donor D195307 than in donor D7F3158, suggesting stronger binding in that donor, the overall distribution looks reproducible between the donors.

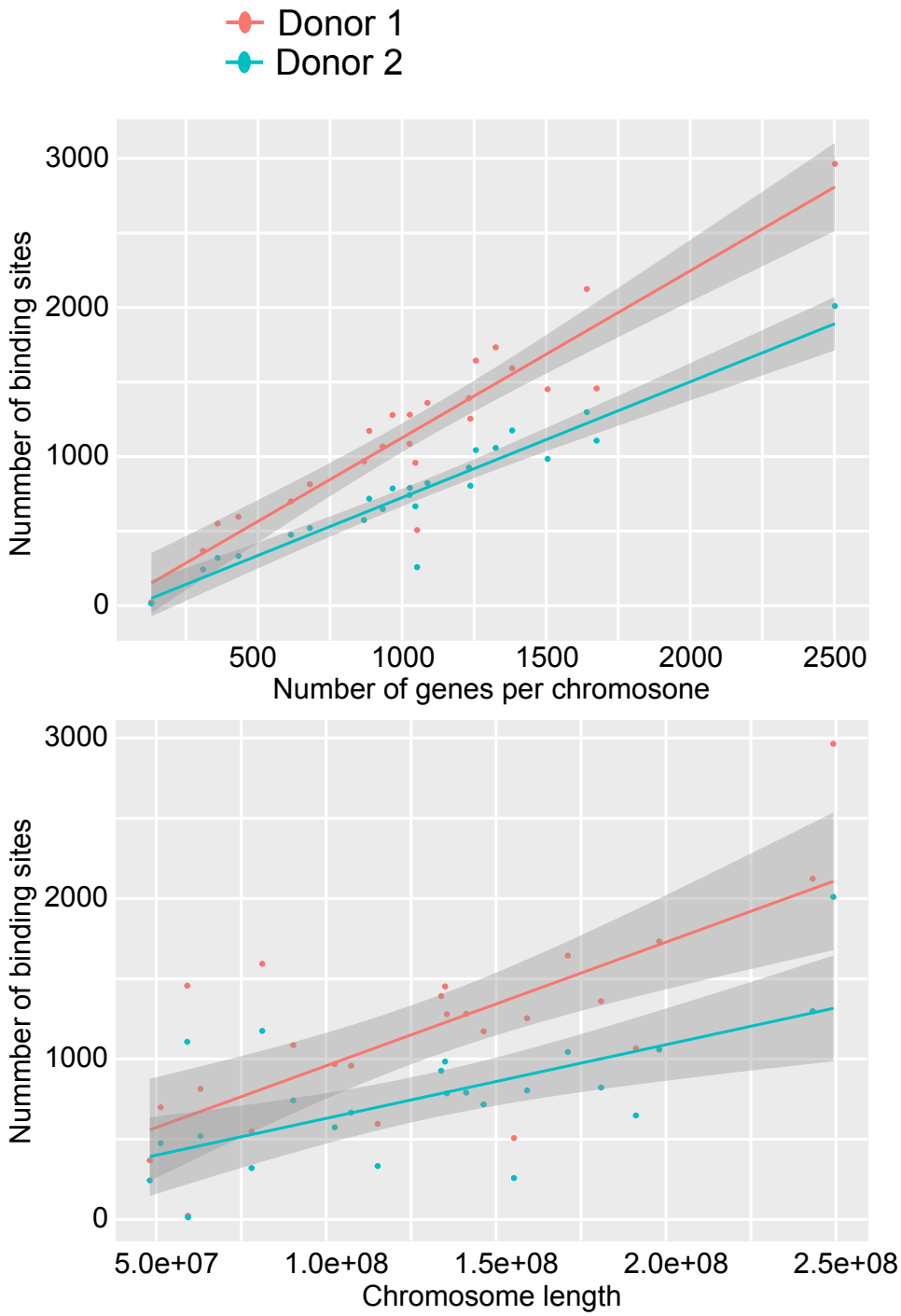


Figure 6.13: Relationship of INTS12 binding to the gene number per each chromosome and chromosome length. Analysis of binding versus number of genes revealed Pearson's correlations of 0.93 and 0.95 in the first (D195307) and second (D7F3158) donor respectively. Instead, correlations of binding sites and chromosome length are weaker being 0.73 and 0.63 for the first and second donor respectively.

6.3.2.5.1 The average pattern of INTS12 binding varies between different gene classes

Since the believed canonical function of INTS12 is processing of snRNAs (Baillat et al. 2005, Chen et al. 2013) the initial prediction was that it would be primarily enriched over the bodies of snRNA genes and less so for the other gene classes. The widespread distribution of INTS12 binding (Figure 6.12), in combination with the overall minor snRNA processing impairment due to INTS12 knockdown (Chapter 4), implied a potential regulatory role in expression of other genes and hence INTS12 binding enrichment over other gene classes was examined.

Among protein coding, snRNA, small nucleolar RNA (snoRNA), microRNA, and long intergenic RNA (lincRNA) genes; protein coding and snRNA genes show the highest enrichment with different patterns of binding over these two main gene classes (Figure 6.14). For protein coding genes the peak binding is proximal to the TSS while for snRNA genes the binding is enriched downstream to the TES suggesting there may be distinct functional activities of INTS12 depending on the class of the genes where INTS12 binding occurs. Enrichment near TES for snRNA genes is in agreement with its putative role as part of snRNA processing machinery which occurs simultaneously to nascent transcription of 3'box element (Baillat et al. 2005).

However, 95% confidence intervals, and therefore error margins, surrounding the mean enrichment is much greater over the snRNA rather than protein coding genes, probably due to the higher number of protein-coding genes to which INTS12 binds in comparison to snRNA genes. This observation further supports the hypothesis that INTS12 has additional roles beyond snRNA processing and is in agreement with the observed molecular signatures described in Chapter 5. Interestingly, in contrast to snRNA loci, for long intergenic RNA genes the peak binding is near the TSS as it is the case for the protein coding genes.

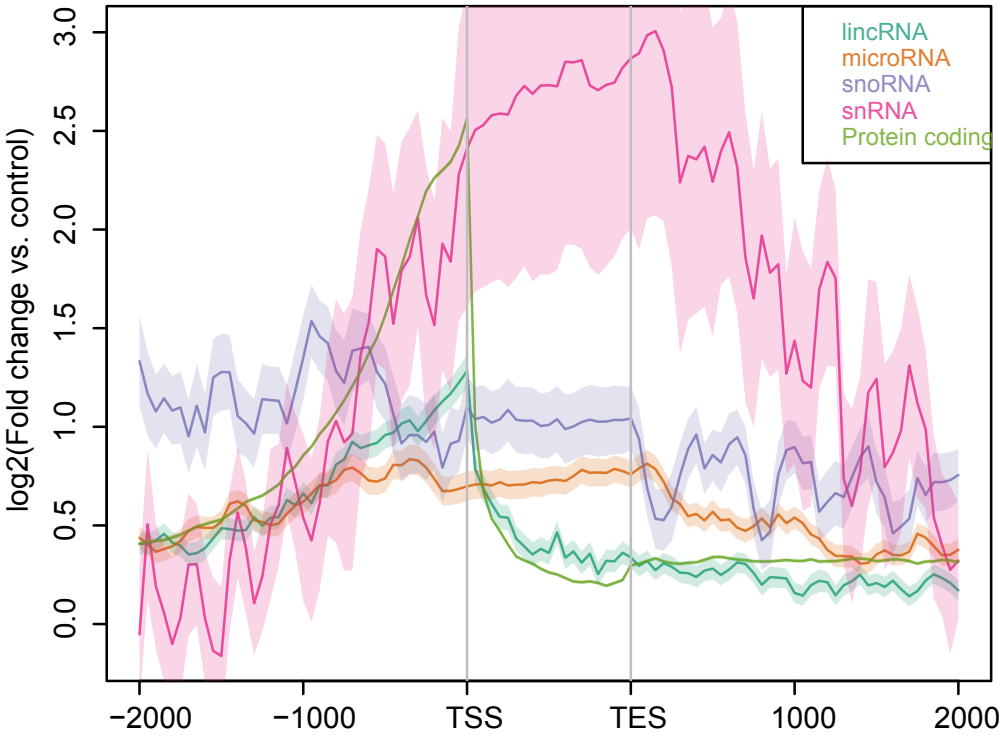


Figure 6.14: Comparison of INTS12 binding in first donor across the gene bodies of protein coding, snRNA, snoRNA, lincRNA, and microRNA genes. The pattern of localization differs between the gene classes suggesting distinct INTS12 activities depending on the type of the transcribed gene. The top two classes with highest enrichment were protein coding and snRNA genes which had peak summits near TSS and TES for the former and latter group respectively. Coloured shadows around the average plots indicate the 95% confidence interval of the plot.

6.3.2.6 Association of INTS12 binding with specific regulatory elements

In addition to investigating association with fixed genomic features, the other aim was to test INTS12's binding localizations relative to specific regulatory elements identified in bronchial epithelial cells. As mentioned in Chapter 1, this kind of element is much more mobile and variable depending on the cell type. Thus, in order to draw valid conclusions regarding the co-localization of these elements and INTS12 binding, HBECs-relevant datasets ought to be used because INTS12 binding was profiled in this cell type. Although bioinformatic searches indicate INTS12's PHD motif domain to be a candidate nucleosomal histone tail binding protein (Chapter 3) no direct experimental evidence exists. Moreover, the weight of evidence argues for a general deficiency of histones within snRNA genes and within the snRNA promoter (Pavelitz

et al. 2008; Egloff et al. 2009). Consequently, the INTS12 PHD finger is unlikely to be coupling INTScom to the snRNA promoter via histone binding. To understand to potential importance of epigenetic histone modifications for INTS12 biology, the intersection of its representative binding with reference localizations of H3K4me3, H3K36me3, and H3K27me3 histone modifications using per-chromosome randomization test (Gel et al. 2016) was investigated.

On a genome-wide scale, H3K4me3 had the highest enrichment with 58% of INTS12 binding co-localizing (Z-score=348; Figure 6.15). H3K36me3 co-localized with 21% of INTS12 binding sites (Z-score=13; Figure 6.15). H3K27me3 overlapped with 4% of INTS12 binding sites which was less than expected by chance (Z-score=-11; Figure 6.15). H3K4me3 is associated with the promoters of actively transcribed genes, while H3K36me3 is enriched in the body of such genes (Bannister and Kouzarides, 2011). On the other hand, H3K27me3 marks silenced regions (Gibney and Nolan, 2010). Therefore, these data provide supporting evidence of recruitment of INTS12 into loci epigenetically marked as transcriptionally active, which may be modulated via its binding to histone 3 and recognition of H3K4me3 modification. This is further supported by the observation that 96% of INTS12 binding occurred in the vicinity of DNaseI accessible chromatin signature (Z-score=223; Figure 6.15) (Thurman et al. 2014).

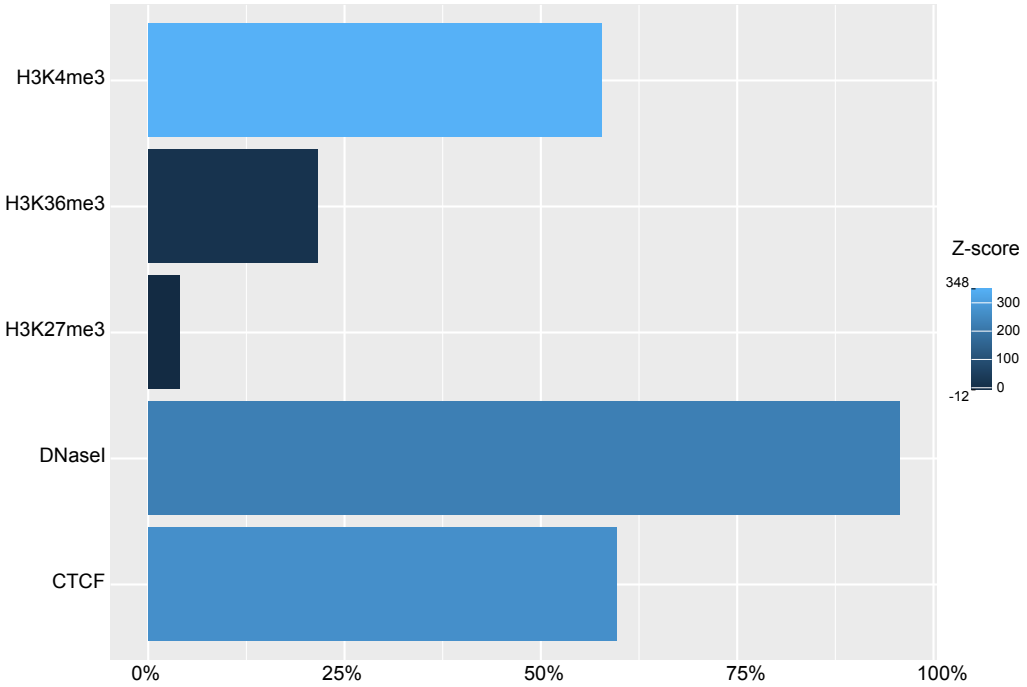


Figure 6.15: Percent of total INTS12 binding sites overlapping with HBEC-specific regulatory elements. Data from the first donor is shown as a representative of the two donors tested. Colour indicates the Z-score of the distance between the observed overlap and the mean of distribution of random overlap permutations. Negative Z-score implies that the observed overlap is less than expected by chance. Higher Z-score implies larger distance to the mean of distribution in a randomization test. Within P<0.05 the minimum Z-score in random permutation walk is 8, 6, 4, 7 and 3 for H3K4me3, H3K36me3, H3K27me3, DNaseI, and CTCF respectively. The two features most prominently localizing with INTS12 are H3K4me3 (Z-score=348) and DNaseI (Z-score=223) both marking transcriptionally active regions.

Interestingly, INTS12 peak regions, defined as ±500bp in both directions from the peak summit, show stronger conservation when compared with neighbouring regions, defined as ±2500bp in both directions from the peak summit (Figure 6.16). INTS12 also overlapped with CTCF insulator protein among 60% of its binding sites (Z-score=264; Figure 6.15), and INTS12 binding sites appeared more evolutionary conserved than CTCF binding sites (Figure 6.16).

