The Idiopathic Pulmonary Fibrosis-associated single nucleotide polymorphism rs35705950 is transcribed in a MUC5B Promoter Associated Long Non-Coding RNA (AC061979.1)

Neatu R.[2], Thompson D.J.[1], Enekwa I.[1], Schwalbe E.C.[1], Fois G.[3], Frick M.[3], Braubach P.[4], Moschos S.A.[1,✉]

Abstract

LncRNAs are involved in regulatory processes in the human genome, including gene expression. The rs35705950 SNP, previously associated with IPF, overlaps the recently annotated lncRNA AC061979.1, a 1,712 nucleotide transcript located within the MUC5B promoter at chromosome 11p15.5. To document the expression pattern of the transcript, we processed 3.9 TBases of publicly available RNA-SEQ data across 27 independent studies involving lung airway epithelial cells. Epithelial lung cells showed expression of this putative pancRNA. The findings were independently validated in cell lines and primary cells. The rs35705950 is found within a conserved region (from fish to primates) within the expressed sequence indicating functional importance. These results implicate the rs35705950-containing AC061979.1 pancRNA as a novel component of the MUC5B expression control minicircuitry.

✉ sterghios.moschos@northumbria.ac.uk

[1]**Faculty of Life and Health Sciences, Department of Applied Sciences, Northumbria University, Ellison Building, Newcastle-Upon-Tyne, Tyne & Wear NE1 8ST, UK** [2]**Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Central Parkway Newcastle Upon Tyne, NE1 3BZ, UK** [3]**Institue of General Physiology, University of Ulm, Albert-Einstein-Allee 11,D-89081 Ulm, Germany** [4]**Institute of Pathology, MHH Hannover, Hannover, Germany**

1 Introduction

Human DNA consists of protein-coding regions and non-coding regions. Protein-coding genomic regions are abundantly transcribed, evolutionary conserved, mutationally sensitive sequences that impact cellular phenotype. These constitute approximately 1% of the human genome [1]. Non-coding regions of DNA, on the other hand, are more complex and can be divided into at least five structural types: (i) binding motifs for regulatory proteins, (ii) non-coding RNAs (ncRNAs), (iii) transposable elements, (iv) highly repetitive DNA - essential in gene regulation and chromosome maintenance, and (v) pseudogenes [2,3]. In 2003, the ENCODE (Encyclopedia of DNA Elements) project was launched to identify and classify functional elements in the human genome including non-coding transcripts. The project continues to grow with the results being made available on Ensembl and UCSC genome browser for both human and mouse [4].

Most of the ncRNAs predicted by ENCODE are expressed at low levels [4]. However, their abundance is not a proxy for their functionality [5]. For example, the predicted lncRNA ENST00000567151 or Viability Enhancing in LUng Cancer Transript (VELUCT) was found at only 0.01 copies per cell. Despite it's low copy number, VELUCT expression was reported to be upregulated by 5.2-fold in lung cancer cells, and its knockdown to reduce the viability of multiple lung cancer cell lines as much as 90% [6].

Recently, lncRNAs (long ncRNAs) have been intensively studied due to their involvement in cancer [7], neurological conditions [8], pulmonary dis-

eases [9], and the regulation of chromosome structure [10]. This research culminated in several published lncRNA databases: NONCODE [11], Lnc2Cancer [12], LncRNADisease [13], and LncRNAdb [14]. LncRNAs can regulate expression via at least two mechanisms: *cis*-acting lncRNA (which regulate expression of adjacent genes) and *trans*-acting lncRNAs (regulating the expression of distant genes on other chromosomes) [15]. Most pancRNAs (promoter associated ncRNAs) to date have been associated with an increased production of mRNA from the adjacent protein-coding gene, suggesting that pancRNAs might contribute to gene expression regulation. Protein-coding genes that possess pancRNAs also exhibit tri-methylated lysine 4 of histone 3 (H3K4me3) and acetylated lysine 27 of histone histone 3 (H3K27ac) whereas the pancRNA-free genes appear to lack such epigenetic signatures [16]. However, how pancRNAs change the expression of protein-coding genes remains unknown.

Furthermore, evidence is amassing concerning the expression of pancRNAs and the occurrence of epigenetic changes. Thus, typically, when single nucleotide polymorphism (SNPs) appear in such non-coding transcript loci, the associated pancRNA secondary structure is disrupted, affecting expression patterns and impacting upon the function [17]. Whilst expression changes in high copy number lncRNA are easy to determine by routine RNA-SEQ, the effect of SNPs resulting in small changes in lncRNA expression levels is harder to study.

The G/T rs35705950 SNP found in the promoter of mucin 5B (MUC5B) on chromosome 11p15.5 [18, 19] has one of the highest ($\sim 40\%$) [20] and most reproducible associations with Idiopathic Pulmonary Fibrosis (IPF) across white, hispanic, and Asian populations [21–37], with homozygous mutants exhibiting higher risk of developing the disease [38] and higher mortality [39]. The polymorphism is implicated in the elevated transcription and translation of MUC5B in both healthy and diseased individuals [40]. This is evidenced via episomal expression of luciferase driven by TT or GG MUC5B promoters cloned from IPF patients in A549 alveolar epithelial cells [41]. Since MUC5B is one of the largest proteins encoded in the human genome, excessive expression is proposed to lead to elevated endoplasmatic reticulum (ER) stress [42] through MUC5B protein recycling and the unfolded protein response, increasing cell sensitivity to exogenous insults and pro-apoptotic phenotypes. This is exacerbated in alveolar lung epithelia where MUC5B aberrant mRNA expression is elevated but MUC5B protein production is not normally observed. Presently, the polymorphism is thought to a) disrupt a 25 CpG motif differential methylated region which is, counterintuitively, hypermethylated in IPF, and b) enhance the binding of the transcription factor Forkhead Box Protein A2 (FOXA2), 32 bp downstream of the SNP as evidenced by chromatin immunoprecipitation [18]. Given the distal effect of the SNP to the FOXA2 binding site and the emerging role of pancRNA in transcription regulation, we sought to determine whether a lncRNA transcript might be implicated in MUC5B expression and its transcriptional dysregulation in the context of the rs35705950 SNP.

To this end we analysed publicly deposited RNA-SEQ datasets. However, most pipelines for novel transcript discovery are focused on small RNA populations or certain RNA species [43], and RNA-SEQ workflows typically involve polyadenylated transcript enrichment. This creates a classical signal to noise ratio detection problem where selective signal acquisition and amplification during sequencing library preparation may reduce non-polyadenylated transcript read frequencies to levels typically ascribed to background noise. Inspired by the application of very long base interferometry in expanding observation dynamic range beyond standard signal to noise ratio limitations through signal integration from multiple sources operating similar data acquisition protocols [44], we applied composite analysis of 3 rd party RNA-SEQ datasets to reveal the existence of such technically occluded transcripts. Overall, we describe a novel and simple computational method for performing such *de novo* lncRNA transcript searches by aggregating data from diverse input sources, and focusing analytical efforts in the RNA-SEQ-verse to specific genomic regions of interest.

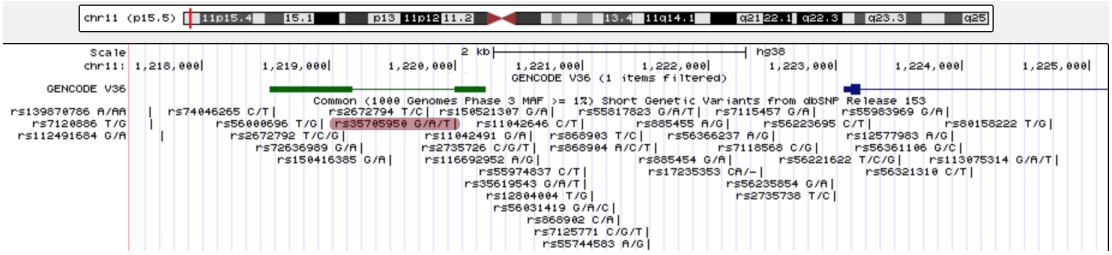

Figure 1: **UCSC genome browser Genomic location of the annotated lncRNA AC061979.1.** The putative pancRNA AC061979.1 - green; the transcription start site of MUC5B - dark blue: thick line - exons; thin line - introns; rs35705950 - highlighted in red.

