3 Results – Genetic Variation in Human *IFITM3*

3.1 Introduction

Although *IFITM* genes were first identified in 1991, it is only since 2009 that the ability of restriction factors IFITM1, 2, and 3 to prevent replication of a broad range of enveloped viruses, including influenza viruses, has been convincingly established[121]. Single nucleotide polymorphisms (SNPs) in important antiviral genes such as *TRIM5α* and *RIG-I* have been used to assess an individual's risk of severe autoimmune or infectious disease[237,238]. However, no studies thus far have investigated the variation of SNPs present in the *IFITM* genes or if any of the SNPs are associated with the patient's response to an infectious disease.

The 2009 H1N1 influenza pandemic provided a unique opportunity to study whether or not SNPs in *IFITM3* are associated with a severe response to IAV, as a large number of patients were hospitalised. Moreover, because this was an exposure to a new IAV, no or little adaptive immunity was present in infected people. This chapter aims to explore the host genetics of hospitalised patients and compare them to ethnically-matched background cohorts.

The aims and objectives of this chapter are as follows:

i. Examine the *IFITM3* locus in the Ensembl database for evidence of SNPs and alternative transcripts
ii. Establish if any SNPs are associated with susceptibility to influenza infection
iii. Investigate the mechanism of action of any SNPs associated with severe influenza
iv. Investigate whether or not alternative splicing of *IFITM3* occurs *in vitro*

3.2 Analysis of Human IFITM3 Transcripts in the Ensembl Database

The coding structure and polymorphisms within human *IFITM3* were initially determined by reference to publically available data. Three protein-encoding transcripts for *IFITM3* were predicted in the Ensembl database (Figure 21); one encodes the full-length wildtype protein (IFITM3_001), which consists of two exons separated by an intron (chromosome 11: 321,050-319,669). The second encodes an N-terminally truncated protein (IFITM3_002) that initiates the open reading frame from the second methionine of exon one (chromosome 11: 321,340-319,773), and the third transcript (IFITM3_004) encodes the same sequence as IFITM3_002, but the 5' UTR is mapped to more than 6 kb further upstream of the gene body (chromosome 11: 327,537-319,773).

By searching dbSNP and 1000 Genomes datasets, 28 exonic SNPs were identified within *IFITM3*, which are summarised in Table 6. Of these 28, 14 were synonymous, 12 were non-synonymous, one resulted in an amino acid deletion, and one resulted in a premature stop codon. Therefore IFITM3 has the potential to vary in both primary sequence and produce alternative transcripts.