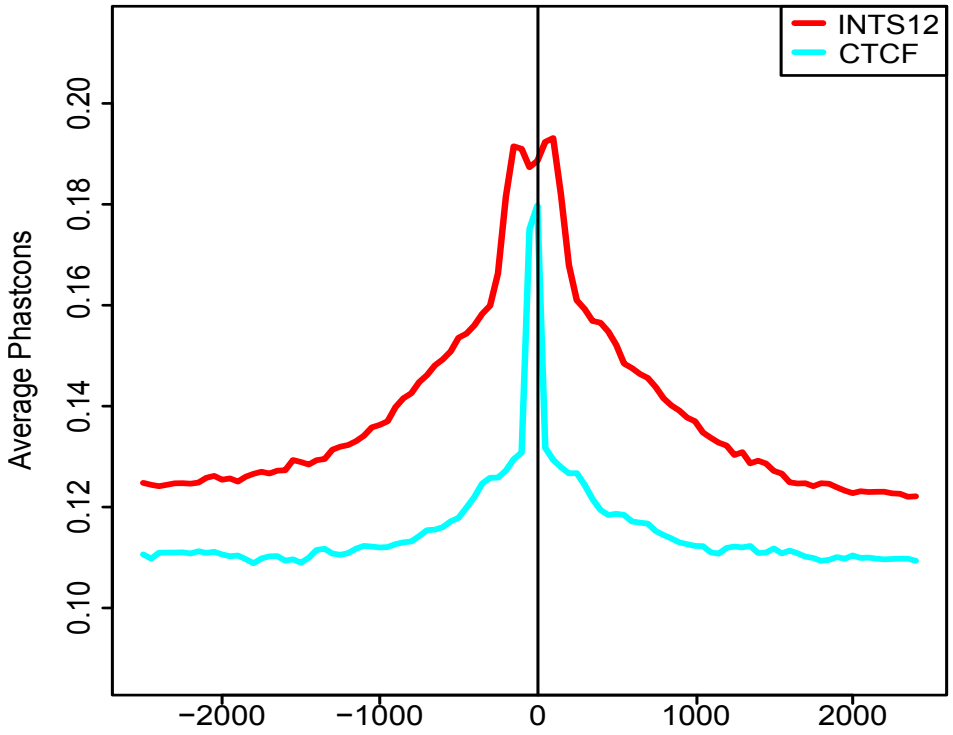


Figure 6.16: Evolutionary conservation of INTS12 binding sites in vertebrates. The figure is showing the phastocons score derived from multiple sequence alignment of vertebrate genomes, across the binding sites of INTS12 (red) and CTCF protein (blue). INTS12 sites appear to have higher conservation in close proximity to the peak summit approx. 250bp in each direction, in comparison to the more distal locations. In contrast CTCF binding locations are more conserved >80bp in each direction. Also, overall INTS12 binding sites are more conserved than the binding of CTCF.

In addition to testing the relationship between cross-binding of INTS12 and specific mobile element sites (Figure 6.15) the overall correlation of their ChIPseq signals on a genome-wide scale was also examined (Figure 6.16). In agreement with initial observations, INTS12 signal most strongly correlated with accessible chromatin (Pearson correlation of 0.83) followed by H3K4me3 (Pearson correlation of 0.74). H3K36me3, CTCF and H3K27me3 had weaker correlations of 0.61, 0.58, and 0.06 respectively. As gene-centric analysis revealed INTS12 binding to be enriched near the TSS (Figure 6.11) the correlation of ChIPseq signals at the promoters defined as TSS±3000bp was also examined. In this analysis the strongest correlation was observed between INTS12 and H3K4me3 (Pearson correlation of 0.8) outweighing the correlation between INTS12 and DNaseI (Pearson correlation of 0.73). Correlations

with H3K36me3, CTCF and H3K27me3 were weak at the promoters being 0.3, 0.3, and -0.29 respectively.

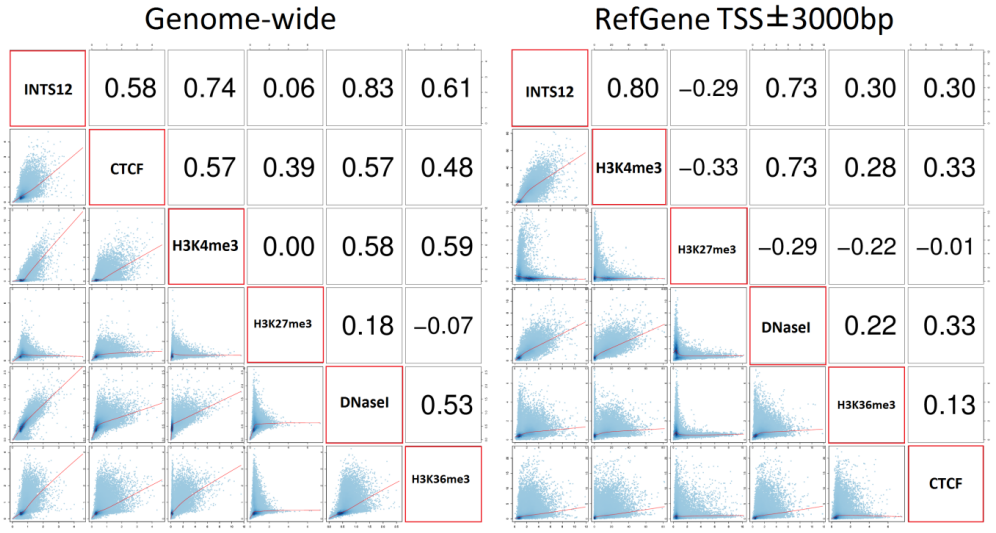


Figure 6.17: Cross-correlations of INTS12 and HBEC specific regulatory elements ChIPseq signals on a genome-wide scale and in the promoter regions (TSS±3000bp). On genome-wide scale the strongest correlation is with DNaseI, while in promoter regions the strongest correlation is with H3K4me3 mark.

6.3.2.7 INTS12 binding in the light of literature data

It has been suggested that a role of INTScom in POLII pause release is conserved between human and *Drosophila* (Gardini et al. 2014), however this hypothesis has not been explored experimentally for INTS12. If it is the case that this function is conserved, then a human homologue of the fly gene displaying the pause release phenomenon ought to show enrichment of INTS12 binding near its TSS. HSP70Aa is a classical heat shock response fly gene displaying the POLII pause release phenomenon and INTS12 was shown not only to bind in close proximity to its TSS but also to have increased binding following heat shock treatment (Chapter 1). It is thus worth asking whether INTS12 binding near human orthologue of HSP70Aa is similarly conserved.

As it can be seen on Figure 6.18, human INTS12 appears to be highly enriched near the TSS of HSPA1A, which is the human orthologue of Drosophila's HSP70Aa. This can be observed in both donors but not in the input control (Figure 6.18). This finding gives preliminary support to the above mentioned conservation hypothesis. On the other hand, the first study that uncovered INTScom recruitment to snRNA but not

protein-coding genes (Baillat et al. 2005) differs in this respect from the data presented herein. In addition to the fact that on a genome-wide scale INTS12 binding shows at least equal, if not improved, enrichment near TSS of protein coding genes relative to TES of snRNA genes (Figure 6.14), a closer look-up into the binding near GAPDH was performed. GAPDH was used by Baillat et al. as a control protein-coding gene lacking INTScom binding, in contrast to U1 and U2 snRNA genes which were highly enriched for INTScom binding (Baillat et al. 2005). Therefore, INTS12 binding near GAPDH was examined. High enrichment near the TSS and consecutive broad peaks towards the 5'end from the gene's TSS were detected (Figure 6.19). Interestingly there is a noticeable binding within the GAPDH gene body.

The discrepancy between these and Baillat et al. data may have arisen from differences between the HeLa and HBEC model systems, or lower immunoprecipitation yield in Baillat et al. study which was not sufficient to detect INTS12 binding in proximity to protein coding genes. Alternatively, non-INTS12 INTScom subunits that were precipitated in Baillat et al. study do not correspond to INTS12 binding (Baillat et al. 2005).

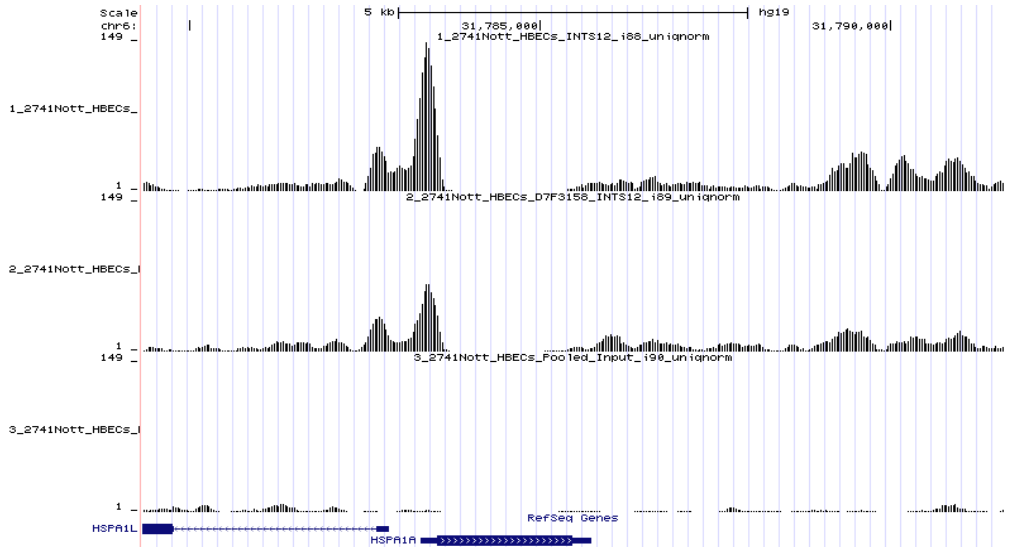


Figure 6.18: INTS12 binding near *HSPA1A* recapitulates the binding observed in fly's orthologous *HSP70Aa* giving preliminary credit to the hypothesis of conservation of INTS12 functional role in POLII pause release.

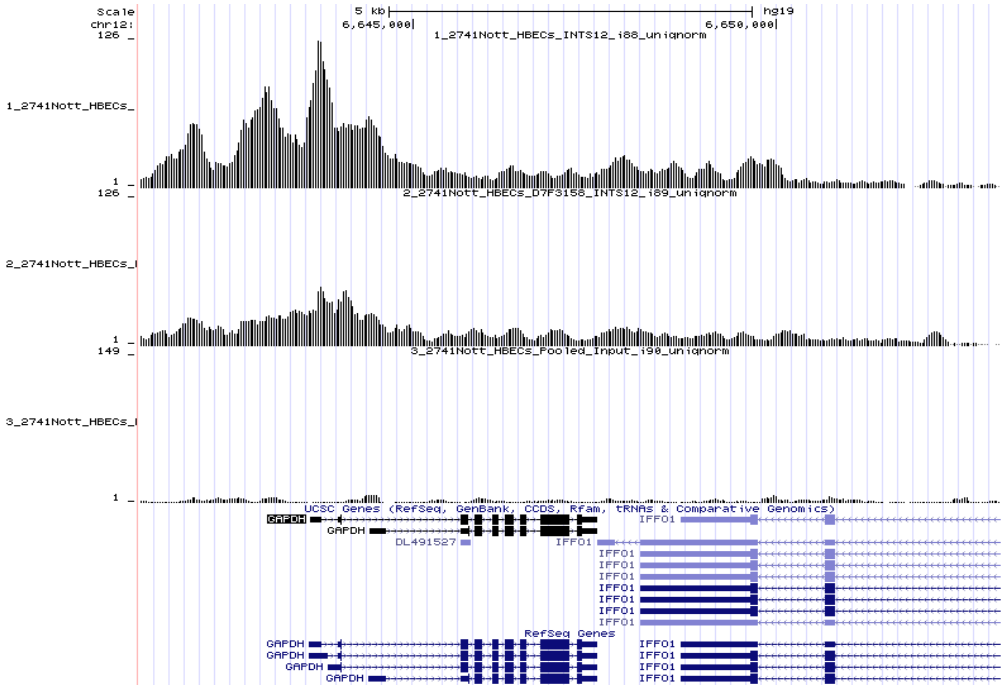


Figure 6.19: INTS12 binding near *GAPDH* disagrees with what was reported for INTScom binding near the same gene in HeLa model.

6.3.2.8 Combination of ChIPseq and RNAseq reveals INTS12 regulome

As INTS12 depletion resulted in downregulation of key protein synthesis pathways (Chapter 5) which appeared to alter cell phenotype through repression of cellular translation and proliferation (Chapter 5), a question was raised asking whether these effects were exerted due to direct regulation or were a secondary epiphenomenon (Figure 6.1). In order to identify INTS12's regulome, i.e. the set of genes directly regulated by INTS12, RNAseq expression data was combined with ChIPseq binding data.

Because INTS12 showed the highest enrichment with DNaseI and H3K4me3 sites, both marking active transcription, and poor correlation with H3K27me3, which marks silenced loci, the key query is whether these observations agree with gene expression signatures detected in basal HBECs. Indeed, INTS12 appeared to have approx. 8-fold higher enrichment of binding near TSS of expressed genes, i.e. having greater than zero FPKM in at least one biological replicate, than genes which are not expressed, i.e. having zero FPKM in three biological replicates (Figure 6.20). Using these criteria there were 14766 and 5129 protein

coding genes which were switched on and off respectively. Importantly, the magnitude of binding corresponded well with the degree of gene expression (Figure 6.21).

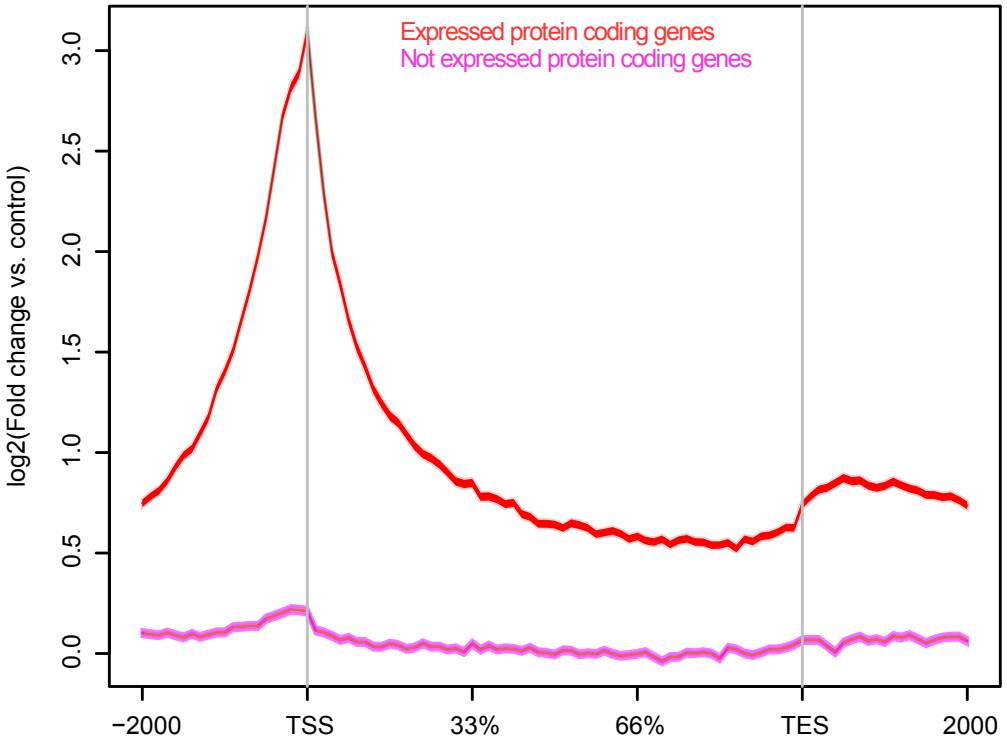


Figure 6.20: INTS12 binding is enriched at transcriptionally active genes as log_2 of binding signal versus input control. The ChIPseq signal was compared between the sets of expressed and silenced genes in basal HBECs of D195307.