2 Materials and Methods

2.1 RNA-SEQ Data Processing for novel ncRNA detection

To determine the existence of a MUC5B pancRNA, we manually collected publicly deposited RNA-SEQ data from 27 independent studies involving alveolar and bronchial samples from primary human tissue and *in vitro* experiments (see Supplementary Data 1). RNA-SEQ reads were mapped to the human reference genome GRCh38.p13 using HISAT2 [45]. Mapped reads were filtered with samtools view [46] and only read pairs mapping to chromosome 11, region 1,202,000-1,220,500 were kept. Subsequently the depth of coverage per base was extracted from all datasets and collapsed. The results were visualised in R Studio (ggplot2). The pipeline can be performed in Galaxy (galaxyproject.org). An extensive step-by-step guide is available as Supplementary Data 2.

2.2 Multiple Sequence Alignment

To demonstrate the evolutionaty importance of the region overlapping the promoter polymorphism rs35705950, we compared the human ncRNA with nucleotide sequences of 10 other species from fish to primates. Rhesus monkey (*Macaca mulatta*), baboon (*Papio anubis*), white-tufted-ear marmoset *Callithrix jacchus*, pig (*Sus scrofa*), sheep (*Ovis aries*), Norvegian rat (*Rattus norvegicus*), house mouse (*Mus musculus*), chicken (*Gallus gallus*) and zebrafish (*Danio rerio*).Alignments of genome sequences were undertaken using AVID and Shuffle-LAGAN programs implemented through mVISTA http://genome.lbl.gov/vista/ mvista/submit.shtml [47] with a match criterion of 70% identity over 50bp [48]. All sequences used in analysis are included in Supplementary Data 3. Subsequently we aligned the same genomic sequences with ClustalO for nucleotide by nucleotide approach.

2.3 Cell culture

A549 cells passage 10-12 were thawed, seeded on t25 flasks at 37°C 5% CO_2 in DMEM/F12 (1:1) (ThermoFisher Scientific, Cramlington, UK) with 10%FBS, + 1% L-Glutamine, 1% Pen/Strep (Merck Life science UK limited, Dorset, UK). Cells were cultivated till 50-60% confluency, then splitted in other t25 flasks until 20-30% confluency was reached (usually 24h). CFBE41o cells (passage 10-12) were thawed, then seeded on t25 flasks at 37°C, 5% CO2 in MEM (Merck Life science UK limited) with 10% FBS, 1% L-Glutamine, 1% Pen/Strep (Merck Life science UK limited). Cells were cultivated till 50-60% then seeded into new t25 flasks until 20-30% or 50-60% confluency was reached for subsequent low confluency or high confluency total RNA extractions respectively.

Primary human airway epithelial primary cells (pHAECs) from several donors (N=1 basal cells, N=4 ALI differentiated cells) were isolated from fresh tissues that were obtained during tumor resections or lung transplantation with fully consent of patients (Ethics approval: ethics committee Medical School Hannover, project no. 2701-2015).

pHAECs basal cells (passage 4) were cultivated on T75 Flasks in Airway Epithelial Cell Basal Medium supplemented with Airway Epithelial Cell Growth Medium Supplemen tPack and with 5 μg/ml Plasmocin prophylactic, 100 μg/ml Primocin and 10 μg/ml Fungin (all from InvivoGen, Toulouse, France). Trypsinization with Promocell DetachKit (Promocell, Heidelberg, Germany) and RNA etxraction was performed at ∼40-50% confluency.

pHAECs basal cell for air liquid interface (passage 2) were expanded as above in T75 flasks till 90% confluency. The cells were than trypsinized and seeded into Transwell filters (6.5 mm diameter, 4 μm pore size,

Corning Costar, Kaiserslauten, Germany). Filters, prior to cell seeding, were coated with 100 μl Collagen Solution (StemCell Technologies, Saint Égrève, France), left to dry under sterile hood overnight. Subsequently the filters were exposed to UV light for 30 min and stored at 4°C.

Cells were resuspended in Growth medium, and 200 μl containing 4x 10^4 cells were added apically to each filter, additional 600µl medium were added basolaterally. The medium was replaced every 48h until 100% confluence was reached. Growth medium was then removed from apical side and on the basolateral side it was replaced by ALI differentiation medium +/- 10 ng/ml IL-13 (IL012; Merck Millipore). Once the ALI interface was established medium was exchanged every second day till day 25-28 on ALI. At the endpoint of cultivation RNA extraction was performed directly on the filter.

2.4 RNA Extraction

RNA extraction was carried out using the miRNeasy mini kit (Qiagen, Manchester, UK). Briefly cells were detached by trypsinisation then resuspended in 0.7 mL Qiazol Lysis reagent with subsequent steps according to the supplier's total RNA extraction protocol.

For pHAECs cells were detached by trypsinisation then resuspended in 2.1 ml Lysis Solution RL from my-Budget RNA Mini Kit (BioBudget, Krefeld, Germany), RNA isolation was done following the manufacturer protocol. If not used immediately after lysis the samples were stored at -80°C. For pHAECs ALI cultures 100 μl of Lysis Solution RL from the same kit was added to the filters apically and the samples were immediately frozen at -80°C.

2.5 DNase Treatment and cDNA synthesis

Total RNA was DNase treated using the PrecisionTM DNase kit Primer Design (Southampton, UK), following the manufacturers protocol. cDNA synthesis was carried out using the High Capacity cDNA reverse Transcription Kit (Thermo Fisher scientific (ThermoFisher Scientific) following the manufacturers protocol. 1000ng of total RNA was loaded into each 20μL cDNA synthesis reaction.

For pHAEC cells the cDNA synthesis from the extracted was performed using SuperScript VILO cDNA Synthesis Kit (Thermo Fisher) following the manufacturer protocol. 400ng of RNA were used for each reaction.

2.6 Real-time Quantitative PCR

Custom Primers and probes (Supplementary data 4) were designed using the PrimerQuestTM tool (Integrated DNA Technologies BVBA, Leuven, Belgium) and validated against an AC061979.1 geneblock (Integrated DNA Technologies) corresponding to the predicted spliced transcript. Inventoried predesigned assays for 18S and MUC5B were purchased from Thermo Fisher Scientific. Real-time quantitative PCR was performed in 10 μL reactions containing 5L TaqMan Fast Advanced Master Mix (2x) (ThermoFisher Scientific), 900nM forward primer, 900nM reverse primer and 250nM probe per reaction and 1μL template on a StepOnePlusTM real-time PCR system (Thermo Fisher Scientific). After a UNG incubation at 50°C for 2 minutes Initial denaturation at 95°C for 2 minutes was followed by 40 cycles of 95°C denaturation for 1 second and 60°C anneal-extension for 20 seconds. Gene expression was calculated according to the delta Ct method [49].