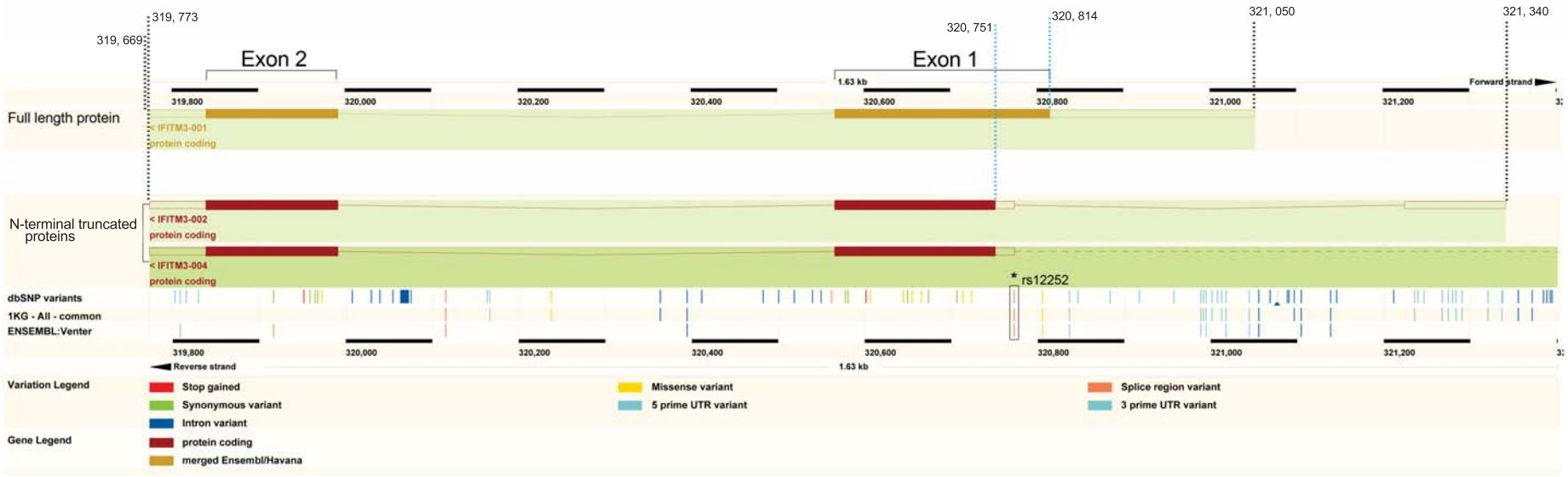

Figure 21: Single nucleotide polymorphisms present in the human *IFITM3* locus

Three protein-coding transcripts are shown in the Ensembl browser on the reverse strand of chromosome 11 of the human genome build version 74.37; one transcript for full-length IFITM3 (gold) and two transcripts encoding N-terminally truncated proteins (red). The boundaries of the 3' and 5' UTRs are highlighted with a black dotted line; note that the 5' region of IFITM3_004 is not shown as it is 6 kb upstream of the locus shown (Chr11:327537). The start of coding regions are highlighted with a blue dotted line. dbSNP and 1000 Genomes data show one stop gain-of-function SNP (red), 12 potential non-synonymous SNPs (yellow) and one splice site SNP (rs12252 highlighted with an ✶) in the translated *IFITM3* sequence (see variation legend).

Table 6: Single nucleotide polymorphisms present in human *IFITM3* gene

aa* Number	DNA Position	SNP	Base Change	aa Change	Major Allele	Minor Allele
2	320808	rs56169757	T/C	N/N	T	C
3	320805	rs1136853	C/A	H/Q	C	A
4	320803	sm42696	C/T	T/I	C	T
9-10	320786-3	rs56398316	TCT/-	FS/S	TCT	ΔTCT
14	320772	rs12252	T/C	S/S	T	C
17	320763	rs56323507	C/T	P/P	C	T
20	320754	rs56020216	T/C	Y/Y	T	C
27	320733	rs55888283	C/G	H/Q	C	G
28	320730	rs142924318	G/A	E/E	G	A
31	320723	rs56227617	G/A	V/M	G	A
34	320713	rs56188107	C/G	A/G	C	G
42	320689	rs55900504	C/T	T/M	C	T
52	320658	rs11553884	C/A	T/T	C	A
55	320649	rs11553885	C/T	P/P	C	T
56	320647	rs55794999	A/G	D/G	A	G
57	320645	rs11553883	C/G	H/D	C	G
58	320640	rs72636984	C/G	V/V	C	G
69	320609	rs12778	A/G	N/D	A	G
70	320606	cosm42691	C/A	P/T	C	A
79	320577	rs55965761	C/G	A/A	C	G
92	319964	rs11539511	C/T	D/D	C	T
95	319957	rs61744108	G/A	G/R	G	A
97	319951	rs113745243	C/T	Q/STOP	C	T
108	319916	rs1060603	C/T	I/I	C	T
113	319903	rs1060675	C/T	L/L	C	T
120	319882	rs1137969	C/T	L/L	C	T
126	319862	rs11539509	G/C	V/V	G	C
129	319855	cosm42690	A/C	F/V	A	C

*aa; amino acid

Adapted from Everitt *et al.* (2012)[3]

3.3 Developing a Robust PCR to Amplify Human IFITM3

To identify polymorphisms in IFITM3, a robust and specific PCR was developed. The PCR was tested on DNA from a lymphoblastoid cell line (LCL), using primers originally designed by Seo *et al.* (2010)[239] (IFITM3_F2 and IFITM3_R2 [Table 3]). However, first attempts using this method resulted in amplification of many non-specific bands (data not shown). The specific band containing exon 1 and 2 and the intron (1.7 kb) was extracted from the gel and purified before sequencing, but this resulted in a very low yield and poor sequencing results.