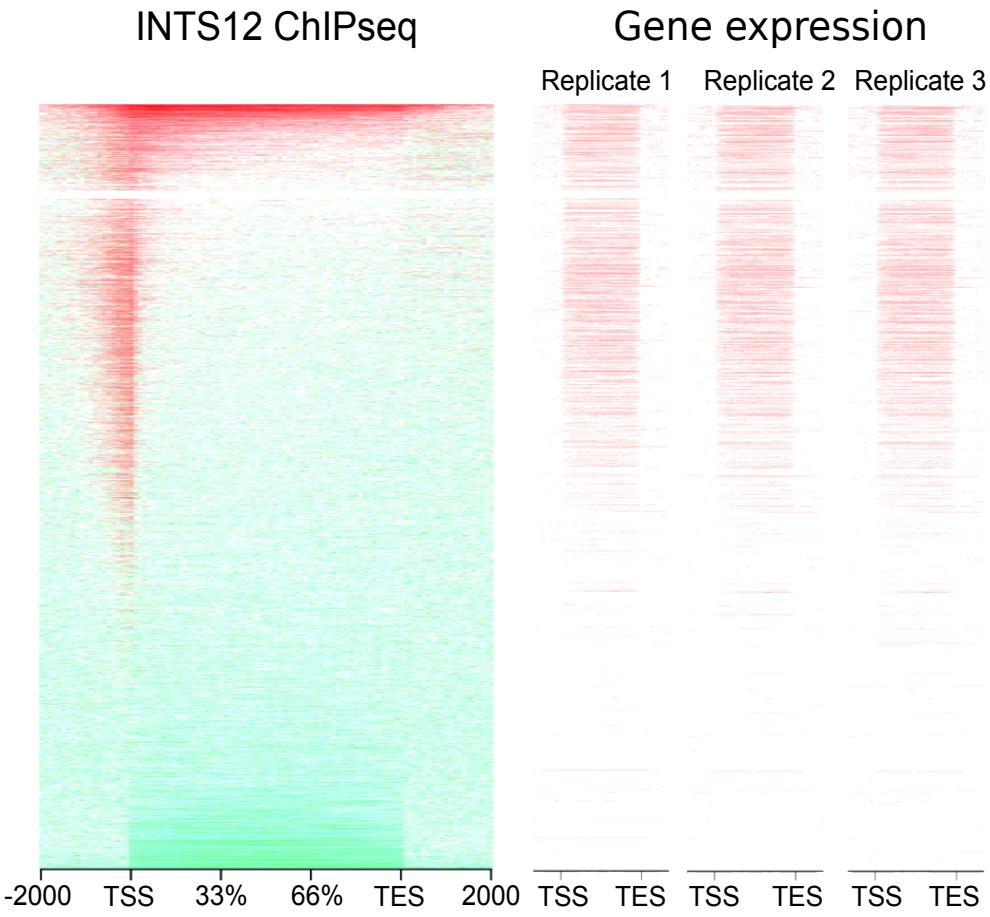


Figure 6.21: Comparison of INTS12 binding vs. corresponding gene expression in basal HBECs. Genes were ordered based on the level of INTS12 ChIPseq signal in D195307. The same sorted gene list was used to evaluate their transcription in basal un-transfected HBECs of the same donor where red colour indicates higher expression derived from read counts on the corresponding gene bodies.

To further investigate this, genes were divided into three groups, namely those which were upregulated, downregulated or not differentially expressed following INTS12 knockdown. Sustained depletion dataset using D195307 was leveraged to define these groups, with genes significantly (FDR<0.05) dysregulated in D-siRNA A and C treatments included in the classification but excluding genes altered by scrambled D-siRNA. There were 868, 1248, and 48156 genes in the upregulated, downregulated and those with no evidence of differential expression groups respectively. Crucially, the ChIPseq dataset that was overlaid upon the RNAseq dataset was from the representative D195307 cells and therefore both omic datasets were generated using the same donor.

On average there was 6-fold, 8-fold and 1.6-fold enrichment of INTS12 binding above genome background for upregulated, downregulated and not differentially expressed genes respectively (Figure 6.22). Thus of the total number of downregulated and upregulated genes 92% and 85% of genes show evidence of INTS12 binding near their TSSs, while only 23% of genes that had no evidence of differential expression showed this localization (Figure 6.23).

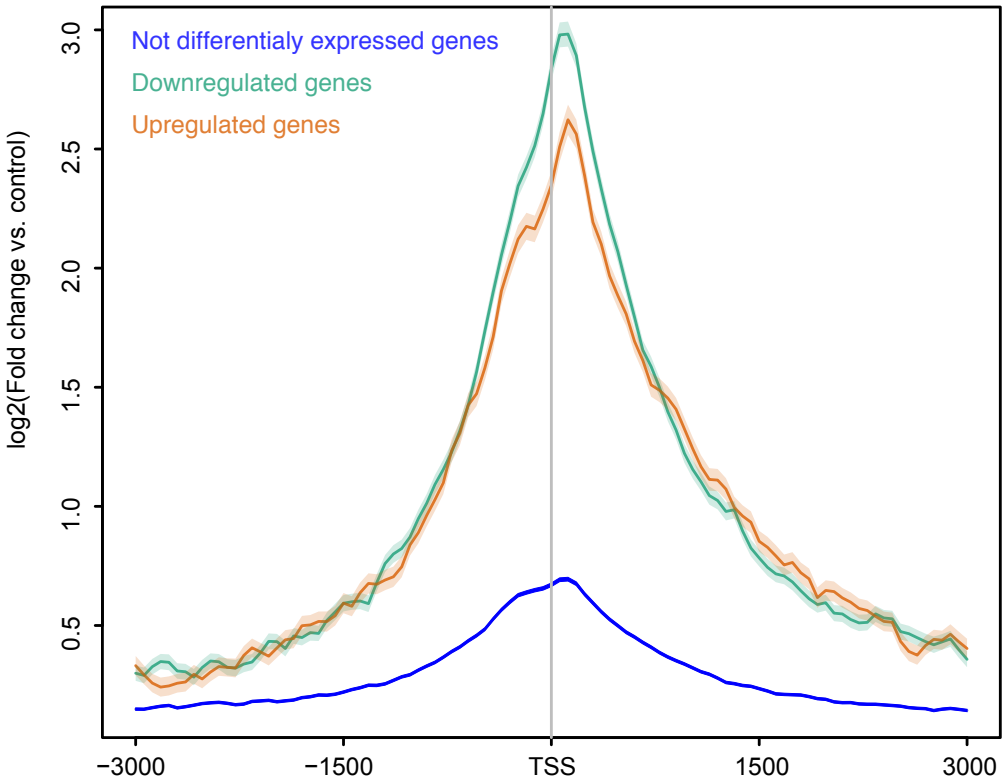


Figure 6.22: Average INTS12 binding profile for differentially expressed genes and genes with no evidence of differential expression following INTS12 depletion in RNAseq. Differentially expressed genes show higher enrichment of binding than not differentially expressed genes, with observable binding bias for downregulated genes explaining more robust effects observed upon downregulated pathways.

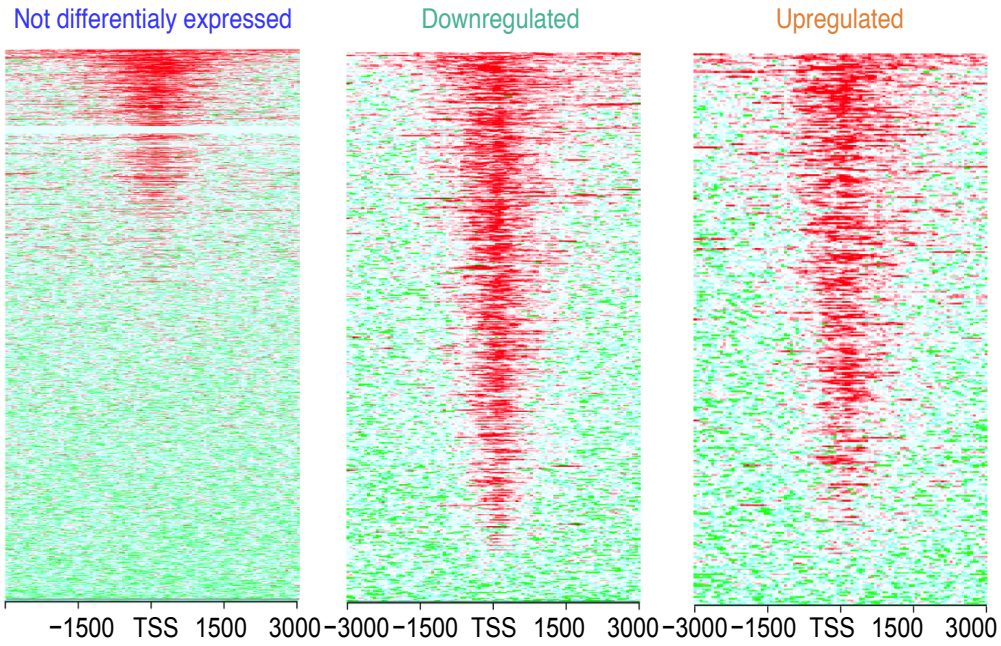


Figure 6.23: Heatmap of INTS12 binding for differentially expressed genes and genes with no evidence of differential expression following INTS12 depletion. Out of total number of downregulated and upregulated genes, 92% and 85% of genes show evidence of INTS12 binding. In comparison, 23% of genes with no evidence of differential expression show evidence of INTS12 binding.

To provide validation for these findings the regulatory potential of INTS12 for each gene based on evidence of binding and significance of differential expression following depletion by either D-siRNA A or D-siRNA C was calculated. The ranked list of genes based on their regulation versus cumulative fraction of genes for a given regulatory potential score was plotted (Figure 6.24; Wang et al. 2013). The sets of upregulated and downregulated genes had significantly higher regulatory potential scores than 'static' genes ($P<0.001$) indicating that genes with evidence of near promoter (TSS±1000bp) INTS12 binding were contributing to altered expression following INTS12 depletion. However, a stronger bias for downregulated genes was observed in both average binding plot (Figure 6.23) and regulatory potential analysis (Figure 6.24), explaining larger effect sizes in gene expression changes and greater number of dysregulated pathways meeting the statistical significance observed among downregulated pathways in the pathway analysis (Chapter 5). Moreover INTS12 was seen to bind to TSS of 4 key genes selected for technical and biological validation of their

downregulation following INTS12 depletion (Figure 6.4) implying INTS12 contribution to their altered expression.

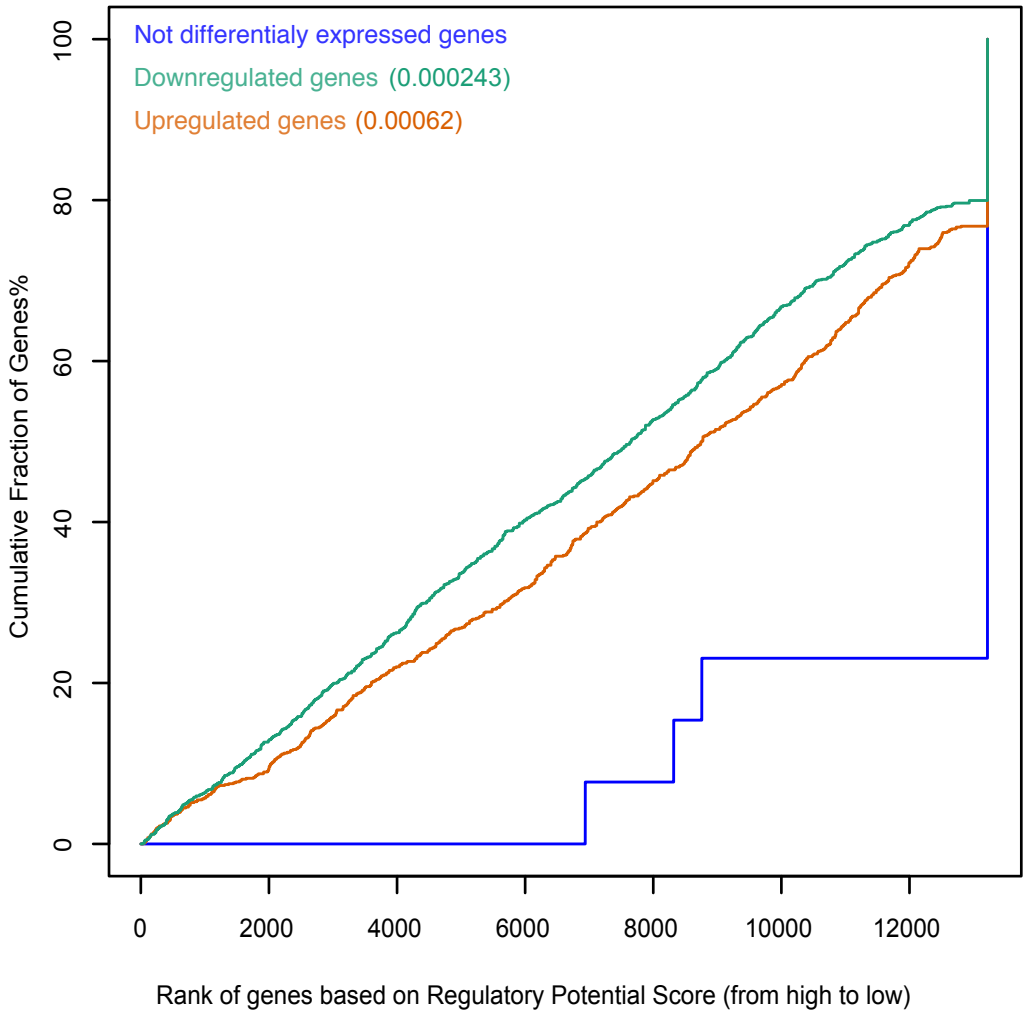


Figure 6.24: Prediction of the activating and repressive function of INTS12. The cumulative fraction of genes is plotted against the regulatory potential, based on significance of representative D-siRNA A differential expression and ChIPseq evidence of binding near genes' TSS. INTS12 depletion was equally likely to induce or suppress gene expression in Kolmogorov-Smirnov test but >90% of downregulated genes had a higher regulatory potential than upregulated genes explaining the more robust effects observed on downregulated pathways.