3 Results

3.1 Rediscovery of the non-coding transcript AC061979.1 in the promoter region of the *MUC5B* gene

To identify a putative non-coding transcript in a "dark" intergenic region on the p-terminus of chromosome 11 on the human genome in the context of lung epithelia, we manually collected and interrogated a total of 3.9 TBases of publicly available RNA-SEQ data involving epithelial lung cells (see Supplementary Data 1) to generate a summative transcriptional signal of the AC061979.1 locus (see figure 2, A). Concomitantly, the GENCODE [50] release 32 (GRCh38) described the putative transcript AC061979.1 (chromosome 11:1,218,530-1,220,242) mapping to the same region (see figure 1). Interestingly, this novel lncRNA is reported to be subject to splicing, with rs35705950 mapping to the 2nd nucleotide of exon 2 and therefore possibly altering AC061979.1 splicing; however, no evidence of splicing was immediately apparent through our analysis (see figure 2) and no rs35705950-containing RNA-SEQ data from lung epithelia were found among the surveyed studies.

Manual inspection of each dataset indicated that almost all samples representing lung epithelial cell lines

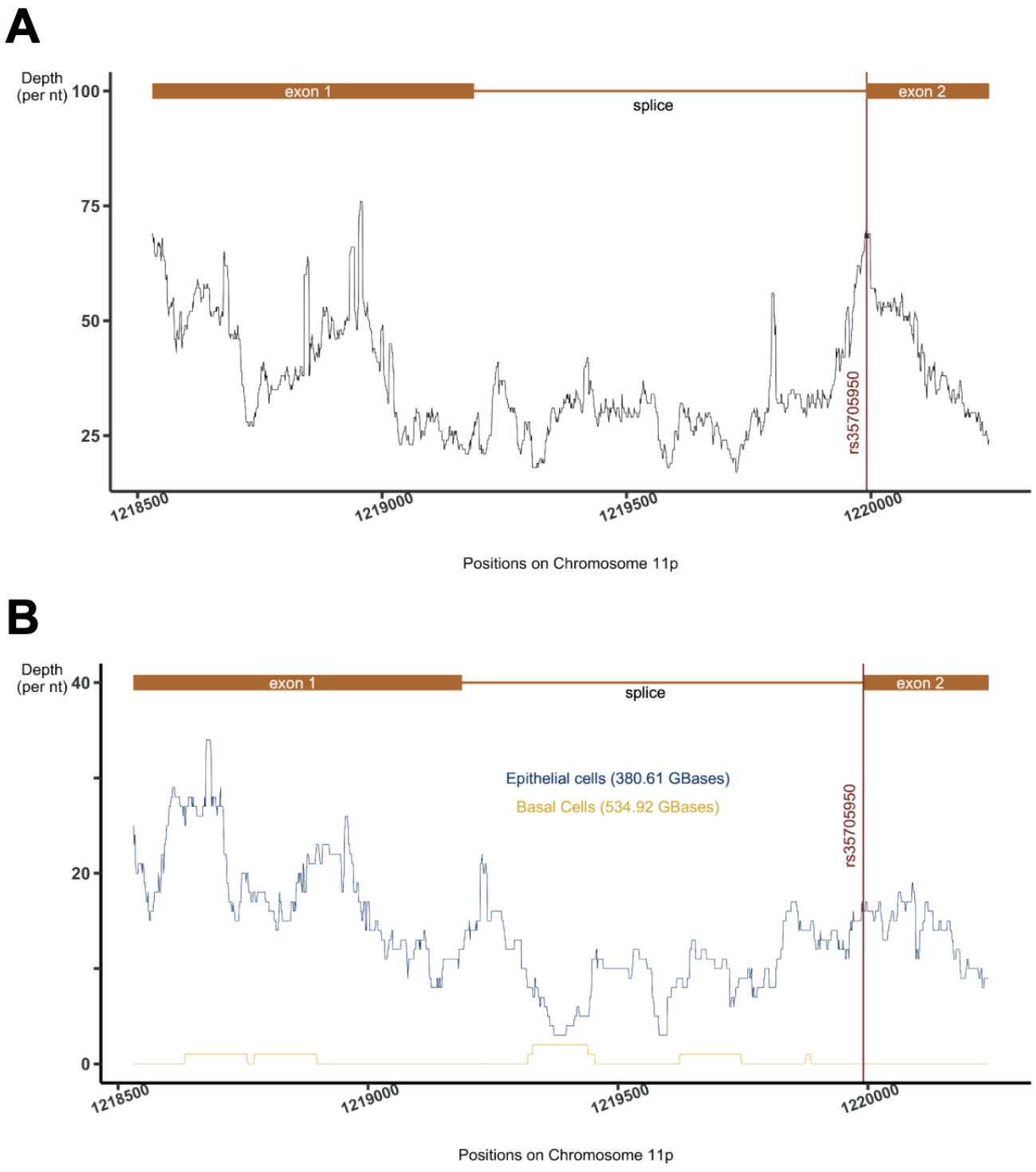

Figure 2: **RNA-SEQ data processing results.** Depth of coverage spanning chromosome 11: 1,218,530-1,220,242 collected from (A) 27 studies (3.9 TBases) and (B) a single dataset (SRP082973) comparing epithelial to basal cell expression. The position of rs35705950 is indicated by a red vertical line and the AC061979.1, primary transcript and spliced exons are indicated in orange.