To resolve this problem, a hemi-nested PCR was developed (Figure 22). One of the original primers, IFITM3_R2, was used in a first round of amplification along with a newly designed forward primer, SES003_F. 2 µl of this reaction was used as template for a second round of PCR using the original forward primer (IFITM3_F2) and short_IFITM3_R2. The second IFITM3_F2 primer anneals to the target DNA just inside of SES003_F (Figure 22), thus reducing non-specific amplification. Since only one product was amplified the PCR reaction was purified directly on a PCR purification column, increasing the final yield. This method was then used to amplify IFITM3 from DNA extracted from patients in an influenza pandemic cohort.

ATAACAATAAAAGGCCTCAGAGGGGAAGGGAATGAGGCAGGAAATTAAATAAAATTAAAATTAAAAAG
AAAGAGAAATAGGTTTTCCTGTATCAGGCTGACTCGTCCCGGAGGCAGCAGCAGACACAGCTGAGACC
CAGGAAAAGTCCTGATAATATTATCTAATGTGCTCTGAGACTCTCCCAGCACTCCCTTAACACAGGGA
GAAGAAAAAACAAATTTTCCTTTGTTTTTGGAATGAGTTTATAGATTCCTGTTCTCTGTAACTAGTGA
CTTCAAGTATTCTGTTTTATCTAAGAAGTACAGTGAAGGTCATGAGACGCCTGAGCAGGCCTGAACGC
.......CGTGTCAGCCAGGATGGTCTCGATCTCCTGACCTCATGATCCTCCCACCTCAGCCTCCCAA
AGTGCTGGGATTACAGGCGTGAGCCGCGGCGCCCGGCAGAGGTGAGGGCTTTGGGGGAACGGTTGTGG
GGCCTGGAGTGTGGAGGCGTCAGCGCAGGCCTGGCAGGAGCCCTGAACCGGGACAGTGGGGTCCTGCA
GCTGCTGGCCTGGGGTGTGGAGACCCCCAACACAGGGGAAGTCTCCAGGACCCCACACCACTAACAAG
ATGAGCCTTGTGCTCCCTTGGGCTCTAGAGAGGAAGCCCCTCTTAGCCCTCAGCCCCTCTTTCCTCCC
TCTCCTAAAGTAATTTGATCCTCAGGAATTTGTTCCGCCCTCATCTGGCCCCGGCCAAATCCCGATTT
GACAAATGCCAGGAAAAGGAAACTGTTGAGAAACCGAAACTACTGGGGAAAGGGAGGGCTCACTGAGA
ACCATCCCAGTAACCCGACCGCCGCTGGTCTTCGCTGGACACCATGAATCACACTGTCCAAACCTTCT
TCTCTCCTGTCAACAGTGGCCAGCCCCCCAACTATGAGATGCTCAAGGAGGAGCACGAGGTGGCTGTG
CTGGGGGCGCCCCACAACCCTGCTCCCCCGACGTCCACCGTGATCCACATCCGCAGCGAGACCTCCGT
GCCCGACCATGTCGTCTGGTCCCTGTTCAACACCCTCTTCATGAACCCCTGCTGCCTGGGCTTCATAG
CATTCGCCTACTCCGTGAAG*GTGCGTATGGCCCCAGGGAATGCTCAGAGGGTGCCGCTGAGCCTGGAG*
CTCCACCTGCCCACATGCTGCCTGGGGTGGGGACTTGTGTGTCCCTGTGACTGTGAGTTTGTGTGCAC
CTCTGTCCCGTGTGTGCCCACGTCAGTGGCTTTGTCTGTGTGATCTGTGTGTGTGTGTGGCTTGGGGA
ATCTGCCCAGTGCAGGTTTAGGAGGAGGCTCCAGGAGGCTGGCTGGCTGGCTCAGAGTCTGTCCCCGG
CTATCCACTAGCCCAGAGCAGTTCTCCCTATAGCCCAGTAAGAAATTACACCTTCACCTTCCAGACTG
GCACCCAGGCTCTCCCAGAAAGTGAGAAGGGAACTCACAGGTGACTTCACCCCATGGTGGGGAGAACA
GCCTGTGCTGAGGTCAAGGCAGAAGGAGGATGAGCCCCGAGGCTCCTGGAGAGTCTGAGCCCGGGTGA
GGAAGGGGAGGAGGTGGTCCCTGATCTCAGGGCGGGGAGAGCCAATGAGGAGACGGAGCCATAGCACG
*CGGCTCTCAGCTGGGGGATCCTGGTCCCCTCACCATCTCCTCTCCCCCAG*TCTAGGGACAGGAAGATG
GTTGGCGACGTGACCGGGGCCCAGGCCTATGCCTCCACCGCCAAGTGCCTGAACATCTGGGCCCTGAT
TCTGGGCATCCTCATGACCATTCTGCTCATCGTCATCCCAGTGCTGATCTTCCAGGCCTATGGATAGA
TCAGGAGGCATCACTGAGGCCAGGAGCTCTGCCCATGACCTGTATCCCACGTACTCCAACTTCCATTC
CTCGCCCTGCCCCCGGAGCCGAGTCCTGTATCAGCCCTTTATCCTCACACGCTTTTCTACAATGGCAT
TCAATAAAGTGCACGTGTTTCTGGTGCTGCTGCGACTTCACCTGGGGAGGGGTCTGGCTGAGGGTTCG
GAGCGTGGTTCTGAGACTGAGCAGGTTGGTCAGCCCCTGCACTGCCCCTTCCGGCCTCTGTGCATCTC
TTGGGGACCGGGCAAGTGCTCAGGCCTTCTGGTTTCGGGCCTCCTGCCGTGAGCAGCAGCTGGATCCA