6.3.2.9 Motif enrichment and its distribution analysis

Having identified INTS12's regulome the next objective was to test for potential DNA recognition signatures. *De novo* differential motif enrichment analysis was performed (Heinz et al. 2010) comparing enrichment at binding sites versus the random background and this identified an enrichment for a xTGAxTCAx signature among 20% and 12% of binding sites which occurred only among 6% and 4.78% of

background genome sequences for the D195307 and D7F3158 donors respectively (Figure 6.25; $P<10^{-900}$). To provide validation for the identified motif, a non-differential *de novo* search analysis was performed leveraging a distinct algorithm (Machanick et al. 2011) and the same motif signature was recapitulated in both donors (Figure 6.26; $P<2x10^{-98}$).

Figure 6.25: The identified motif signature using HOMER's differential enrichment algorithm ($P<10^{-900}$).

Figure 6.26: Independently identified motif using MEME's non-differential enrichment algorithm ($P<2x10^{-98}$).

The identified sequence was compared to currently known motifs (Gupta et al. 2007) and was found to be identical to a motif previously identified as enriched among activator protein 1 (AP1) binding sites (Hull et al. 2013). AP1 is a heterodimeric transcription factor known to regulate transcription in response to inflammatory stimuli (Hess et al. 2004). Interestingly, this signature was also described to be enriched among the binding sites of activating transcription factor 3 (ATF3), nuclear basic leucine zipper (BATF) as well as jun dimerization protein 2 (JDP2) (Wang et al. 2012). BATF mediates dimerization with members of the jun proteins acting as a negative regulator of ATF transcriptional axis (Dorsey et al. 1995). ATF3 is involved in the cellular stress response processes, and JDP2 is ATF3's paralogue (Weidenfeld-Baranboim et al. 2009). Among other nuclear proteins resembling INTS12 enriched motif are nuclear factor erythroid 2 (NFE2) and Fos-related antigen 2 (FOSL2) (Consortium, 2007). In conclusion, the *de novo* identified motif in the

INTS12 ChIPseq dataset appears to be a common signature in a diverse range of regulators of gene transcription.

6.3.2.9.1 Distribution analysis of the enriched INTS12 signature

Central motif enrichment analysis can potentially identify whether the studied protein shows evidence of direct or cooperative DNA binding based on the probability distribution of the enriched motif among its binding sites (Machanick and Bailey, 2011). This approach was used to test the most likely binding pattern of INTS12. Although the motif appears to be centrally distributed, the site probability is relatively broad (±158bp from the peak summit) suggesting that much of the binding via the identified motif occurs in cooperation with other molecules including those sharing the identified common signature (Figure 6.27).

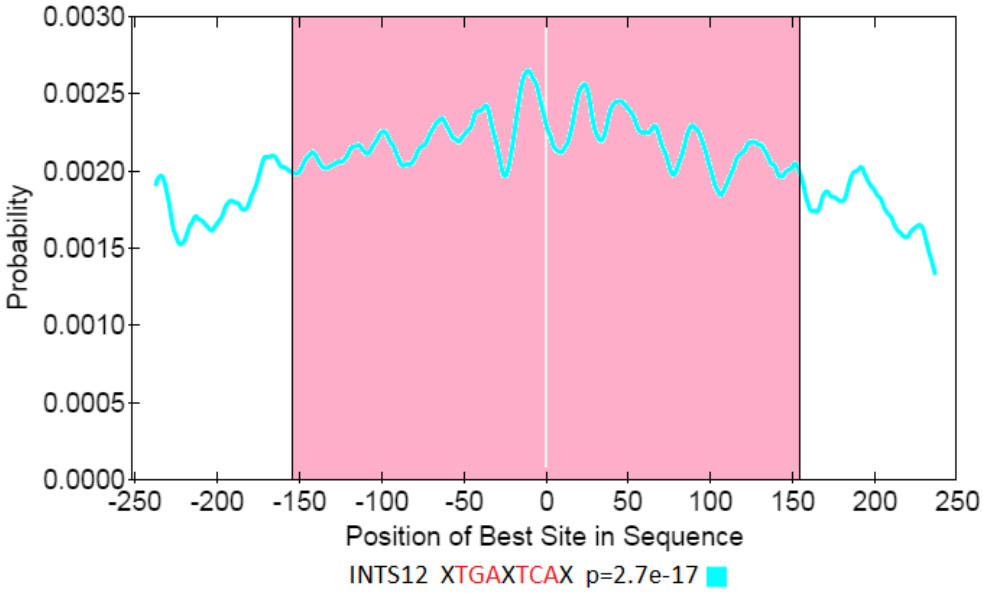


Figure 6.27: Probability distribution of INTS12 binding enriched DNA motif TGAxTCA across the sites at which it is present. Position at zero represents peak summit and motif appears to be centrally enriched 158bp in each direction from this summit. The site probability curve is very broad indicating indirect or cooperative binding to the DNA. The figure shows the motif, its central enrichment P-value and the width of the enriched region.

6.4 Discussion

The data presented in Chapter 5 showed that there is molecular and phenotypic evidence to suggest that INTS12 is a regulator of pathways

important in protein synthesis. This Chapter began with a key question of whether translational misbalance due to INTS12 depletion is a secondary epiphenomenon or a direct primary response. In an attempt to provide an answer to this question and to get an insight on the regulatory roles of INTS12, a ChIPseq approach was performed for this protein. As a proof of principle, pilot ChIPseq was conducted and showed that INTS12 displays enough genome binding to yield high quality reads for factor analysis. Moreover, INTS12 antibody was shown to be specific by IF in combination with gene knockdown. Based on that finding, INTS12 ChIPseq was performed and the data obtained analysed separately as well as in combination with RNAseq data.

Examination of binding association with fixed features of the human genome revealed the highest enrichment at intron, intergenic and promoter regions. However, gene-centric analysis revealed a clear localization near the TSS. Sub-setting genes into protein coding, lincRNA, snRNA, snoRNA, and microRNA showed that binding differed between these classes. Interestingly, the highest enrichment was observed for protein coding and snRNA genes with peak summits near TSS and TES respectively. A higher enrichment near TES for snRNA genes goes hand-in-hand with the proposed INTS12 canonical role in snRNA processing. On the other hand, enrichment near TSS for protein coding genes suggests INTS12 to be having a different functional role over this class of genes implying INTS12 to be pleiotropic.

Co-localization and ChIPseq signal correlations were also performed for HBEC relevant features as INTS12 binding was profiled in this cell type. On a genome-wide scale and within promoter regions INTS12 showed prominent associations with DNaseI and H3K4me3, both marking actively transcribed regions. In agreement with this observation, INTS12 was more highly recruited to expressed genes than silent loci and the degree of binding corresponded well with the magnitude of basal gene expression. DNA motif analysis revealed enrichment for a common signature enriched at binding sites of other regulators of gene expression. Motif distribution analysis indicated cooperative binding of INTS12, however this remains to be further explored experimentally.

Finally, INTS12 ChIPseq experiments showed a statistically rigorous preferential enrichment for TSS of genes with evidence of differential expression following INTS12 depletion identifying a set of genes directly regulated by INTS12.

Thus coming back to the initial question, it is possible to say that INTS12 is likely to be a direct regulator of pathways important in protein synthesis because (i) gene expression changes induced by knockdown can be causally attributed to the deficiency in INTS12 levels as INTS12 was experimentally manipulated (ii) INTS12 TSS binding is enriched for these differentially expressed genes. In other words, INTS12's regulome was defined and comprises genes belonging to the translational pathways. Interestingly, binding was biased towards the downregulated pathways: 90% of downregulated genes had higher regulatory potential scores than upregulated genes and the remaining 10% of genes had as high regulatory scores as upregulated genes, explaining a more robust effect seen upon the downregulated pathways in pathway analysis. Taken together these data suggest a key role for INTS12 in regulating gene expression with a prominent role for protein synthesis pathways.

7. General discussion

The studies described in this thesis were designed in an attempt to understand the biological mechanisms behind a previously identified genetic association signal for lung function at 4q24 using some of the *in silico*, *in vitro*, and *in vivo* approaches described in Chapter 1. This locus harbours genetic variation strongly and reproducibly associated with lung function parameters and risk of COPD (Repapi et al. 2010, Hancock et al. 2010, Castaldi et al. 2011, Wain et al. 2015). The most recent study has shown that there are at least three independent association signals within this locus, one located over the gene *TET2*, one over the gene for *NPNT*, and a third peak situated over the genes for *GSTCD* and *INTS12* (Wain et al. 2015). This thesis sought to understand the functional basis for the *GSTCD*/*INTS12* region signal. Analysis of lung eQTL dataset (Lonsdale et al. 2013) suggested *INTS12* may be involved in the phenotypic manifestation of altered lung function or pathogenesis of COPD (see Chapter 3). Therefore, the key question to answer is how might variation in the expression of *INTS12* contribute to these phenotypes? To help answer this question, work was undertaken aiming to illuminate the molecular and cellular functions of this candidate gene in the lung. This was achieved using both hypothesis driven (described in Chapter 4) and hypothesis free approaches (described in Chapter 5 and 6).

7.1 Thesis conclusions

The challenge in the post-GWAS era is the identification of causal genetic variants contributing to the variation of the considered trait. Because association does not imply causation, it does not follow that variants with the most significant P-values are the key ones. As outlined in Chapter 1, one possible approach to address this challenge is to map other variants in strong linkage with the highly significant variant in order to reconstruct an entire haplotype with an evidence of association. Such a block can then be intersected with the available functional annotation in this region which may pinpoint the possible functional variant. It is important to keep in mind that this method does not prove that a

particular genetic variant associated with a particular functional annotation to be causative. It merely raises the likelihood that this variant is contributory. More conclusive studies require experimental manipulation of a variant or expression of candidate gene linked to the putative variant controlling its transcription and/or translation.

eQTL analyses in a range of tissues (Obeidat et al. 2013) and human lung (Table 3.1) provide evidence that the genetic association signal for lung function and risk of COPD at 4q24 (Repapi et al. 2010, Hancock et al. 2010, Castaldi et al. 2011, Wain et al. 2015), may in part be due to altered expression of INTS12. This was further supported by results presented in Chapter 3 of this thesis which showed that, in human lung, a subset of SNPs tagging the above mentioned 4q24 haplotype, are genome-wide significant for lung function and are eQTLs for *INTS12* but not neighbouring *GSTCD*.

The first aim of this thesis was to investigate the potential role of INTS12 in contributing to lung function via the snRNA processing pathway. As the specific INTS12 requirement for snRNA 3'end formation was demonstrated in *Drosophila* only, its potential contribution to this activity was tested in HBECs. Out of four snRNA genes only one showed evidence of misprocessing following INTS12 depletion (Chapter 4). In the light of the fact that across the metazoans PHD of INTS12 is very well conserved (Chapter 3) and that it is dispensable for snRNA processing in the fly cells (Chapter 4), it was conjectured that INTS12 has novel, yet to be defined functions. Given that there is no homologous INTS12 gene in *C.elegans* and *S.cerevisiae* genomes, but strong conservation in mammals and a lethal effect of its knockout in *M.musculus* (Obeidat et al. 2013), INTS12 may be important for the evolution of metazoan tissue differentiation and specialization. Moreover, INTS12 has sequence similarity to regulators of gene expression (Chapter 3). Thus the second aim of the thesis was to uncover novel functions and regulatory properties of INTS12 that are important in cell homeostasis using a hypothesis free approach.

Following knockdown with two D-siRNAs, marked downregulation of pathways critical in protein synthesis was observed (Chapter 5). To

further investigate the functional importance of this, additional experiments were undertaken and showed that suppression of INTS12 affects the relevant cell phenotype by reducing the rate of protein synthesis and proliferation (Chapter 5). As part of the RNAseq analysis three novel INTS12 splice variants were discovered, validated and corrected by PCR and Sanger sequencing (Chapter 5). The important point learned from the splice analysis was that for the low abundance variants there is a risk of erroneous splice junction inference during transcriptome assembly as out of novel variants, one agreed with Sanger sequence and aligned at the expected position on the electrophoresis gel while the other two had discrepancies between RNAseq and Sanger inferred structures.

INTS12 ChIPseq was undertaken in order to delve deeper into the possible mechanism behind the observed dysregulation of protein synthesis related pathways following gene knockdown. The key question was whether observed effects are primary responses or secondary epiphenomena. Because INTS12 ChIPseq has not been performed before, a pilot experiment was executed. The success of the pilot ChIPseq demonstrated that there is enough interaction between this protein and genomic DNA for high quality sequencing data to be generated. Genome-wide analyses showed INTS12 binding to be high at introns, intergenic regions, and promoters. Gene-centric analyses revealed INTS12 binding to be near transcriptional start sites. For protein-coding genes the peak binding is proximal to the TSS while for snRNA genes the binding is enriched downstream to the TES suggesting distinct functional activities for INTS12 depending on the class of the genes where INTS12 binding occurs. Co-localization studies revealed association with accessible chromatin and H3K4me3 both marking transcriptionally active regions.

In agreement with this observation, the degree of binding was found to correspond with the magnitude of basal gene expression and binding sites were enriched for differentially expressed loci. Crucially, INTS12 ChIPseq experiments showed a preferential enrichment for TSS of differentially expressed genes uncovering INTS12's regulome which

comprises genes forming part of several protein synthesis relevant pathways, in particular the tRNA synthetases pathway. RNAseq of INTS12 depleted cells and INTS12 ChIPseq taken together, identify a significant unrecognised role for INTS12 in protein synthesis regulation via direct regulation of genes belonging to protein synthesis relevant pathways.

INTS12 is a member of the INTScom which itself has been shown to be implicated in various molecular and cellular processes (Baillat et al. 2005, Tao et al. 2009, Rutkowski and Warren 2009, Takata et al. 2012, Jodoin, Sitaram et al. 2013, Otani et al. 2013, Gardini et al. 2014) but it remains unclear whether all INTScom subunits are required for some of all these processes. For example, although INTS3 and INTS4 are both INTScom members, INTS3 is entirely dispensable while INTS4 is fundamentally required for the snRNA processing (Ezzeddine et al. 2011). What remains to be elucidated is how INTScom perturbations yield such specific yet so diverse phenotypes. It has been suggested that the mechanism behind that is the alteration of snRNA 3'-end formation affecting the splicing of mRNAs belonging to genes of particular functional groups explaining the specific phenotypic effects (Tao et al. 2009, Otani et al. 2013, Jodoin, Sitaram et al. 2013), including the maintenance of epithelial cilia (Jodoin, Shboul et al. 2013). For example, it has been argued that the induced downregulation of INTS5, INTS9, and INTS11 in zebrafish causes impaired haematopoiesis due to aberrant splicing of smad1 and smad5 via a dominant negative form of Smad transcripts (Tao et al. 2009). However, given the facts that INTS11 depletion results in a loss of perinuclear dynein whilst there was no enrichment for misprocessed transcripts encoding dynein-dynactin subunits, adaptor molecules or dynein-binding cassettes in HeLa cells (Jodoin, Sitaram et al. 2013) and our own observation of minor effect of INTS12 knockdown on snRNA processing concurrent with misbalanced protein synthesis in HBECs, it is possible to say that this hypothesis seems unlikely in a human model.