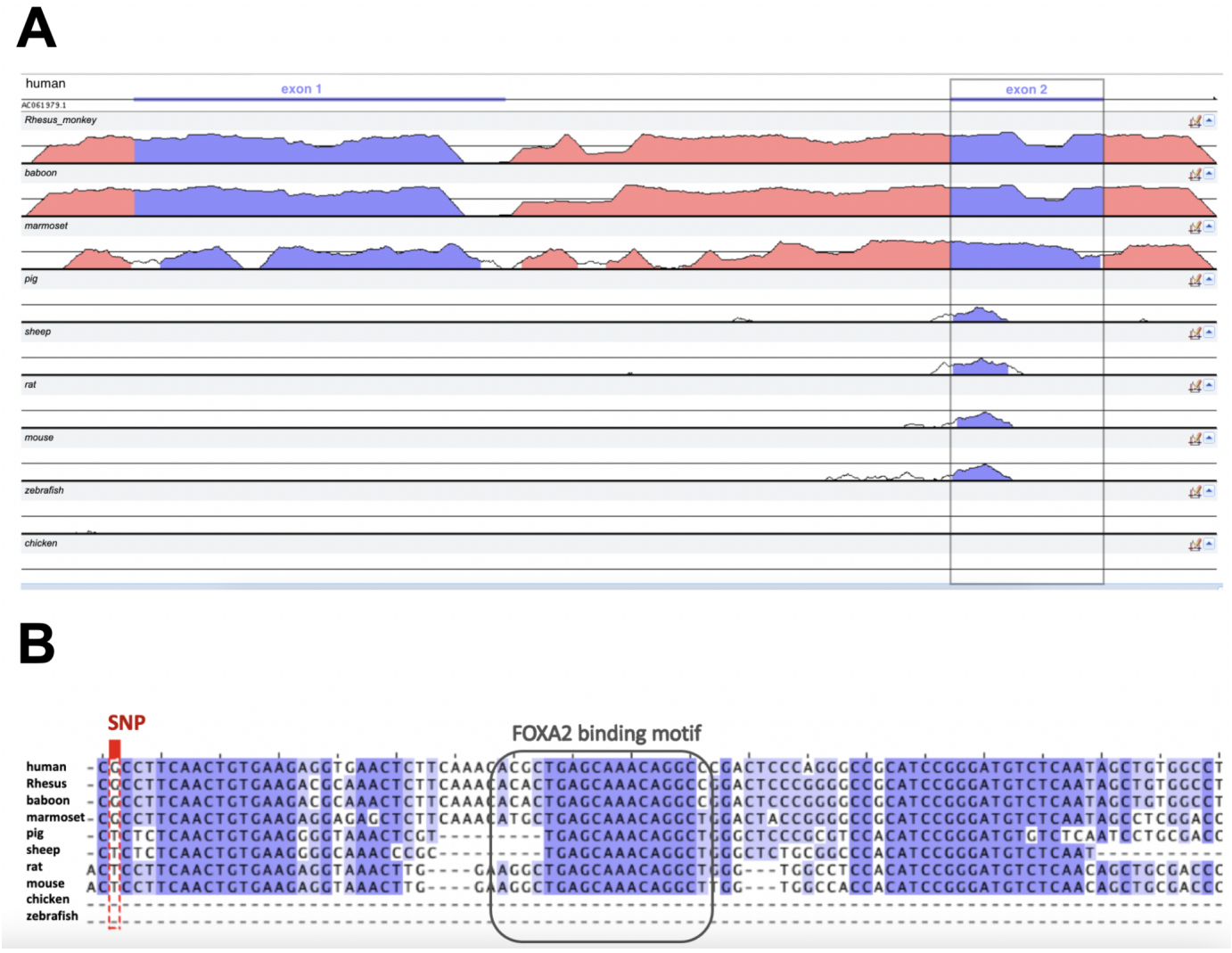

Figure 3: **Conservation of the MUC5B-MUC5AC intergenic region across 10 species.** **A.** The genomic sequences were aligned using AVID in mVISTA: global pair-wise alignment between ∼2,000 nt spanning the human AC061979.1 transcript and the whole intergenic region of the other species (∼20,000 nt; Supplementary Data 3). Coloured peaks (purple: AC061979.1 exons; pink: intergenic regions) indicate at least 50bp with 70% similarity. The grey rectangle indicates the conserved exon across mammals. **B.** Multiple Sequence Alignment by ClustoalO in Jalview shows that 100 nucleotides downstream of rs35705950 (red rectangle) there are i) 15-25 bp conserved across mammals (purple shades by nucleotide similarity percentage), ii) a FOXA2 binding site (grey rectangle), and iii) a third conserved region approximately 10 nt downstream of the FOXA2 binding site.

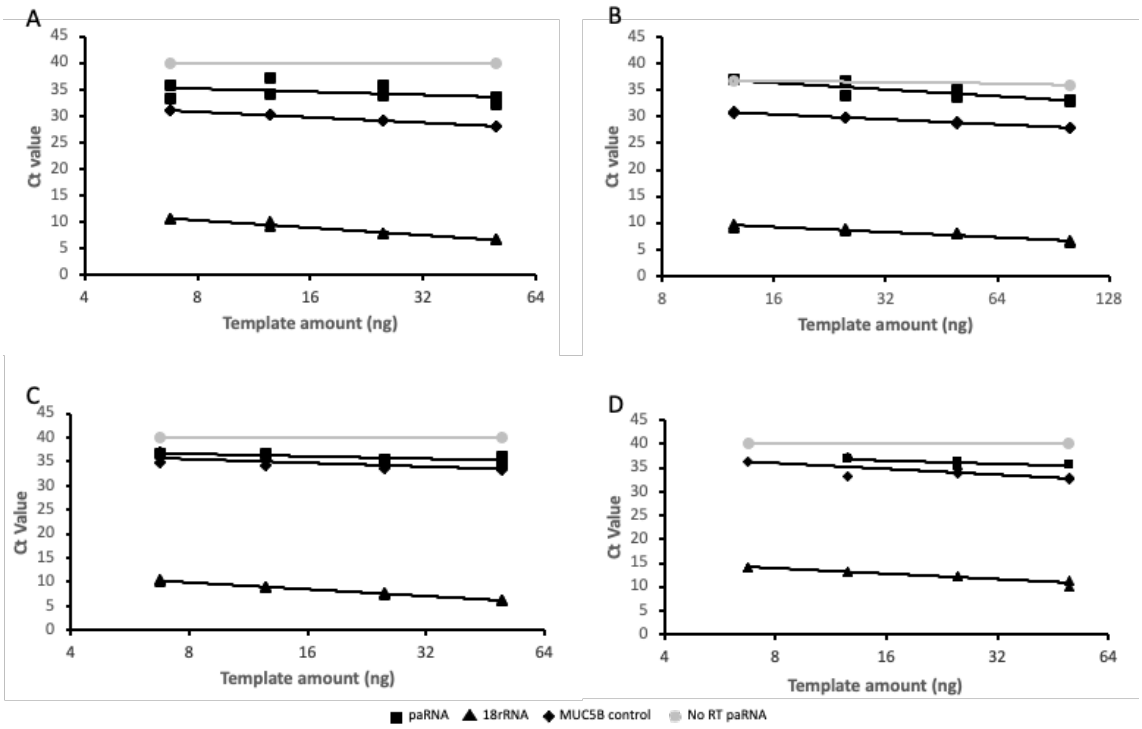

Figure 4: **AC061979.1 expression validation in cell culture.**Two-fold serial dilutions of A549 (A, B) and CFBE41o- (C, D) RNA extracts obtained from low (<30%; A, C) and high (>70%; B, D) confluence cells were analysed by probe hydrolysis RT-qPCR for 18S rRNA (triangle), MUC5B (diamond) and AC061979.1 paRNA (square) expression across 2-fold serial dilutions, with a no RT of AC061979.1 reaction set included as a negative control (grey circles). Data are expressed in log2-linear scale and are representative of 3 independent biological experiments and dual technical replicate Ct's points shown.

or primary lung epithelial cells showed evidence of expression in the locus. Of particular interest, however, was a dataset (SRP082973) that isolated only basal cells from the epithelium. Thus, within the same study, we compared basal cells and the epithelium (mix of basal, ciliated, columnar and secretory cells) RNA expression [51]. Interestingly, whilst reads from epithelial cell extracts mapped liberally to the AC061979.1 locus, sporadic alignments of only a couple of reads were detected among basal cell RNA 2). Given that basal cells are a sub-type of human airway epithelial cells not involved in mucous production, that act as stem cells for the other sub-types (ciliated, columnar and secretory cells) [52], these results potentially show the activation of ncRNA AC061979.1 after cell differentiation. Taken together these results suggest that expression at AC061979.1 is detectable in lung epithelia irrespective of the biological origin of the data or the precise sequencing protocol used, minimising the risk of batch-associated effects.

To determine the evolutionary importance of the DNA sequence harbouring the rs35705950 polymorphism, the human reference genome (GRCh38.p13) was aligned against nine vertebrate genome seuqences: six mammals (Rhesus monkey, baboon, marmoset, pig, sheep, rat, mouse), one fish (zebrafish) and one bird (chicken). High similarity was observed in exonic regions across primates, with phylogenetically distant mammals showing conservation only at the 5' end of the second exon (see figure 3A). This region appears to harbour at least 3 conserved loci, including a FOXA2 binding motif, found across mammalian species (see figure 3B).

Although AC061979.1 transcript abundance in other species is limited by the lack of RNA-SEQ datasets to interrogate, taken together these results indicate a functional significance for this pancRNA, with the G/T rs35705950 SNP possibly being involved in differential splicing of AC061979.1.

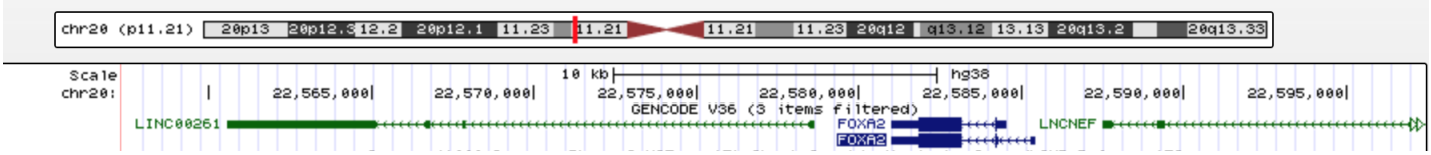

Figure 5: **UCSC genome browser Genomic location of the annotated lncRNA LINC00261 and FOXA2.** LINC00261 and lncRNA neighboring enhancer of FOXA2 (LNCNEF) - green; FOXA2 - dark blue.