NNN = SES003_F; NNN = IFITM3_F2; NNN = IFITM3_R2; NNN = Short_IFITM3_R2; NNN = Start and stop codons; NNN = Internal start codon; *NNN* = Intron; … = concatenated sequence (1295 bp); NNN = promoter elements; NNN = ISRE

Figure 22: The primer binding sites for amplification of human *IFITM3* by hemi-nested PCR

Primer SES003_F binds 404 bp upstream of the start site and primer IFITM3_R2 binds 328 bp downstream of the stop codon. These primers amplify a 1.78 kb fragment, which is used as the template for the subsequent PCR using primers IFITM3_F2 and short_IFITM3_R2.

3.4 Sequencing Human IFITM3 from Clinical Samples

The Mechanisms of Severe Acute Influenza Consortium (MOSIAC) study recruited a single cohort of 250 individuals hospitalised between November 2009 and February 2011 with severe acute respiratory infection (SARI), during the second and third waves of the influenza pandemic in the UK. In addition the genetics of influenza susceptibility in Scotland (GenISIS) consortium recruited SARI patients in Scotland during the pandemic. Both these cohorts provided a unique opportunity to study how influenza causes illness and how patient management can be improved.

These consortiums collected the DNA and meta-data (sex, age, weight, pre-existing medical conditions) of individuals who required admission to hospital as a result of pandemic H1N1/09 or seasonal influenza virus infection in 2009–2010. From these collections, patients with significant co-morbidities and those non-Caucasians (n=31) were excluded, leaving 60 Caucasian SARI patients for this study.