An alternative model for INTScom, where its subunits have different activities despite their physical association in the same complex and with

POLII, seems more likely to be true. Consequently, individual INTScom members are pleiotropic (Rutkowski and Warren, 2009) and have distinct functions which may explain the plethora of phenotypes observed following INTScom perturbation. Such a hypothesis allows for a more specific verifiable prediction then the very generic model involving snRNA processing which could account for virtually any cellular function or phenotype. Regardless of whether INTS12 is truly involved in snRNA 3'-end formation, allele carriers with low INTS12 expression and therefore at risk of lowered lung function (Obeidat et al. 2013) are unlikely to have decreased levels of INTS12 to an extent resulting in snRNA misprocessing. Thus the potential effects of INTS12 on lung function may not be explained via snRNA processing pathway. The discovered role for INTS12 in translational control suggests that altered levels of INTS12 in specific allele carriers may have widespread effects on protein synthesis control via gene regulation of relevance to lung development and repair; i.e. shown to be important for lung function and health based on the survey of literature data. This opens the door for possible future recall-by-genotype or gene knockout models that could look for this specific outcome.

7.2 Pathways forward – preliminary explorations and considerations

7.2.1 The effects of full length and serine-rich domain missing INTS12 overexpression on gene expression

Chapter 5 presents evidence that INTS12 depletion in bronchial epithelial cells results in a greater number of downregulated rather than upregulated pathways (Chapter 5). This observation is in keeping with the higher regulatory potential scores among the set of repressed genes versus the set of induced genes (Chapter 6). As the gene-centric analysis of genome-wide INTS12 binding revealed clear localization in proximity to the TSS, one possible molecular mechanistic hypothesis is that INTS12 has transcription controlling properties. Consequently, since

removal of this nuclear protein resulted in a reduced expression of many members the tRNA synthetases pathway including *MARS* and *GARS* as well as *ATF4* and *ASNS* of the PERK pathway, it follows that these genes might be upregulated when INTS12 is overexpressed. Such a result would imply that INTS12 is their activator.

In a preliminary attempt to shed light on this question, a full length and truncated INTS12 with a missing serine-rich domain were transiently overexpressed (see section 3.6.2). The rational for overexpressing the truncated protein is to understand whether this missing part of the protein is required for such activation. Surprisingly, overexpression of full length INTS12 consistently reduced the expression of these genes although this effect was not statistically significant ($P>0.05$). No effect on the *MARS*, *GARS*, *ATF4* and *ASNS* expression was observed in cells overexpressed with truncated version of INTS12. Future work could clarify the effects of INTS12 overexpression on airway cells. One possible approach would be to perform global gene expression profiling following these manipulations and compare the results to the gene expression profiling following gene knockdown. This could provide further insights into the regulatory properties of INTS12. It is especially worth investigating whether genes that are differentially expressed due to INTS12 depletion are dysregulated when INTS12 is overexpressed and to compare the effects of full length and truncated variants.

7.2.1.1 Proposal of INTS12 function hypothesis

If it is the case that both highly elevated and depleted INTS12 cause reduction in the expression of genes belonging to its regulome, it is possible that the level of INTS12 has to be very finely tuned, probably relative to other factors which it has to bind with, in order to fulfil its function. INTS12 could be bridging two protein complexes which is a plausible scenario: first considering the fact that it is already part of INTScom complex, and secondly it highly associates with K3K4me3 histone tail modification (Chapter 6) implying a likely physical interaction. If the function of INTS12 is to bind INTScom/POLII and certain histone tail modifications thus tethering transcriptional complex to the promoters,

then at very low levels of INTS12, no bridging can occur and INTScom would not be recruited to the promoter resulting in reduced transcription. As the levels of INTS12 rise, more of the INTScom/POLII is recruited to the promoters until a plateau is reached, where all suitably modified promoter sites have recruited the complex or all available INTScom/POLII is bound to promoters. As the levels of INTS12 exceed the optimum, dominant negative effects may be starting to exert their influence. For instance, promoters may have INTS12 bound without the tethering of INTScom/POLII due to compromised biding caused by INTS12 overloading. So at high concentrations, INTS12 might become a repressor of transcription at target regulome genes.

7.2.2 *In vivo* approaches

Up to this point of this thesis all the methods used to study INTS12 function and understand mechanistic basis for association signal at 4q24 were either *in silico* or *in vitro*. Out of the repertoire of various methods described in Chapter 1, I have not used to date an *in vivo* approach. Thanks to the efforts of IMPC (see section 1.6.3.1) there is publically available detailed information regarding the characteristics of approx. 20,000 protein-coding knockout mouse models. This resource was used in order to undertake a preliminary exploration of the phenotypic manifestations of homozygous and heterozygous INTS12 knockout mouse models.

Interestingly, heterozygous models show a small but significant decreased erythrocyte mean cell volume in both males and females ($P<0.0001$). This finding is interesting in the light of fact that INTS5 has been implicated in zebrafish haemopoiesis (Tao et al. 2009). The fact that models with one copy of INTS12 removed have decreased erythrocyte volume suggests that perhaps INTS12 is also required for mammalian haematopoiesis and could be used as starting point for future studies. Another phenotype that manifests itself in the heterozygotes is the marginally increased level of circulating magnesium ($P<0.0001$). However, this has been observed in females only, whereas in males circulating magnesium may actually be decreased ($P=0.1$). It is

not clear at this stage why there is such a difference between males and females and it may be worth investigating the reasons for this dichotomy. The relevance of these findings to the lung is also unclear, and to date there are no reported lung phenotypes in heterozygous INTS12 mice. On the other hand, the homozygous INTS12 knock out mouse shows pre-weaning lethality (Obeidat et al. 2013) suggesting a critical role for INTS12 in early development.

As the expression of *MARS* and *GARS* of the tRNA synthetases pathway, and *ASNS* and *ATF4* from the PERK pathway were downregulated following D-siRNA treatment targeting INTS12, a prediction was made that these genes would also have reduced expression in the heterozygous INTS12 knockout model. cDNA representing the total RNA content of mouse lung was obtained from a small number of heterozygous and wild type models and relevant gene expression was measured. As expected, there was ~64% reduction in *INTS12* expression in the animals that had one copy of the gene removed but this decrease was not statistically significant (P=0.2). However, no effect on the expression of target genes was seen in the knockout mouse model (n=2).

It is possible that such effects cannot be observed with ~50% of gene being expressed because one functional copy may be masking the functional consequences of the gene removal. Moreover, in the *in vitro* knockdown experiments the degree of INTS12 suppression achieved was greater than 80% (Chapter 3). More experiments are warranted to ascertain the case as these data can only be considered preliminary (n=2). If it is the case that in heterozygous model there is no apparent effect on the expression of target genes, both conditional, complete and tissue specific INTS12 knockouts could be used instead. Having established whether target genes are differentially downregulated in either inducible or non-inducible models it would also be worth investigating whether the rate of protein synthesis is altered in the lung in a manner analogous to the human knockdown experiments (see Chapter 5). One potential issue in such experiments is whether or not INTS12 has the same role in murine as in human tissue. The high degree

of interspecies sequence homology would support this, although as noted earlier the role of INTS12 in *Drosophila* appears to be significantly different than in human cells.

7.2.2.1 *In vivo* efforts in clinical translation

Another possible future avenue for further exploration could be a recall-by-genotype studies where human cells obtained from individuals on a known genetic background can be stratified and tested for a possible differences in protein synthesis rates. It may be worth testing whether cells harbouring alleles associated with lower lung function, and therefore lower INTS12 expression, intrinsically have lower rates of protein synthesis.

7.3 Summary

Lung eQTL analyses suggests INTS12 to be a likely gene whose variable expression is contributing to variation in lung function. I conclude that INTS12 is a pleiotropic gene with at least two different functions depending on the class of genes where its binding occurs. In agreement with the canonical function, over snRNA genes INTS12 is likely to contribute to their 3'-end formation. However, in contrast to what was reported in *Drosophila*, INTS12 requirement for snRNA processing is moderate in HBECs. Gene depletion resulted in dysregulation of a core subset of genes of relevance to lung physiology at two time points and in two different donors. Finally, the data presented herein identify a significant and previously unrecognized role for INTS12 in protein synthesis control via direct regulation of key protein coding genes belonging to the related cellular pathways.

Appendix

Donors

INTS12 knockdown was optimized in HBECs from D7F3206. 48h RNAseq experiment was performed upon D7F3206. 120h RNAseq experiment was performed upon D195307. Target gene expression identified in 120h RNAseq was biologically validated in D7F3206. Protein synthesis was performed in D195307. ChIPseq was performed in D195307 and D7F3158.

- D7F3206 was a 50 years old male Caucasian who was a smoker.
- D195307 was a 19 years old male who was not a smoker.
- D7F3158 was a 56 years old male Caucasian who was a smoker.

Tables

Table 1: The sequences of D-siRNAs used to suppress the INTS12 expression.

Oligo	Sequence
D-siRNA #A	5'-GGAAUGGAAAUAGUGGAACAUCAGG-3'
D-siRNA #B	5'-GGCAAUCAAUUAGUAGAAUGUCAGG-3'
D-siRNA #C	5'-GCGUUUAAGAGAACAGAAGUCAAGA-3'

Table 2: Pre-designed TaqMan qPCR assays for housekeeping genes (Life Technologies).

Target gene	Cat. num.
GAPDH	4310884E
HPRT1	4310890E
TfR	4331182E

Table 3: Pre-designed SYBR® Green qPCR assays (Sigma-Aldrich).

Target gene	Oligo	Sequence
MARS	Forward primer	5'-TACCCATTACTGCAAGATCC-3'
	Reverse primer	5'-CTTGCTGTTTCAGTACAGTC-3'
GARS	Forward primer	5'-GTGTTAGTGGTCTGTATGAC-3'
	Reverse primer	5'-GTCTTTAAAACTGGCTCAGG-3'
ASNS	Forward primer	5'-GATTGGCTGCCTTTTATCAG-3'
	Reverse primer	5'-AATTGCAAATGTCTGGAGAG-3'
ATF4	Forward primer	5'-CCTAGGTCTCTTAGATGATTACC-3'
	Reverse primer	5'-CAAGTCGAACTCCTTCAAATC-3'
LEP	Forward primer	5'-TCAATGACATTTCACACACG-3'
	Reverse primer	5'-TCCATCTTGGATAAGGTCAG-3'

Table 4: ChIP-PCR primer sequences.

Target gene site/primer	Primer sequence
POR forward	5'-CAGGGTCCGAGCTGTAGAAG-3'
POR reverse	5'-CCGGCAGAGAAATGAAAGTG-3'
NBPF1 forward	5'-CACCTACGCCTCCCAGTACC-3'
NBPF1 reverse	5'-GCCTTGGGTTATCCTGACAC-3'
ACTB forward	5'-AACTCTCCCTCCTCCTCTTCC-3'
ACTB reverse	5'-CCTCTCCCCTCCTTTTGC-3'
Untr12 forward	5'-TGAGCATTCCAGTGATTTATTG-3'
Untr12 reverse	5'-AAGCAGGTAAAGGTCCATATTTC-3'

Table 5: In-house designed TaqMan qPCR assays.

Target gene	Oligo	Sequence
INTS12	Forward primer	5'-CTCCAGCTGTCAAAGATCCATT-3'
	Reverse primer	5'-GAGAGCTGCTGGATTCTGAAGT-3'
	Probe	5'-TGGCTGCAAAAGCTGCCCATCCAG-3'

Table 6: In-house designed SYBR® Green qPCR assays.

Target gene	Oligo	Sequence
Immature U1	Forward primer	5'-GATGTGCTGACCCCTGCGATTTC-3'
	Reverse primer	5'-GTCTGTTTTTGAAACTCCAGAAAGTC-3'
Immature U2	Forward primer	5'-TTGCAGTACCTCCAGGAACGG-3'
	Reverse primer	5'-CAGGGAAGCAGTTAAGTTAAGCC-3'
Immature U4	Forward primer	5'-AGCTTTGCGCAGTGGCAGTATCG-3'
	Reverse primer	5'-AAGCCTCTGTTGTTCAACTGC-3'
Immature U5	Forward primer	5'-TACTCTGGTTTCTCTTCAGATCGC-3'
	Reverse primer	5'-TTCTATTGTTGGATTACCAC-3'

Developed python programs

gene.perXLOC_exp_parser.py

```
fname = raw_input("Enter input file name (if in local dir) or path:"'\n')

try:
    fh = open(fname)
except:
    print 'The input file ' + fname + ' does not exist DONE'
    exit()

output = raw_input("Enter output file name (if in local dir) or path" + '\n' + "WARNING:Any existing data will be erased: "'\n')
try:
    oh = open(output, "w")
except:
    'The output file ' + output + 'does not exist. DONE'

for line in fh:
    line = line.rstrip()
    line = line.split()
    if line[1].startswith('gene'):
        oh.write('NAME' + '       ' + line[2] + '       ' + line[3] + '       ' + line[4] + '\n')
        continue
    elif line[1].startswith('"'):
        line[1] = line[1].lstrip('"')
        line[1] = line[1].rstrip('"')
        genes = line[1].split(',')
        for gene in genes:
            oh.write(gene+ '       ' + line[2]+ '       ' + line[3]+ '       ' + line[4] + '\n')
    elif line[1].startswith('-'):
        continue
    else:
        oh.write(line[1] + '       ' + line[2] + '       ' + line[3] + '       ' + line[4] + '\n')
print 'DONE'
```