3.2 MUC5B pancRNA expression validation

To independently validate the expression of AC061979.1 we first designed probe hydrolysis RT-qPCR assays for the putative spliced variant and holotranscript. These assays exhibited amplification efficiencies of 128.8% and 90.6% when tested against serial dilutions of a spliced AC061979.1 geneblock or A549 cell DNA extracts, respectively. Next, we obtained RNA extracts from adenocarcinoma human alveolar type II epithelial cells (A549 cells) and cystic fibrosis bronchial epithelial cells (CFBE41o) representing alveolar and bronchial epithelial cells, respectively. To account for the potential impact of contact inhibition effects on AC061979.1 expression, total RNA was extracted at low (<30%) and high (>70%) confluence, and expression of the two AC061979.1 variants was assessed against 18S rRNA and MUC5B, across serial dilutions of total RNA. These analyses indicated that only the AC061979.1 full transcript was detectable albeit at very low copy number (see figure 4). Thus, at 50 ng of RNA input per RT-PCR reaction, in A549 cells the paRNA ΔCt to 18S was 22.44 (+/-5.94) at high confluence vs 24.12 (+/- 3.16) at low confluence, whereas in CFBE41o- the ΔCt was 21.92 (+/-5.53) at high confluence vs 28.59 (+/- 0.81) at low confluence (n=3). Of note, where RNA extraction resulted in higher Ct values for 18S, the capacity to detect the paRNA transcript was lost as concentrations dropped below the assay limit of detection, justifying the very high load of RNA template in the RT-PCR reactions.

The same assays were perfomed with undifferentiated (basal) and ALI differentiated HAEpCs. The basal cells were cultivated to a confluence of 40-50% before extracting total RNA. Of the two AC061979.1 transcripts only the full variant was detectable (36.19 +/- 0.29), whereas the spliced variant was not detectable. Interestingly MUC5B levels were below the assay detection limit, 18S was detected at a Ct of 10.63 (+/-0.14). ALI differentiated HAEpCs in control conditions showed similar values to the basal cells (36.84 +/-1.38) whereas in IL-13 stimulated cells both AC061979.1 variants were below detection limit. In differentiated epithelia Cts for 18S were low for both controls and IL-13 treated cells (7.97+/-0.06 and 8.53+/-0.08 respectively). Similarly, to the basal cells MUC5B was not detectable.

4 DISCUSSION

MUC5B dysregulation presently appears to be mechanistically involved in the development of the underlying pathology particularly in the context of the IPF-associated SNP rs35705950. It contributes to mucus overproduction and expression in the alveolar microenvironment, leading to micro-injuries to alveolar epithelium and, across the lifetime of a carrier, excessive cell death and fibrosis [53]. Whilst in one study the polymorphism was found in 51% of the patients with IPF, but in only 23% of the control group [18], it is unclear at present if onset of disease among rs35705950 positive controls is a matter of time, lifestyle, or additional genetic variability. However, the strong association and high incidence rate of the polymorphism in IPF make a compelling case for lifestyle management and preventative chronic or genome modifying treatments targeting MUC5B overexpression, such as small interfering RNA, antisense or genome/prime editing.

Helling *et al.* (2017) reported a binding motif for FOXA2, located 32 bp downstream of rs35705950, which overlaps the second putative exon of the pancRNA AC061979.1 as reported in GENCODE v32 (see figure 1). The protein-coding gene for FOXA2 originates on chromosome 20, p11.21, between the lncRNA LINC00261 and LNCNEF. Whilst LINC00261 (a.k.a. DEANR1 [54], FALCOR [55], LCAL62 [56]) is widely studied for its role in non-small cell lung cancer, no study exists on LNCNEF to date. LINC00261 is an endoderm-associated lncRNA that recruits SMAD2/3 to induce the expression of FOXA2 [54, 56]. FOXA2 transcription factor is known to have a role in lung development and home-

ostasis [57], MUC5B expression and IPF [58] however research in this area is limited.

Our RNA-SEQ-verse survey did not return an explicit splicing signal in line with RNA-SEQ observations associated with high copy number RNAs, however if the proposed splicing event is confirmed, the location of the SNP raises the possibility that aberrant AC061979.1 splicing might be occurring in the context of the rs35705950 SNP. In turn, improper AC061979.1 splicing could be driving aberrant biochemistry on the locus such as the FOXA2 association demonstrated by Helling *et al.* (2017). Such a finding would introduce the additional option of a splice-correcting treatment in preventing the onset of IPF among rs35705950 carriers. Importantly, this oligonucleotide therapeutic modality has been approved for clinical use in Duchenne's Muscular Dystrophies (eteplirsen) and spinal muscular atrophy (nusinersen) without the need for drug delivery solutions that otherwise plague efficacious oligonucleotide therapies for the lung [59].

LncRNAs can form complex biological systems by binding to other RNA molecules, regulatory proteins, or DNA. FENDRR is a lncRNA expressed in the nascent lateral mesoderm, in the promoter of Forkhead Box F1 (FOXF1) where it forms a triple helix with double-stranded DNA and increases the occupancy of the Polycomb repressive complex 2 (PRC2) at this site. Rescue experiments on FENDRR-knockdown cells wherein a construct expressing the lncRNA was placed randomly in the genome showed its biological role and, that the transcript acts in *trans* [60]. Similarly, LINC00261-null cells were rescued by viruses expressing FOXA2, in the transcriptional activation of FOXA2, which is upstream of LINC00261 [54]. It is thus possible that the mechanism behind MUC5B regulation involves an assembly between the pancRNA AC061979.1 and other regulatory proteins or transcripts interacting with the promoter region of MUC5B acting in *cis* or in *trans*, including competitive binding of FOXA2 or SMAD2/3. Although Helling *et al.* (2017) did not asses the importance of SMAD2/3 in MUC5B expression, Feldman *et al.* (2019) showed that phosphorylated SMAD levels are low in mucosecretory cells, and the inflammatory TGF-beta-dependendent SMAD signaling inhibition enhanced mucin expression, as well as goblet cell metaplasia and hyperplasia, supporting a role for SMAD proteins in MUC5B expression regulation [61]. Interactions with SMAD2/3 in the promoter of MUC5B and AC061979.1 are indeed possible due to the presence of the canonical SMAD Binding Element (SBE) CAGAC within the intronic region of the pancRNA, and the newly described GGC(GC)(CG) motif also known as 5GC SBE [62] within the first exon of AC061979.1 [61]. Moreover, SMAD2/3 does not necessarily need to occupy either of these SBEs on chromatin, because SMAD2/3 does not occupy the SBEs located within the LINC00261 gene [54]: instead, it interacts with LINC00261directly at least under some experimental conditions [55]. LINC00261 is, therefore, an example cis-acting ncRNA, whereas other lncRNAs such as EMT-associated lncRNA induced by TGF*beta*1 (ELIT-1) act in trans to bind SMAD to SMAD Binding Elements (SBEs) such as the CAGAC box [63]. Disruption of this proposed SMAD2/3, AC061979.1, and possibly FOXA2 ribonucleoprotein and chromatin interaction network at the MUC5B promoter by rs35705950, for example due to aberrant splicing, could explain MUC5B overexpression in IPF, given the pivotal role of SMAD proteins in resolving goblet cell metaplasia and hyperplasia in inflammatory pulmonary disease [61].