expression_Table_parser.py

```
fname = raw_input("Enter input file name (if in local dir) or path:"'\n')

try:
    fh = open(fname)
except:
    print 'The input file ' + fname + ' does not exist DONE'
    exit()

output = raw_input("Enter output file name (if in local dir) or path" + '\n' + "WARNING:Any existing data will be erased: "'\n')
try:
    oh = open(output, "w")
except:
    'The output file ' + output + 'does not exist. DONE'

for line in fh:
    line = line.rstrip()
    line = line.split()
    if line[1].startswith('gene'):
        oh.write('NAME' + '       ' + line[2] + '       ' + line[3] + '       ' + line[4] + '\n')
        continue
    elif line[1].startswith('"'):
        line[1] = line[1].lstrip('"')
        line[1] = line[1].rstrip('"')
        genes = line[1].split(',')
        for gene in genes:
            oh.write(gene+ '       ' + line[2]+ '       ' + line[3]+ '       ' + line[4] + '\n')
    elif line[1].startswith('-'):
        continue
    else:
        oh.write(line[1] + '       ' + line[2] + '       ' + line[3] + '       ' + line[4] + '\n')
print 'DONE'

for line in fh:
    if line.startswith('NAME'):
        continue
    line = line.rstrip()
    line = line.split()
    gene = line[0]
    if gene not in dct_map: #creats the dictionary where key is gene_name while value is list containing expression data. Redundant genes
have multiple FPKMs nested together within the sample.
        dct_map[gene] = [1, line[1], line[2], line[3], line[4], line[5], line[6], line[7], line[8], line[9], line[10], line[11], line[12]]
    else:
        dct_map[gene][0] = dct_map[gene][0] + 1
        dct_map[gene][1] = str(dct_map[gene][1]) + ' ' + line[1]
        dct_map[gene][2] = str(dct_map[gene][2]) + ' ' +  line[2]
        dct_map[gene][3] = str(dct_map[gene][3]) + ' ' +  line[3]
        dct_map[gene][4] = str(dct_map[gene][4]) + ' ' +  line[4]
        dct_map[gene][5] = str(dct_map[gene][5]) + ' ' +  line[5]
        dct_map[gene][6] = str(dct_map[gene][6]) + ' ' +  line[6]
        dct_map[gene][7] = str(dct_map[gene][7]) + ' ' +  line[7]
        dct_map[gene][8] = str(dct_map[gene][8]) + ' ' +  line[8]
        dct_map[gene][9] = str(dct_map[gene][9]) + ' ' +  line[9]
        dct_map[gene][10] = str(dct_map[gene][10]) + ' ' +  line[10]
        dct_map[gene][11] = str(dct_map[gene][11]) + ' ' +  line[11]
        dct_map[gene][12] = str(dct_map[gene][12]) + ' ' +  line[12]

print 'FINISHED GENERATING DICTIONARY '


output = raw_input("ENTER initial output file name" + '\n' + "The file will be created in local dir. If the file exists then it will be
overwritten!" + '\n')
oh = open(output, "w")



for gene,data in dct_map.items():
    if data[0] == 1:
         oh.write(gene + ' ' + str(dct_map[gene][0]) + ' ' + dct_map[gene][1] + ' ' + dct_map[gene][2] + ' ' + dct_map[gene][3] + ' ' +
dct_map[gene][4] + ' ' + dct_map[gene][5] + ' ' + dct_map[gene][6] + ' ' + dct_map[gene][7] + ' ' + dct_map[gene][8] + ' ' + dct_map[gene]
[9] + ' ' + dct_map[gene][10] + ' ' + dct_map[gene][11] + ' ' + dct_map[gene][12] + '\n')
    else:
        oh.write(gene + ' ' + str(dct_map[gene][0]) + ' ' + dct_map[gene][1] + ' ' + dct_map[gene][2] + ' ' + dct_map[gene][3] + ' ' +
dct_map[gene][4] + ' ' + dct_map[gene][5] + ' ' + dct_map[gene][6] + ' ' + dct_map[gene][7] + ' ' + dct_map[gene][8] + ' ' + dct_map[gene]
[9] + ' ' + dct_map[gene][10] + ' ' + dct_map[gene][11] + ' ' + dct_map[gene][12] + '\n')

print 'FINIFHED WRITING INITIAL OUTPUT FILE'

fh.close() #closing file handles
oh.close()

fh2 = open(output)

output2 = raw_input("ENTER FINAL OUTPUT file name" + '\n' + "The file will be created in local dir. If the file exists then it will be
overwritten!" + '\n')
oh2 = open(output2, "w") #makes file handle in writing mode for final file output

oh2.write('gene_name UT1 UT2 UT3 NC1 NC2 NC3 A1 A2 A3 C1 C2 C3''\n')



for generaw in fh2:
    generaw = generaw.rstrip()
    generaw = generaw.split()
    generaw = list(str_to_float(generaw)) #converts numeric strings into floats if possible
    if generaw[1] == 1:
        oh2.write(str(generaw[0]) + ' ' + str(generaw[2])+ ' ' +str(generaw[3])+ ' ' +str(generaw[4])+ ' ' +str(generaw[5])+ ' ' +str
(generaw[6])+ ' ' +str(generaw[7])+ ' ' +str(generaw[8])+ ' ' +str(generaw[9])+ ' ' +str(generaw[10])+ ' ' +str(generaw[11])+ ' ' +str
(generaw[12])+ ' ' +str(generaw[13])+ '\n')
    if generaw[1] > 1:
       generaw[2:] = block_of_num_sum_returner(generaw[2:],12)
       oh2.write(str(generaw[0]) + ' ' + str(generaw[2])+ ' ' +str(generaw[3])+ ' ' +str(generaw[4])+ ' ' +str(generaw[5])+ ' ' +str(generaw
[6])+ ' ' +str(generaw[7])+ ' ' +str(generaw[8])+ ' ' +str(generaw[9])+ ' ' +str(generaw[10])+ ' ' +str(generaw[11])+ ' ' +str(generaw[12])
+ ' ' +str(generaw[13])+ '\n')
```

pathway_database_parser.py

```
fname = raw_input("Enter pathways file name or path: ")

try:
    fh = open(fname)
except:
    print 'The file ' + fname + ' does not exist. DONE'
    exit()

genes_dct = dict()

for line in fh:
    line = line.rstrip()
    line = line.split() # creats a list of items
    line.remove(line[0]) # removes first item
    line.remove(line[0]) # removes second item (of original)
    for gene in line:
        genes_dct[gene] = genes_dct.get(gene, 0) + 1 # counts number of times genes occures


gene_lst = list()
for gene,count in genes_dct.items():
    gene_lst.append(gene)

print 'FINISHED CREATING GENES LIST'
output = raw_input("ENTER output file name" + '\n' + "The file will be created in local dir. If the file exists then it will be overwitten!"
+ '\n')
oh = open(output, "w")

for gene in gene_lst:
    oh.write(gene + '\n')

print 'DONE'
```

genes_extraction_from_my_exp_Table.py

```
exp_table_fname = raw_input("Enter gene expression table file name or path: ")

try:
    etf = open(exp_table_fname)
except:
    print 'The file does not exist. DONE'
    exit()

KEGG_limited_genes = raw_input("Enter to-be-extracted gene names file name or path: ")

try:
    klg = open(KEGG_limited_genes)
except:
    print 'The file does not exist. DONE'
    exit()

output = raw_input("ENTER output file name" + '\n' + "The file will be created in local dir. If the file exists then it will be overwitten!"
+ '\n')
oh = open(output, "w")

oh.write('gene_name UT1 UT2 UT3 NC1 NC2 NC3 A1 A2 A3 C1 C2 C3''\n')

KEGG_lst = list()
for gene in klg:
    gene = gene.rstrip()
    KEGG_lst.append(gene)

for line in etf:
    line = line.rstrip()
    if line.startswith('gene_name'):
        continue
    line = line.split()
    if line[0] in KEGG_lst:
        oh.write(str(line[0])+ ' ' + str(line[1]) + ' ' + str(line[2]) + ' ' + str(line[3])+ ' ' + str(line[4]) + ' ' + str(line[5]) + ' ' +
str(line[6]) + ' ' + str(line[7]) + ' ' + str(line[8]) + ' ' + str(line[9]) + ' ' + str(line[10]) + ' ' + str(line[11]) + ' ' + str(line[12])
+ ' ' + '\n')

print 'DONE'
```

max_min_pathway_genes_counter.py

```
f = raw_input('Enter file: ')
fh = open(f)
max_number = None
min_number = None
for line in fh:
    line = line.rstrip()
    line = line.split()
    #print line
    genes_per_path = len(line)-2
    if max_number == None or genes_per_path > max_number:
        max_number = genes_per_path
    elif min_number == None or genes_per_path < min_number:
        min_number = genes_per_path

print 'The largest pathway has ' + str(max_number) + ' genes'
print 'The smallest pathway has ' + str(min_number) + ' genes'
```

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., and 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., and 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
- Amarzguioui, M., Lundberg, P., Cantin, E., Hagstrom, J., Behlke, M.A. and Rossi, J.J. (2006) Rational design and in vitro and in vivo delivery of Dicer substrate siRNA, Nature Protocols, 1(2), pp. 508–517. doi: 10.1038/nprot.2006.72
- Bailey, T. L. and Machanick, P. (2012) Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res, 40, e128.
- 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., and 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., and Kouzarides, T. (2011) Regulation of chromatin by histone modifications. Cell Res, 21(3), 381-395. doi: 10.1038/cr.2011.22.
- Bartel, D.P. (2009) MicroRNAs: Target recognition and regulatory functions, Cell, 136(2), pp. 215–233. doi: 10.1016/j.cell.2009.01.002.
- Bellusci, S., Furuta, Y., Rush, M. G., Henderson, R., Winnier, G., and Hogan, B. L. (1997) Involvement of Sonic hedgehog (Shh) in mouse embryonic lung growth and morphogenesis. Development, 124(1), 53-63.
- Benjamini, Y., Hochberg Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Journal of the Royal Statistical Society. 57 (1): 289–300.

- Bienz, M. (2006) The PHD finger, a nuclear protein-interaction domain, Trends in Biochemical Sciences, 31(1), pp. 35–40. doi: 10.1016/j.tibs.2005.11.001.
- Bradford, M.M. (1976) Rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, Anal. Biochem. 72: 248–254.
- 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., and 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.
- Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J. and Wittwer, C.T. (2009) The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments, Clinical Chemistry, 55(4), pp. 611–622. doi: 10.1373/clinchem.2008.112797.
- 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.
- 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., Ezzeddine, N., Waltenspiel, B., Albrecht, T.R., Warren, W.D., Marzluff, W.F. and Wagner, E.J. (2012) An RNAi screen identifies additional members of the Drosophila integrator complex and a

requirement for cyclin C/Cdk8 in snRNA 3-end formation, RNA, 18(12), pp. 2148–2156. doi: 10.1261/rna.035725.112.

- Chen, J., Waltenspiel, B., Warren, W. D., and 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, T.-W., Li, H.-P., Lee, C.-C., Gan, R.-C., Huang, P.-J., Wu, T.H., Chang, Y.-F. and Tang, P. (2014) ChIPseek, a web-based analysis tool for chIP data, BMC Genomics, 15(1), p. 539. doi: 10.1186/1471-2164-15-539.
- 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.
- 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., 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.
- 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.
- Ciprandi, G., Capasso, M., Tosca, M., Salpietro, C., Salpietro, A., Marseglia, G., and 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.

- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. and Mortazavi, A. (2016) A survey of best practices for RNA-seq data analysis, Genome Biology, 17(1) doi: 10.1186/s13059-016-0881-8.
- 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., 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., and 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.
- Criner, G.J., Cordova, F., Sternberg, A.L. and Martinez, F.J. (2011) The national emphysema treatment trial (NETT), American Journal of Respiratory and Critical Care Medicine, 184(8), pp. 881–893. doi: 10.1164/rccm.201103-0455ci.
- Croft, D., Mundo, A.F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M.R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L. and DEustachio, P. (2013) The Reactome pathway knowledgebase, Nucleic Acids Research, 42(D1), pp. D472–D477. doi: 10.1093/nar/gkt1102.
- Dawkins, P. A., and Stockley, R. A. (2001) Animal models of chronic obstructive pulmonary disease. Thorax, 56(12), 972-977.
- Deaton, A. M., and Bird, A. (2011) CpG islands and the regulation of transcription. Genes Dev, 25(10), 1010-1022. doi: 10.1101/gad.2037511.
- DeMeo, D. L., and 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.

- 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., Mariani, T., Bhattacharya, S., Srisuma, S., Lange, C., Litonjua, A., Bueno, R., Pillai, S.G., Lomas, D.A., Sparrow, D., Shapiro, S.D., Criner, G.J., Kim, H.P., Chen, Z., Choi, A.M.K., Reilly, J. and Silverman, E.K. (2009) Integration of Genomic and genetic approaches Implicates IREB2 as a COPD susceptibility gene, The American Journal of Human Genetics, 85(4), pp. 493–502. doi: 10.1016/j.ajhg.2009.09.004.
- Dieffenbach, C.W., Lowe, T.M. and Dveksler, G.S. (1993) General concepts for PCR primer design, Genome Research, 3(3), pp. S30–S37. doi: 10.1101/gr.3.3.s30.
- Diez-Villanueva, A., Mallona, I., and 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., 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.
- Echeverri, C.J., Beachy, P.A., Baum, B., Boutros, M., Buchholz, F., Chanda, S.K., Downward, J., Ellenberg, J., Fraser, A.G., Hacohen, N., Hahn, W.C., Jackson, A.L., Kiger, A., Linsley, P.S., Lum, L., Ma, Y., Mathey-Prévôt, B., Root, D.E., Sabatini, D.M., Taipale, J., Perrimon, N. and Bernards, R. (2006) Minimizing the risk of reporting false positives in large-scale RNAi screens, Nature Methods, 3(10), pp. 777–779. doi: 10.1038/nmeth1006-777.
- Edwards, J., Belvisi, M., Dahlen, S. E., Holgate, S., and 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.

- Egloff, S., Al-Rawaf, H., OReilly, D. and Murphy, S. (2009) Chromatin structure is implicated in “late” elongation checkpoints on the U2 snRNA and -Actin genes, Molecular and Cellular Biology, 29(14), pp. 4002–4013. doi: 10.1128/mcb.00189-09.
- Egloff, S., OReilly, D. and Murphy, S. (2008) Expression of human snRNA genes from beginning to end, Biochemical Society Transactions, 36(4), pp. 590–594. doi: 10.1042/bst0360590.
- Eising, J.B., Uiterwaal, C.S.P.M., Evelein, A.M.V., Visseren, F.L.J. and van der Ent, C.K. (2013) Relationship between leptin and lung function in young healthy children, European Respiratory Journal, 43(4), pp. 1189–1192. doi: 10.1183/09031936.00149613.
- Engström, P.G., Steijger, T., Sipos, B., Grant, G.R., Kahles, A., Alioto, T., Behr, J., Bertone, P., Bohnert, R., Campagna, D., Davis, C.A., Dobin, A., Gingeras, T.R., Goldman, N., Guigó, R., Harrow, J., Hubbard, T.J., Jean, G., Kosarev, P., Li, S., Liu, J., Mason, C.E., Molodtsov, V., Ning, Z., Ponstingl, H., Prins, J.F., Rätsch, G., Ribeca, P., Seledtsov, I., Solovyev, V., Valle, G., Vitulo, N., Wang, K., Wu, T.D. and Zeller, G. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data, Nature Methods, 10(12), pp. 1185–1191. doi: 10.1038/nmeth.2722.
- Eriksson, S. (1965) Studies in alpha 1-antitrypsin deficiency. Acta Med Scand Suppl, 432, 1-85.
- Esposito, A.M. and Kinzy, T.G. (2014) In vivo [35 S]-Methionine incorporation, Laboratory Methods in Enzymology: Protein Part A, , pp. 55–64. doi: 10.1016/b978-0-12-420070-8.00005-2.
- 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.
- Ezzeddine, N., Chen, J., Waltenspiel, B., Burch, B., Albrecht, T., Zhuo, M., Warren, W.D., Marzluff, W.F. and Wagner, E.J. (2010) A subset of Drosophila integrator proteins is essential for efficient U7 snRNA and Spliceosomal snRNA 3-end formation, Molecular and Cellular Biology, 31(2), pp. 328–341. doi: 10.1128/mcb.00943-10.
- Falfan-Valencia, R., Pavon-Romero, G. F., Camarena, A., Garcia Mde, L., Galicia-Negrete, G., Negrete-Garcia, M. C. and Teran, L. M. (2012)

The IL1B-511 Polymorphism (rs16944 AA Genotype) Is Increased in Aspirin-Exacerbated Respiratory Disease in Mexican Population. J Allergy (Cairo) 741313.

- Farris, J.S. (2008) Parsimony and explanatory power, Cladistics, 24(5), pp. 825–847. doi: 10.1111/j.1096-0031.2008.00214.
- Favorov, A., Mularoni, L., Cope, L. M., Medvedeva, Y., Mironov, A. A., Makeev, V. J., and 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.
- Fisher, R.A. (1945) A new test for 2 × 2 tables, Nature, 156(3961), pp. 388–388. doi: 10.1038/156388a0.
- Forterre, P. (2015) The universal tree of life: An update, Frontiers in Microbiology, 6. doi: 10.3389/fmicb.2015.00717.
- Frank, J. A., Pittet, J. F., Wray, C. and Matthay, M. A. (2008) Protection from experimental ventilator-induced acute lung injury by IL-1 receptor blockade. Thorax, 63, 147-53.
- Gardini, A., Baillat, D., Cesaroni, M., Hu, D., Marinis, J.M., Wagner, E.J., Lazar, M.A., Shilatifard, A. and Shiekhattar, R. (2014) Integrator regulates Transcriptional initiation and pause release following activation, Molecular Cell, 56(1), pp. 128–139. doi: 10.1016/j.molcel.2014.08.004.
- Gel, B., Diez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A. and Malinverni, R. (2016) regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics, 32 289-91.
- 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. and Nolan, C. M. (2010) Epigenetics and gene expression. Heredity (Edinb), 105, 4-13.
- 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.

- Gish, W. and States, D.J. (1993) Identification of protein coding regions by database similarity search, Nature Genetics, 3(3), pp. 266–272. doi: 10.1038/ng0393-266.
- Glaab, E., Baudot, A., Krasnogor, N. and Valencia, A. (2010) Extending pathways and processes using molecular interaction networks to analyse cancer genome data. BMC Bioinformatics, 11, 597.
- Grutters, J. C., Sato, H., Pantelidis, P., Ruven, H. J., Mcgrath, D. S., Wells, A. U., Van Den Bosch, J. M., Welsh, K. I. and Du Bois, R. M. (2003) Analysis of IL6 and IL1A gene polymorphisms in UK and Dutch patients with sarcoidosis. Sarcoidosis Vasc Diffuse Lung Dis 20 20-7.
- Gunther, E. J., Belka, G. K., Wertheim, G. B., Wang, J., Hartman, J. L., Boxer, R. B., and 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.
- Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. and Noble, W. S. (2007) Quantifying similarity between motifs. Genome Biol, 8, R24.
- Hackett, T. L., Holloway, R., Holgate, S. T., and 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., 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.
- 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.
- 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.

- 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.
- Harding, H.P., Zhang, Y., Zeng, H., Novoa, I., Lu, P.D., Calfon, M., Sadri, N., Yun, C., Popko, B., Paules, R., Stojdl, D.F., Bell, J.C., Hettmann, T., Leiden, J.M. and Ron, D. (2003) An integrated stress response regulates amino acid metabolism and resistance to Oxidative stress, Molecular Cell, 11(3), pp. 619–633. doi: 10.1016/s1097-2765(03)00105-9.
- Hata, T. and Nakayama, M. (2007) Targeted disruption of the murine large nuclear KIAA1440/Ints1 protein causes growth arrest in early blastocyst stage embryos and eventual apoptotic cell death, Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1773(7), pp. 1039–1051. doi: 10.1016/j.bbamcr.2007.04.010.
- Hayter, A.J. (1986) The maximum Familywise error rate of Fishers least significant difference test, Journal of the American Statistical Association, 81(396), p. 1000. doi: 10.2307/2289074.
- Hedges, S.B., Dudley, J. and Kumar, S. (2006) TimeTree: A public knowledge-base of divergence times among organisms, Bioinformatics, 22(23), pp. 2971–2972. doi: 10.1093/bioinformatics/btl505.
- 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.
- Heinzmann, A., Ahlert, I., Kurz, T., Berner, R. and Deichmann, K. A. (2004) Association study suggests opposite effects of polymorphisms within IL8 on bronchial asthma and respiratory syncytial virus bronchiolitis. J Allergy Clin Immunol, 114, 671-6.

- Hellman, L. M., and 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., and Sundquist, J. (2008) 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.
- Hernandez, N. (2001) Small nuclear RNA genes: A model system to study fundamental mechanisms of transcription, Journal of Biological Chemistry, 276(29), pp. 26733–26736. doi: 10.1074/jbc.r100032200.
- Hersh, C.P., DeMeo, D.L. and Silverman, E.K. (2008) National emphysema treatment trial state of the art: Genetics of emphysema, Proceedings of the American Thoracic Society, 5(4), pp. 486–493. doi: 10.1513/pats.200706-078et.
- Hirschhorn, J. N., and 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.
- House, J. S., Li, H., DeGraff, L. M., Flake, G., Zeldin, D. C., and 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., and 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., and 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., and 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.
- Hull, R.P., Srivastava, P.K., DSouza, Z., Atanur, S.S., Mechta-Grigoriou, F., Game, L., Petretto, E., Cook, H.T., Aitman, T.J. and Behmoaras, J. (2013) Combined chIP-seq and transcriptome analysis identifies AP-1/JunD as a primary regulator of oxidative stress and IL-1β synthesis in macrophages, BMC Genomics, 14(1), p. 92. doi: 10.1186/1471-2164-14-92.
- 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., and Pauwels, R. (2002) Global Initiative for Chronic Obstructive Lung Diseases (GOLD) Pulm Pharmacol Ther, 15(4), 353-355.
- 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., and 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
- Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G. and Linsley, P.S. (2003) Expression profiling reveals off-target gene regulation by RNAi, Nature Biotechnology, 21(6), pp. 635–637. doi: 10.1038/nbt831.

- Jackson, A.L. and Linsley, P.S. (2010) Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application, Nature Reviews Drug Discovery, 9(1), pp. 57–67. doi: 10.1038/nrd3010.
- Jacobsen, L. (2004) FuGENE 6 Transfection reagent: The gentle power, Methods, 33(2), pp. 104–112. doi: 10.1016/j.ymeth.2003.11.002.
- Janciauskiene, S. M., Bals, R., Koczulla, R., Vogelmeier, C., Kohnlein, T., and 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.
- Jiang, H. and Wong, W.H. (2008) SeqMap: Mapping massive amount of oligonucleotides to the genome, Bioinformatics, 24(20), pp. 2395–2396. doi: 10.1093/bioinformatics/btn429.
- Jiang, C. and Pugh, B. F. (2009) Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet, 10, 161-72.
- Jodoin, J.N., Shboul, M., Albrecht, T.R., Lee, E., Wagner, E.J., Reversade, B. and Lee, L.A. (2013) The snRNA-processing complex, integrator, is required for ciliogenesis and dynein recruitment to the nuclear envelope via distinct mechanisms, Biology Open, 2(12), pp. 1390–1396. doi: 10.1242/bio.20136981.
- Jodoin, J.N., Sitaram, P., Albrecht, T.R., May, S.B., Shboul, M., Lee, E., Reversade, B., Wagner, E.J. and Lee, L.A. (2013) Nuclear-localized asunder regulates cytoplasmic dynein localization via its role in the integrator complex, Molecular Biology of the Cell, 24(18), pp. 2954–2965. doi: 10.1091/mbc.e13-05-0254.
- Joung, J. K., and 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.
- Kanehisa, M. (2000) KEGG: Kyoto encyclopedia of genes and Genomes, Nucleic Acids Research, 28(1), pp. 27–30. doi: 10.1093/nar/28.1.27.
- Kheirallah, A. K., Miller, S., Hall, I. P. and Sayers, I. (2016) Translating Lung Function Genome-Wide Association Study (GWAS) Findings: New Insights for Lung Biology. Adv Genet, 93, 57-145.
- 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.
- Kim, W.J., Lim, J.H., Lee, J.S., Lee, S.-D., Kim, J.H. and Oh, Y.-M. (2015) Comprehensive analysis of Transcriptome Sequencing data in the lung tissues of COPD subjects, International Journal of Genomics, 2015, pp. 1–9. doi: 10.1155/2015/206937.
- Klimentidis, Y. C., Vazquez, A. I., de Los Campos, G., Allison, D. B., Dransfield, M. T., and 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.
- Knight, J.C. (2014) Approaches for establishing the function of regulatory genetic variants involved in disease, Genome Medicine, 6(10) doi: 10.1186/s13073-014-0092-4.
- 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., and Schork, N. J. (1994) Genetic dissection of complex traits. Science, 265(5181), 2037-2048.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J.,

LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G.R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater,

G., Smit, A.F.A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Patrinos, A. and Morgan, M.J. (2001) Initial sequencing and analysis of the human genome, Nature, 409(6822), pp. 860–921. doi: 10.1038/35057062.

- 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.
- Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biology, 10(3), p. R25. doi: 10.1186/gb-2009-10-3-r25.
- Latchman, D. S. (1997) Transcription factors: an overview. Int J Biochem Cell Biol, 29(12), 1305-1312.
- Laurell, C. B. and Eriksson, S. (2013) The electrophoretic alpha1-globulin pattern of serum in alpha1-antitrypsin deficiency. 1963. COPD, 10 Suppl 1, 3-8.
- Lebecque, P., Kiakulanda, P., and 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, T. I., and Young, R. A. (2000) Transcription of eukaryotic protein-coding genes. Annu Rev Genet, 34, 77-137. doi: 10.1146/annurev.genet.34.1.77.
- Lee, B. Y., Cho, S., Shin, D. H., and 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.2010.11.00.
- 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.

- Lewitter, F. I., Tager, I. B., McGue, M., Tishler, P. V., and Speizer, F. E. (1984) Genetic and environmental determinants of level of pulmonary function. Am J Epidemiol, 120(4), 518-530.
- Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores, Genome Research, 18(11), pp. 1851–1858. doi: 10.1101/gr.078212.108.
- Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.
- Li, Y., and 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.
- 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.
- 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.
- Lipman, N.S., Jackson, L.R., Trudel, L.J. and Weis-Garcia, F. (2005) Monoclonal versus Polyclonal antibodies: Distinguishing characteristics, applications, and information resources, ILAR Journal, 46(3), pp. 258–268. doi: 10.1093/ilar.46.3.258.
- Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y., Pape, U.J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y. and Liu, X.S. (2011) Cistrome: An integrative platform for transcriptional regulation studies, Genome Biology, 12(8), p. R83. doi: 10.1186/gb-2011-12-8-r83.
- Livak, K. J. and Schmittgen, T. D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2-ddCt method. Methods 25, 402-8.
- Lodish, H., Lodish, U.H., Berk, U.A., Matsudaira, U.P. and Matsudaira, P. (2004) Molecular cell biology, 5e + working with molecular cell biology. New York, NY, United States: W.H. Freeman and Company.
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M.,

Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J. and Moore, H.F. (2013) The Genotype-Tissue expression (GTEx) project, Nature Genetics, 45(6), pp. 580–585. doi: 10.1038/ng.2653.

- 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., and 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.
- Ma, W. and Wong, W.H. (2011) The analysis of ChIP-seq data, Synthetic Biology, Part A, pp. 51–73. doi: 10.1016/b978-0-12-385075-1.00003-2.

- Ma, W., Noble, W.S. and Bailey, T.L. (2014) Motif-based analysis of large nucleotide data sets using MEME-ChIP, Nature Protocols, 9(6), pp. 1428–1450. doi: 10.1038/nprot.2014.083.
- Machanick, P. and Bailey, T. L. (2011) MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics 27, 1696-7.
- Madrigal, P. and Krajewski, P. (2012) Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. Front Genet, 3 230.
- Maeder, M. L., Linder, S. J., Cascio, V. M., Fu, Y., Ho, Q. H., and 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.
- Maier, T., Güell, M. and Serrano, L. (2009) Correlation of mRNA and protein in complex biological samples, FEBS Letters, 583(24), pp. 3966–3973. doi: 10.1016/j.febslet.2009.10.036.
- Majewski, J., and 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.
- Makinde, T., Murphy, R.F. and Agrawal, D.K. (2007) The regulatory role of TGF-β in airway remodeling in asthma, Immunology and Cell Biology, 85(5), pp. 348–356. doi: 10.1038/sj.icb.7100044.
- 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.
- Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L. and Tong, L. (2006) Polyadenylation factor CPSF-73 is the pre-mRNA 3-end-processing endonuclease, Nature, 444(7121), pp. 953–956. doi: 10.1038/nature05363.
- Marchini, J., and Howie, B. (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet, 11(7), 499-511. doi: 10.1038/nrg2796.
- Marciniak, S. J., Garcia-Bonilla, L., Hu, J., Harding, H. P. and Ron, D. (2006) Activation-dependent substrate recruitment by the eukaryotic translation initiation factor 2 kinase PERK. J Cell Biol, 172 201-9.