To date, the FOXA2 binding site 32 bp downstream of rs35705950 has been shown to bind FOXA2 in episomal reporter systems but not by genome editing or CHIP-SEQ [18]. Our own genome editing efforts with 3 separate single guide RNAs to introduce the rs35705950 G/T transversion at Chr11:1,219,991 in A549 cells in support of CHIP-SEQ, RIP-SEQ, and proteomic experiments to resolve the MUC5B transcriptional complex have so far proven to irreparably affect cell viability or fail in generating any detectable editing either by T7-EI or sequencing assays. Furthermore, no verified G/T or T/T lung epithelial cell line is currently available to support such mechanistic studies. As lncRNA-protein interaction research is a hot research topic, recent studies have focused on developing computational methods for predicting these complex networks [64–67]. It is thus anticipated that with increasing understanding of lncRNA biology and characterisation of lncRNA structures and families, additional insights in AC061979.1 function might be obtained.

In this study, we have developed a simple-to-use method for the targeted mining of the RNA-SEQ dataverse for lncRNA transcripts irrespective of their polyadenylation status. Our method is achievable on a public server in Galaxy (galaxyproject.org) with an extensive easy-to-follow guide available (see Supplementary information). It takes as input Sequence Read Archive (SRA) codes and the output is a .TXT file reporting the depth of coverage per position mak-

ing end user memory requirements compatible with standard desktop/laptop computers or even smartphones. However, it can be adapted to run on a cluster without a Graphical User Interface (GUI). Using this method, we have been able to amass evidence through the analysis of 3.9 TBase of RNA-SEQ data across 27 publications documenting the expression of a novel pancRNA overlapping the IPF-associated rs35705950 SNP implicated in MUC5B overexpression, annotated as AC061979.1 by GENCODE. The results were replicated by qRT-PCR in A549 cells and CFBE41o submerged cultures as well as in pHAECs.

5 Author Contributions

- **Neatu R.** collected the data, performed the analysis and wrote the manuscript.
- **Thompson D.J.** validated the pipeline by replicating the results on the SRP082973 dataset and revised the manuscript.
- **Enekwa I.** attempted the generation of rs35705950 hetero- and homozygote A549 derivative cell lines by genome editing and reviewed the manuscript.
- **Schwalbe E.C.** contributed to the final version of the manuscript.
- **Braubach P.** collected and provided donor tissue for isolation of HAECs
- **Fois G.** performed ALI culture of HAEpCs, sample collection and data analysis.
- **Frick M.** designed the air liquid interface experiments, supervised Dr Fois, analysed the data, and reviewed the manuscript.
- **Moschos S.A.** conceived of the presented idea, supervised the project, offered critical feedback and co-authored the manuscript.

6 Conflict of interest statement.

None declared.

References

[1] Sean B Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, 2008.

[2] Y Grace Chen, Ansuman T Satpathy, and Howard Y Chang. Gene regulation in the immune system by long noncoding rnas. *Nature immunology*, 18(9):962–972, 2017.

[3] Alexander F Palazzo and T Ryan Gregory. The case for junk dna. *PLoS Genet*, 10(5):e1004351, 2014.

[4] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57, 2012.

[5] Alexander F Palazzo and Eliza S Lee. Noncoding rna: what is functional and what is junk? *Frontiers in genetics*, 6:2, 2015.

[6] Jana Seiler, Marco Breinig, Maïwen Caudron-Herger, Maria Polycarpou-Schwarz, Michael Boutros, and Sven Diederichs. The lncrna veluct strongly regulates viability of lung cancer cells despite its extremely low abundance. *Nucleic acids research*, 45(9):5458–5469, 2017.

[7] Ping Ji, Sven Diederichs, Wenbing Wang, Sebastian Böing, Ralf Metzger, Paul M Schneider, Nicola Tidow, Burkhard Brandt, Horst Buerger, Etmar Bulk, et al. Malat-1, a novel noncoding rna, and thymosin β 4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, 22(39):8031–8041, 2003.

[8] Tayyeb Bahrami, Mohammad Taheri, Mir Davood Omrani, and Morteza Karimipoor. Associations between genomic variants in lncrna-trpm2-as and lncrna-hnf1a-as1 genes and risk of multiple sclerosis. *Journal of Molecular Neuroscience*, pages 1–6, 2020.

[9] Parameet Kumar, Chaitali Sen, Kathryn Peters, Raymond A Frizzell, and Roopa Biswas. Comparative analyses of long non-coding rna profiles in vivo in cystic fibrosis lung airway and parenchyma tissues. *Respiratory Research*, 20(1):284, 2019.

[10] Anne-Valerie Gendrel and Edith Heard. Fifty years of x-inactivation research. *Development*, 138(23):5049–5055, 2011.

[11] Yi Zhao, Hui Li, Shuangsang Fang, Yue Kang, Wei Wu, Yajing Hao, Ziyang Li, Dechao Bu, Ninghui Sun, Michael Q Zhang, et al. Noncode 2016: an informative and valuable data source of long non-coding rnas. *Nucleic acids research*, 44(D1):D203–D208, 2016.

[12] Shangwei Ning, Jizhou Zhang, Peng Wang, Hui Zhi, Jianjian Wang, Yue Liu, Yue Gao, Maoni Guo, Ming Yue, Lihua Wang, et al. Lnc2cancer: a manually curated database of experimentally supported lncrnas associated with various human cancers. *Nucleic acids research*, 44(D1):D980–D985, 2016.

[13] Geng Chen, Ziyun Wang, Dongqing Wang, Chengxiang Qiu, Mingxi Liu, Xing Chen, Qipeng Zhang, Guiying Yan, and Qinghua Cui. Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic acids research*, 41(D1):D983–D986, 2012.

[14] Xiu Cheng Quek, Daniel W Thomson, Jesper LV Maag, Nenad Bartonicek, Bethany Signal, Michael B Clark, Brian S Gloss, and Marcel E Dinger. lncrnadb v2. 0: expanding the reference database for functional long noncoding rnas. *Nucleic acids research*, 43(D1):D168–D173, 2015.

[15] Lina Ma, Vladimir B Bajic, and Zhang Zhang. On the classification of long non-coding rnas. *RNA biology*, 10(6):924–933, 2013.

[16] Masahiro Uesaka, Kiyokazu Agata, Takao Oishi, Kinichi Nakashima, and Takuya Imamura. Evolutionary acquisition of promoter-associated non-coding rna (pancrna) repertoires diversifies species-dependent gene activation mechanisms in mammals. *BMC genomics*, 18(1):285, 2017.

[17] Linda Minotti, Chiara Agnoletto, Federica Baldassari, Fabio Corrà, and Stefano Volinia. Snps and somatic mutation on long non-coding rna: new frontier in the cancer studies? *High-throughput*, 7(4):34, 2018.

[18] Britney A Helling, Anthony N Gerber, Vineela Kadiyala, Sarah K Sasse, Brent S Pedersen, Lenore Sparks, Yasushi Nakano, Tsukasa Okamoto, Christopher M Evans, Ivana V Yang, et al. Regulation of muc5b expression in idiopathic pulmonary fibrosis. *American journal of respiratory cell and molecular biology*, 57(1):91–99, 2017.

[19] Christopher M Evans, Tasha E Fingerlin, Marvin I Schwarz, David Lynch, Jonathan Kurche, Laura Warg, Ivana V Yang, and David A Schwartz. Idiopathic pulmonary fibrosis: a genetic disease that involves mucociliary dysfunction of the peripheral airways. *Physiological reviews*, 96(4):1567–1591, 2016.