- Mardis, E. R. (2011) A decade's perspective on DNA sequencing technology. Nature, 470(7333), 198-203. doi: 10.1038/nature09796.
- Marnetto, D., Molineris, I., Grassi, E. and Provero, P. (2014) Genome-wide identification and characterization of fixed human-specific regulatory regions. Am J Hum Genet, 95, 39-48.
- Masters, J. R. (2000) Human cancer cell lines: fact and fantasy. Nat Rev Mol Cell Biol, 1(3), 233-236. doi: 10.1038/35043102
- Matera, A.G., Terns, R.M. and Terns, M.P. (2007) Non-coding RNAs: Lessons from the small nuclear and small nucleolar RNAs, Nature Reviews Molecular Cell Biology, 8(3), pp. 209–220. doi: 10.1038/nrm2124.
- 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., and 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., and 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, L. A., Wert, S. E., Clark, J. C., Xu, Y., Perl, A. K., and 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.
- 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, S., Melén, E., Merid, S.K., Hall, I.P. and Sayers, I. (2016) Genes associated with polymorphic variants predicting lung function are differentially expressed during human lung development, Respiratory Research, 17(1) doi: 10.1186/s12931-016-0410-z.
- Moore, K.L. and Persaud, T.V.N. (2003) The developing human: Clinically oriented embryology. 7th edn. Philadelphia, PA: Saunders (W.B.) Co.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nature Methods, 5(7), pp. 621–628. doi: 10.1038/nmeth.1226.
- Morton, N.E. (1955) Sequential tests for the detection of linkage. American Journal of Human Genetics. 7 (3): 277–318.
- Mugal, C.F., Wolf, J.B.W. and Kaj, I. (2013) Why time matters: Codon evolution and the temporal dynamics of dN/dS, Molecular Biology and Evolution, 31(1), pp. 212–231. doi: 10.1093/molbev/mst192.
- Nei M. and Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3(5):418-26.
- Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and 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.
- 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., Nelson, C.P., Stewart, C.E., Swan, C., Wain, L.V., Artigas, M.S., Melén, E., Ushey, K., Hao, K., Lamontagne, M., Bossé, Y., Postma, D.S., Tobin, M.D., Sayers, I. and Hall, I.P. (2013) GSTCD and INTS12 regulation and expression in the human lung, PLoS ONE, 8(9), p. e74630. doi: 10.1371/journal.pone.0074630.
- Ober, C., Abney, M., and 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.
- Ohta, T. (2006) Gene families: Multigene families and Superfamilies, Encyclopedia of Life Sciences. doi: 10.1038/npg.els.0005126.
- O'Reilly, D., Dienstbier, M., Cowley, S.A., Vazquez, P., Drozdz, M., Taylor, S., James, W.S. and Murphy, S. (2012) Differentially expressed, variant U1 snRNAs regulate gene expression in human cells, Genome Research, 23(2), pp. 281–291. doi: 10.1101/gr.142968.112.
- Otani, Y., Nakatsu, Y., Sakoda, H., Fukushima, T., Fujishiro, M., Kushiyama, A., Okubo, H., Tsuchiya, Y., Ohno, H., Takahashi, S.-I., Nishimura, F., Kamata, H., Katagiri, H. and Asano, T. (2013) Integrator complex plays an essential role in adipose differentiation, Biochemical and Biophysical Research Communications, 434(2), pp. 197–202. doi: 10.1016/j.bbrc.2013.03.029.
- Page, M. (1999) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on Phylogenies, Systematic Biology, 48(3), pp. 612–622. doi: 10.1080/106351599260184.
- Palmer, L. J., Celedon, J. C., Chapman, H. A., Speizer, F. E., Weiss, S. T., and 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.
- Park, P.J. (2009) ChIP–seq: Advantages and challenges of a maturing technology, Nature Reviews Genetics, 10(10), pp. 669–680. doi: 10.1038/nrg2641.
- Pauwels, R.A., Buist, A.S., Calverley, P.M.A., Jenkins, C.R. And Hurd, S.S. (2001) Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease, American Journal of Respiratory and Critical Care Medicine, 163(5), pp. 1256–1276. doi: 10.1164/ajrccm.163.5.2101039.
- Pavelitz, T., Bailey, A.D., Elco, C.P. and Weiner, A.M. (2008) Human U2 snRNA genes exhibit a persistently open Transcriptional state and promoter disassembly at Metaphase, Molecular and Cellular Biology, 28(11), pp. 3573–3588. doi: 10.1128/mcb.00087-08.

- Pearson, T. A., and 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., and McMahon, A. P. (1998) Sonic hedgehog regulates branching morphogenesis in the mammalian lung. Curr Biol, 8(19), 1083-1086.
- Petrache, I. (2009) Safety and efficacy of alpha-1-antitrypsin augmentation therapy in the treatment of patients with alpha-1-antitrypsin deficiency, Biologics: Targets and Therapy, , p. 193. doi: 10.2147/btt.2009.3088.
- 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.
- Ploner A (2015) Heatplus: Heatmaps with row and/or column covariates and colored clusters. R package version 2.18.0, https://github.com/alexploner/Heatplus.
- Pond, K. and Frost, S. (2005) Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Mol Biol Evol 22, 1208-22.
- 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.

- Probert, K., Miller, S., Kheirallah, A.K., Hall, I.P. (2015) Developmental genetics of the COPD lung. COPD Research and Prectice, 1:10.
- Ptashne, M., and Gann, A. (1997) Transcriptional activation by recruitment. Nature, 386(6625), 569-577. doi: 10.1038/386569a0
- Qin, L., Tan, Y., Hu, C., Liu, X. and He, R. (2015) Leptin is Oversecreted by respiratory Syncytial virus-infected bronchial Epithelial cells and regulates th2 and th17 cell differentiation, International Archives of Allergy and Immunology, 167(1), pp. 65–71. doi: 10.1159/000436966.
- Quinlan, A. R., and 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., and 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., and Speizer, F. E. (1987) Assessment of genetic and nongenetic influences on pulmonary function. A twin study. Am Rev Respir Dis, 135(1), 217-222.
- Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K. and Crapo, J.D. (2010) Genetic Epidemiology of COPD (COPDGene) study design, COPD: Journal of Chronic Obstructive Pulmonary Disease, 7(1), pp. 32–43. doi: 10.3109/15412550903499522.
- 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.
- Roberts, A.J., Kon, T., Knight, P.J., Sutoh, K. and Burgess, S.A. (2013) Functions and mechanics of dynein motor proteins, Nature Reviews Molecular Cell Biology, 14(11), pp. 713–726. doi: 10.1038/nrm3667.

- Rockman, M. V., and Kruglyak, L. (2006) Genetics of global gene expression. Nat Rev Genet, 7(11), 862-872. doi: 10.1038/nrg1964.
- Rutkowski, R. J. and Warren, W. D. (2009) Phenotypic analysis of deflated/Ints7 function in Drosophila development. Dev Dyn 238, 1131-9.
- Saitou N. and Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406-425.
- Sambamurthy, N., Leme, A. S., Oury, T. D., and 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., and Joung, J. K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol, 32(4), 347-355. doi: 10.1038/nbt.2842.
- Sandford AJ., Joos L., Pare P.D. (2002) Genetic risk factors for chronic obstructive pulmonary disease. Curr. Opin. Pulm. Med. 8:87–94.
- Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors, Proceedings of the National Academy of Sciences, 74(12), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.
- Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and 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.
- Shen, L., Shao, N., Liu, X. and Nestler, E. (2014) ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. BMC Genomics, 15 284.
- Shin, H., Liu, T., Manrai, A.K. and Liu, X.S. (2009) CEAS: Cis-regulatory element annotation system, Bioinformatics, 25(19), pp. 2605–2606. doi: 10.1093/bioinformatics/btp479.
- Shlyueva, D., Stampfel, G., and 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., and 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
- Smith, H. W., and 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, 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.

- 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., Johnson, T., Zhao, J.H., Albrecht, E., Dominiczak, A.F., Kerr, S.M., Smith, B.H., Cadby, G., Hui, J., Palmer, L.J., Hingorani, A.D., Wannamethee, S.G., Whincup, P.H., Ebrahim, S., Smith, G.D., Barroso, I., Loos, R.J.F., Wareham, N.J., Cooper, C., Dennison, E., Shaheen, S.O., Liu, J.Z., Marchini, J., Dahgam, S., Naluai, Å.T., Olin, A.-C., Karrasch, S., Heinrich, J., Schulz, H., McKeever, T.M., Pavord, I.D., Heliövaara, M., Ripatti, S., Surakka, I., Blakey, J.D., Kähönen, M., Britton, J.R., Nyberg, F., Holloway, J.W., Lawlor, D.A., Morris, R.W., James, A.L., Jackson, C.M., Hall, I.P. and Tobin, M.D. (2011) Effect of Five genetic variants associated with lung function on the risk of chronic obstructive lung disease, and their joint effects on lung function, American Journal of Respiratory and Critical Care Medicine, 184(7), pp. 786–795. doi: 10.1164/rccm.201102-0192oc.
- Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S., Parrinello, H., Cuvier, O. and Benkirane, M. (2014) Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes, Nature Communications, 5, p. 5531. doi: 10.1038/ncomms6531.
- Stewart, C. E., Nijmeh, H. S., Brightling, C. E., and 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., Torr, E.E., Mohd Jamili, N.H., Bosquillon, C. and Sayers, I. (2012) Evaluation of differentiated human bronchial Epithelial cell

culture systems for asthma research, Journal of Allergy, 2012, pp. 1–11. doi: 10.1155/2012/943982.

- Stoller, J. K., and 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.
- Suki, B. (2005) Biomechanics of the lung parenchyma: Critical roles of collagen and mechanical forces, Journal of Applied Physiology, 98(5), pp. 1892–1899. doi: 10.1152/japplphysiol.01087.2004.
- Sun, W., and 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
- Takata, H., Nishijima, H., Maeshima, K. and Shibahara, K. (2012) The integrator complex is required for integrity of Cajal bodies, Journal of Cell Science, 125(1), pp. 166–175. doi: 10.1242/jcs.090837.
- 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.
- Tao, S., Cai, Y. and Sampath, K. (2009) The integrator subunits function in hematopoiesis by modulating Smad/BMP signaling, Development, 136(16), pp. 2757–2765. doi: 10.1242/dev.034959.
- Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z.,

Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E. and Stamatoyannopoulos, J.A. (2012) The accessible chromatin landscape of the human genome, Nature, 489(7414), pp. 75–82. doi: 10.1038/nature11232.

- Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L. and Pachter, L. (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq, Nature Biotechnology, 31(1), pp. 46–53. doi: 10.1038/nbt.2450.
- Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nature Biotechnology, 28(5), pp. 511–515. doi: 10.1038/nbt.1621.
- Uguen, P. (2003) The 3 ends of human pre-snRNAs are produced by RNA polymerase II CTD-dependent RNA processing, The EMBO Journal, 22(17), pp. 4544–4554. doi: 10.1093/emboj/cdg430.
- 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 den Borst, B., Souren, N.Y.P., Loos, R.J.F., Paulussen, A.D.C., Derom, C., Schols, A.M.W.J. and Zeegers, M.P. (2012) Genetics of maximally attained lung function: A role for leptin?, Respiratory Medicine, 106(2), pp. 235–242. doi: 10.1016/j.rmed.2011.08.001.
- Van Durme, Y. M., Eijgelsheim, M., Joos, G. F., Hofman, A., Uitterlinden, A. G., Brusselle, G. G., and 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.
- Villar Álvarez, F., Troncoso Acevedo, M.F. and Peces-Barba Romero, G. (2009) Evaluation of COPD Longitudinally to identify predictive

surrogate end-points (ECLIPSE), Revista de Patología Respiratoria, 12(1), pp. 48–49. doi: 10.1016/s1576-9895(09)70092-x.

- Visscher, P. M., Brown, M. A., McCarthy, M. I., and 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.V., Shrine, N., Miller, S., Jackson, V.E., Ntalla, I., Artigas, M.S., Billington, C.K., Kheirallah, A.K., Allen, R., Cook, J.P., Probert, K., Obeidat, M., Bossé, Y., Hao, K., Postma, D.S., Paré, P.D., Ramasamy, A., Mägi, R., Mihailov, E., Reinmaa, E., Melén, E., O'Connell, J., Frangou, E., Delaneau, O., Freeman, C., Petkova, D., McCarthy, M., Sayers, I., Deloukas, P., Hubbard, R., Pavord, I., Hansell, A.L., Thomson, N.C., Zeggini, E., Morris, A.P., Marchini, J., Strachan, D.P., Tobin, M.D. and Hall, I.P. (2015) 'Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): A genetic association study in UK Biobank', The Lancet Respiratory Medicine, 3(10), pp. 769–781. doi: 10.1016/s2213-2600(15)00283-0.
- 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, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: A revolutionary tool for transcriptomics, Nature Reviews Genetics, 10(1), pp. 57–63. doi: 10.1038/nrg2484.
- Wang, K., Li, M., and Hakonarson, H. (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet, 11(12), 843-854. doi: 10.1038/nrg2884
- Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., Cheng, Y., Rando, O.J., Birney, E., Myers, R.M., Noble, W.S., Snyder, M. and Weng, Z. (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors, Genome Research, 22(9), pp. 1798–1812. doi: 10.1101/gr.139105.112.

- Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., Tang, Q., Meyer, C. A., Zhang, Y. and Liu, X. S. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protoc, 8, 2502-15.
- Wang, C., Gong, B., Bushel, P.R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z., Meehan, J., Li, X., Yang, L., Li, H., Łabaj, P.P., Kreil, D.P., Megherbi, D., Gaj, S., Caiment, F., van Delft, J., Kleinjans, J., Scherer, A., Devanarayan, V., Wang, J., Yang, Y., Qian, H.-R., Lancashire, L.J., Bessarabova, M., Nikolsky, Y., Furlanello, C., Chierici, M., Albanese, D., Jurman, G., Riccadonna, S., Filosi, M., Visintainer, R., Zhang, K.K., Li, J., Hsieh, J.-H., Svoboda, D.L., Fuscoe, J.C., Deng, Y., Shi, L., Paules, R.S., Auerbach, S.S. and Tong, W. (2014) The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance, Nature Biotechnology, 32(9), pp. 926–932. doi: 10.1038/nbt.3001.
- Ward, L. D., and 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.
- 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., and 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.
- Williams, C.R., Baccarella, A., Parrish, J.Z. and Kim, C.C. (2016) Trimming of sequence reads alters RNA-Seq gene expression estimates, BMC Bioinformatics, 17(1) doi: 10.1186/s12859-016-0956-2.
- 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.
- Wong, Y. Y., Moon, A., Duffin, R., Barthet-Barateig, A., Meijer, H. A., Clemens, M. J. and De Moor, C. H. (2010) Cordycepin inhibits protein synthesis and cell adhesion through effects on signal transduction. J Biol Chem 285 2610-21.
- Wood, A.M. and Stockley, R.A. (2006) Respiratory Research, 7(1), p. 130. doi: 10.1186/1465-9921-7-130.
- Wray, N., Visscher, P. (2008) Estimating Trait Heritability. Nature Education. 1 (1): 29.
- Xie, X. H., Law, H. K., Wang, L. J., Li, X., Yang, X. Q. and Liu, E. M. (2009) Lipopolysaccharide induces IL-6 production in respiratory syncytial virus-infected airway epithelial cells through the toll-like receptor 4 signalling pathway. Pediatr Res, 65, 156-62.
- Xu, S., Grullon, S., Ge, K., and 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
- Yamamoto, J., Hagiwara, Y., Chiba, K., Isobe, T., Narita, T., Handa, H. and Yamaguchi, Y. (2014) DSIF and NELF interact with integrator to specify the correct post-transcriptional fate of snRNA genes, Nature Communications, 5. doi: 10.1038/ncomms5263.
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W. and Liu,

X.S. (2008) Model-based analysis of chIP-seq (MACS), Genome Biology, 9(9), p. R137. doi: 10.1186/gb-2008-9-9-r137.

- Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. and Liu, X. (2014) Comparison of RNA-Seq and Microarray in Transcriptome profiling of activated T cells, PLoS ONE, 9(1), p. e78644. doi: 10.1371/journal.pone.0078644.
- Zhou, J. J., Cho, M. H., Castaldi, P. J., Hersh, C. P., Silverman, E. K., and 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.