[20] Max A Seibold, Anastasia L Wise, Marcy C Speer, Mark P Steele, Kevin K Brown, James E Loyd, Tasha E Fingerlin, Weiming Zhang, Gunnar Gudmundsson, Steve D Groshong, et al. A common muc5b promoter polymorphism and pulmonary fibrosis. *New England Journal of Medicine*, 364(16):1503–1512, 2011.

[21] Imre Noth, Yingze Zhang, Shwu-Fan Ma, Carlos Flores, Mathew Barber, Yong Huang, Steven M Broderick, Michael S Wade, Pirro Hysi, Joseph Scuirba, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. *The Lancet respiratory medicine*, 1(4):309–317, 2013.

[22] Amy Dressen, Alexander R Abbas, Christopher Cabanski, Janina Reeder, Thirumalai R Ramalingam, Margaret Neighbors, Tushar R Bhangale, Matthew J Brauer, Julie Hunkapiller, Jens Reeder, et al. Analysis of protein-altering variants in telomerase genes and their association with muc5b common variant status in patients with idiopathic pulmonary fibrosis: a candidate gene sequencing study. *The Lancet Respiratory Medicine*, 6(8):603–614, 2018.

[23] Brian D Hobbs, Rachel K Putman, Tetsuro Araki, Mizuki Nishino, Gunnar Gudmundsson, Vilmundur Gudnason, Gudny Eiriksdottir, Nuno Rodrigues Zilhao Nogueira, Josée Dupuis, Hanfei Xu, et al. Overlap of genetic risk between interstitial lung abnormalities and idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine*, 200(11):1402–1413, 2019.

[24] Chunli Wang, Yi Zhuang, Wenwen Guo, Lili Cao, Huan Zhang, Lizhi Xu, Yimei Fan, Deping Zhang, and Yaping Wang. Mucin 5b promoter polymorphism is associated with susceptibility to interstitial lung diseases in chinese males. *PloS one*, 9(8):e104919, 2014.

[25] Gary M Hunninghake, Hiroto Hatabu, Yuka Okajima, Wei Gao, Josée Dupuis, Jeanne C Latourelle, Mizuki Nishino, Tetsuro Araki, Oscar E Zazueta, Sila Kurugol, et al. Muc5b promoter polymorphism and interstitial lung abnormalities. *New England Journal of Medicine*, 368(23):2192–2200, 2013.

[26] Joanne J van der Vis, Reinier Snetselaar, Karin M Kazemier, Liesbeth ten Klooster, Jan C Grutters, and Coline HM van Moorsel. Effect of

m uc5b promoter polymorphism on disease predisposition and survival in idiopathic interstitial pneumonias. *Respirology*, 21(4):712–717, 2016.

[27] Rongrong Wei, Chong Li, Min Zhang, Yava L Jones-Hall, Jamie L Myers, Imre Noth, and Wanqing Liu. Association between muc5b and tert polymorphisms and different interstitial lung disease phenotypes. *Translational Research*, 163(5):494–502, 2014.

[28] Anna L Peljto, Yingze Zhang, Tasha E Fingerlin, Shwu-Fan Ma, Joe GN Garcia, Thomas J Richards, Lori J Silveira, Kathleen O Lindell, Mark P Steele, James E Loyd, et al. Association between the muc5b promoter polymorphism and survival in patients with idiopathic pulmonary fibrosis. *Jama*, 309(21):2232–2239, 2013.

[29] Anna L Peljto, Moises Selman, Dong Soon Kim, Elissa Murphy, Laura Tucker, Annie Pardo, Jung Su Lee, Wonjun Ji, Marvin I Schwarz, Ivana V Yang, et al. The muc5b promoter polymorphism is associated with idiopathic pulmonary fibrosis in a mexican cohort but is rare among asian ancestries. *Chest*, 147(2):460–464, 2015.

[30] Carmel J Stock, Hiroe Sato, Carmen Fonseca, Winston AS Banya, Philip L Molyneaux, Huzaifa Adamali, Anne-Marie Russell, Christopher P Denton, David J Abraham, David M Hansell, et al. Mucin 5b promoter polymorphism is associated with idiopathic pulmonary fibrosis but not with development of lung fibrosis in systemic sclerosis or sarcoidosis. *Thorax*, 68(5):436–441, 2013.

[31] Raphael Borie, Bruno Crestani, Philippe Dieude, Hilario Nunes, Yannick Allanore, Caroline Kannengiesser, Paolo Airo, Marco Matucci-Cerinic, Benoit Wallaert, Dominique Israel-Biet, et al. The muc5b variant is associated with idiopathic pulmonary fibrosis but not with systemic sclerosis interstitial lung disease in the european caucasian population. *PloS one*, 8(8):e70621, 2013.

[32] Amit Kishore, Veronika Žižková, Lenka Kocourková, Jana Petrkova, Evangelos Bouros, Hilario Nunes, Vladimíra Loštáková, Joachim Müller-Quernheim, Gernot Zissel, Vitezslav Kolek, et al. Association study for 26 candidate loci in idiopathic pulmonary fibrosis patients from four european populations. *Frontiers in immunology*, 7:274, 2016.

[33] Qing-Qing Zhu, Xin-Lin Zhang, Si-Min Zhang, Shao-Wen Tang, Hai-Yan Min, Long Yi, Biao Xu, and Yong Song. Association between the muc5b promoter polymorphism rs35705950 and idiopathic pulmonary fibrosis: a meta-analysis and trial sequential analysis in caucasian and asian populations. *Medicine*, 94(43), 2015.

[34] Yasushi Horimasu, Shinichiro Ohshimo, Francesco Bonella, Sonosuke Tanaka, Nobuhisa Ishikawa, Noboru Hattori, Nobuoki Kohno, Josune Guzman, and Ulrich Costabel. Muc 5 b promoter polymorphism in j apanese patients with idiopathic pulmonary fibrosis. *Respirology*, 20(3):439–444, 2015.

[35] Yanhan Deng, Zongzhe Li, Juan Liu, Zheng Wang, Yanyan Cao, Yong Mou, Bohua Fu, Biwen Mo, Jianghong Wei, Zhenshun Cheng, et al. Targeted resequencing reveals genetic risks in patients with sporadic idiopathic pulmonary fibrosis. *Human Mutation*, 39(9):1238–1245, 2018.

[36] Susan K Mathai, Stephen Humphries, Jonathan A Kropski, Timothy S Blackwell, Julia Powers, Avram D Walts, Cheryl Markin, Julia Woodward, Jonathan H Chung, Kevin K Brown, et al. Muc5b variant is associated with visually and quantitatively detected preclinical pulmonary fibrosis. *Thorax*, 74(12):1131–1139, 2019.

[37] Jose M Lorenzo-Salazar, Shwu-Fan Ma, Jonathan Jou, Pei-Chi Hou, Beatriz Guillen-Guio, Richard J Allen, R Gisli Jenkins, Louise V Wain, Justin M Oldham, Imre Noth, et al. Novel idiopathic pulmonary fibrosis susceptibility variants revealed by deep sequencing. *ERJ open research*, 5(2), 2019.

[38] Camille Moore, Rachel Z Blumhagen, Ivana V Yang, Avram Walts, Julie Powers, Tarik Walker, Makenna Bishop, Pamela Russell, Brian Vestal, Jonathan Cardwell, et al. Resequencing study confirms that host defense and cell senescence gene variants contribute to the risk of idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine*, 200(2):199–208, 2019.

[39] Haiming Jiang, Yejia Hu, Li Shang, Yuzhu Li, Lihua Yang, and Yuguo Chen. Association between muc5b polymorphism and susceptibility and severity of idiopathic pulmonary fibrosis. *International journal of clinical and experimental pathology*, 8(11):14953, 2015.

[40] Carmel J Stock, Caterina Conti, Ángeles Montero-Fernandez, Gaetano Caramori, Philip L Molyneaux, Peter M George, Maria Kokosi, Vaslis Kouranos, Toby M Maher, Felix Chua, et al. Interaction between the promoter muc5b polymorphism and mucin expression: is there a difference according to ild subtype? *Thorax*, 75(10):901–903, 2020.

[41] Yasushi Nakano, Ivana V Yang, Avram D Walts, Alan M Watson, Britney A Helling, Ashley A Fletcher, Abigail R Lara, Marvin I Schwarz, Christopher M Evans, and David A Schwartz. Muc5b promoter variant rs35705950 affects muc5b expression in the distal airways in idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine*, 193(4):464–466, 2016.

[42] Gang Chen, Carla MP Ribeiro, Ling Sun, Kenichi Okuda, Takafumi Kato, Rodney C Gilmore, Mary B Martino, Hong Dang, Aiman Abzhanova, Jennifer M Lin, et al. Xbp1s regulates muc5b in a promoter variant–dependent pathway in idiopathic pulmonary fibrosis airway epithelia. *American journal of respiratory and critical care medicine*, 200(2):220–234, 2019.

[43] Sebastiano Di Bella, Alessandro La Ferlita, Giovanni Carapezza, Salvatore Alaimo, Antonella Isacchi, Alfredo Ferro, Alfredo Pulvirenti, and Roberta Bosotti. A benchmarking of pipelines for detecting ncrnas from rna-seq data. *Briefings in Bioinformatics*, 2019.

[44] Kazunori Akiyama, Antxon Alberdi, Walter Alef, Keiichi Asada, Rebecca Azulay, Anne-Kathrin Baczko, David Ball, Mislav Baloković, John Barrett, Dan Bintley, et al. First m87 event horizon telescope results. iii. data processing and calibration. *The Astrophysical Journal Letters*, 875(1):L3, 2019.

[45] Daehwan Kim, Joseph M Paggi, Chanhee Park, Christopher Bennett, and Steven L Salzberg. Graph-based genome alignment and genotyping with hisat2 and hisat-genotype. *Nature biotechnology*, 37(8):907–915, 2019.

[46] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[47] Kelly A Frazer, Lior Pachter, Alexander Poliakov, Edward M Rubin, and Inna Dubchak. Vista: computational tools for comparative genomics. *Nucleic acids research*, 32(suppl_2):W273–W279, 2004.

[48] Guang Li and Peter WH Holland. The origin and evolution of argfx homeobox loci in mammalian radiation. *BMC evolutionary biology*, 10(1):1–10, 2010.

[49] Kenneth J Livak and Thomas D Schmittgen. Analysis of relative gene expression data using real-time quantitative pcr and the 2- $\delta\delta$ct method. *methods*, 25(4):402–408, 2001.

[50] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. Gencode reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019.

[51] Haijun Zhang, Jing Yang, Matthew S Walters, Michelle R Staudt, Yael Strulovici-Barel, Jacqueline Salit, Jason G Mezey, Philip L Leopold, and Ronald G Crystal. Mandatory role of hmga1 in human airway epithelial normal differentiation and post-injury regeneration. *Oncotarget*, 9(18):14324, 2018.

[52] Neil R Hackett, Renat Shaykhiev, Matthew S Walters, Rui Wang, Rachel K Zwick, Barbara Ferris, Bradley Witover, Jacqueline Salit, and Ronald G Crystal. The human airway epithelial basal cell transcriptome. *PloS one*, 6(5), 2011.

[53] Luca Richeldi, Harold R Collard, and Mark G Jones. Idiopathic pulmonary fibrosis. *The Lancet*, 389(10082):1941–1952, 2017.

[54] Wei Jiang, Yuting Liu, Rui Liu, Kun Zhang, and Yi Zhang. The lncrna deanr1 facilitates human endoderm differentiation by activating foxa2 expression. *Cell reports*, 11(1):137–148, 2015.

[55] Daniel T Swarr, Michael Herriges, Shanru Li, Mike Morley, Sharlene Fernandes, Anusha Sridharan, Su Zhou, Benjamin A Garcia, Kathleen Stewart, and Edward E Morrisey. The long noncoding rna falcor regulates foxa2 expression to maintain lung epithelial homeostasis and promote regeneration. *Genes & development*, 33(11-12):656–668, 2019.

[56] Ha X Dang, Nicole M White, Emily B Rozycki, Brooke M Felsheim, Mark A Watson, Ramaswamy Govindan, Jingqin Luo, and

Christopher A Maher. Long non-coding rna lcal62/linc00261 is associated with lung adenocarcinoma prognosis. *Heliyon*, 6(3):e03521, 2020.

[57] Woosuk Choi, Shawn Choe, and Gee W Lau. Inactivation of foxa2 by respiratory bacterial pathogens and dysregulation of pulmonary mucus homeostasis. *Frontiers in immunology*, 11:515, 2020.

[58] Qinghua Zhang, Yan Wang, Danhua Qu, Jinyan Yu, and Junling Yang. The possible pathogenesis of idiopathic pulmonary fibrosis considering muc5b. *BioMed research international*, 2019, 2019.

[59] Manish Kumar and Sterghios Moschos. Oligonucleotide therapies for the lung: ready to return to the clinic? *Molecular Therapy*, 25(12):2604–2606, November 2017.

[60] Phillip Grote and Bernhard G Herrmann. The long non-coding rna fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis. *RNA biology*, 10(10):1579–1585, 2013.

[61] Michael B Feldman, Michael Wood, Allen Lapey, and Hongmei Mou. Smad signaling restricts mucous cell differentiation in human airway epithelium. *American journal of respiratory cell and molecular biology*, 61(3):322–331, 2019.

[62] Pau Martin-Malpartida, Marta Batet, Zuzanna Kaczmarska, Regina Freier, Tiago Gomes, Eric Aragón, Yilong Zou, Qiong Wang, Qiaoran Xi, Lidia Ruiz, et al. Structural basis for genome wide recognition of 5-bp gc motifs by smad transcription factors. *Nature communications*, 8(1):1–15, 2017.

[63] Satoshi Sakai, Tatsuya Ohhata, Kyoko Kitagawa, Chiharu Uchida, Takuya Aoshima, Hiroyuki Niida, Tetsuro Suzuki, Yasumichi Inoue, Keiji Miyazawa, and Masatoshi Kitagawa. Long noncoding rna elit-1 acts as a smad3 cofactor to facilitate tgfβ/smad signaling and promote epithelial–mesenchymal transition. *Cancer research*, 79(11):2821–2838, 2019.

[64] Sarah C Pyfrom, Hong Luo, and Jacqueline E Payton. Plaidoh: a novel method for functional prediction of long non-coding rnas identifies cancer-specific lncrna activities. *BMC genomics*, 20(1):137, 2019.

[65] Yun Xiao, Jingpu Zhang, and Lei Deng. Prediction of lncrna-protein interactions using hetesim scores based on heterogeneous networks. *Scientific reports*, 7(1):1–12, 2017.

[66] Qi Zhao, Yue Zhang, Huan Hu, Guofei Ren, Wen Zhang, and Hongsheng Liu. Irwnrlpi: integrating random walk and neighborhood regularized logistic matrix factorization for lncrna-protein interaction prediction. *Frontiers in genetics*, 9:239, 2018.

[67] Lihong Peng, Fuxing Liu, Jialiang Yang, Xiaojun Liu, Yajie Meng, Xiaojun Deng, Cheng Peng, Geng Tian, and Liqian Zhou. Probing lncrna–protein interactions: Data repositories, models, and algorithms. *Frontiers in Genetics*, 10, 2019.