The *IFITM3* gene from these patients was amplified from DNA extracted from the peripheral blood by hemi-nested PCR. Of these, 53 samples produced single bands with enough material to sequence (Figure 23A). *IFITM3* was distinguished from *IFITM2* by the presence of a double phenylalanine at amino acid position 8 and 9 (Figure 23B), and at least 5-fold coverage of the SNP was required for accurate genotyping. 45 patients (84.9 %) carried majority alleles for all 28 known SNPs in the coding sequence of the gene, but the remaining eight possessed known variants (Figure 24). Of these, four were heterozygous (CT) at rs12252 and three were homozygous for the ancestral C allele. Three of the four heterozygotes were also heterozygous at rs1136853 (C to T change). One further patient had an alternative allele for rs56227617 (G to C), however this did not encode the described valine to a methionine change, but an alternative change to leucine, as surrounding bases were also mutated.

Analysis of the prevalence of the minority A allele at rs1136853 in the SARI patients showed that it did not differ significantly from the Hardy-Weinberg equilibrium (Table 7) and also that the proportion of heterozygotes in the study group (5.66 %) did not differ significantly from the proportion of heterozygotes in the control European population (4.75 %).

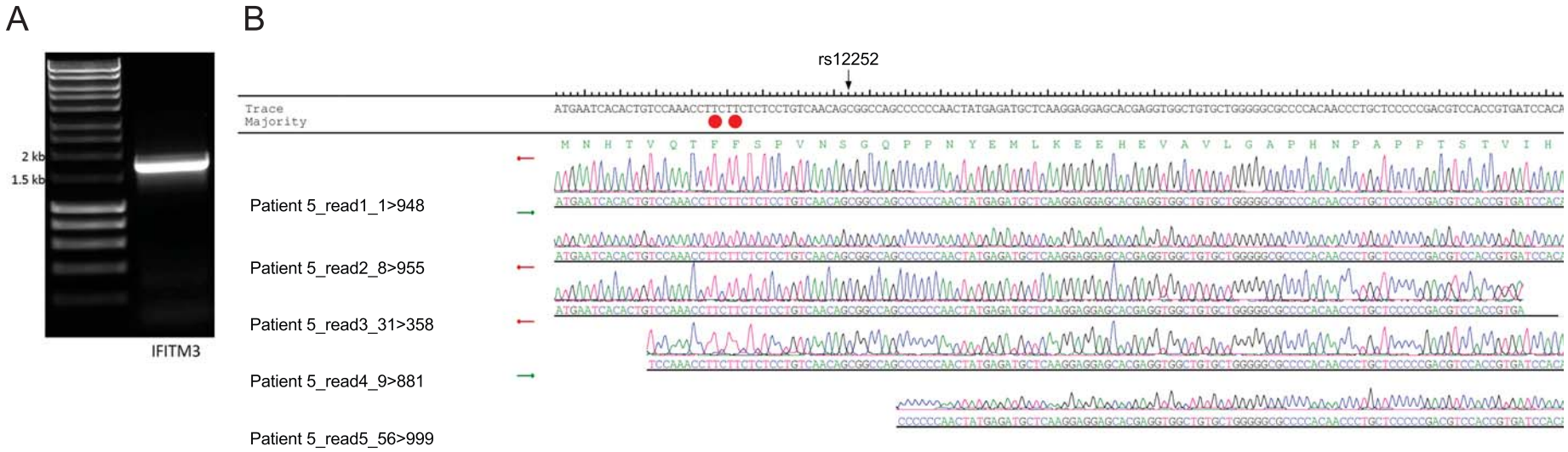

Figure 23: Capillary sequencing of human *IFITM3*

The full *IFITM3* gene (exon 1, exon 2, and the intron) from SARI patients was amplified by hemi-nested PCR to produce a 1.7 kb amplicon (A). DNA amplicons were sequenced by capillary sequencing to a depth of 4-5-fold coverage (B). The red circles highlight the consecutive phenylalanines that differentiate human *IFITM3* from *IFITM2*, and the black arrow denotes SNP rs12252 in this sample.

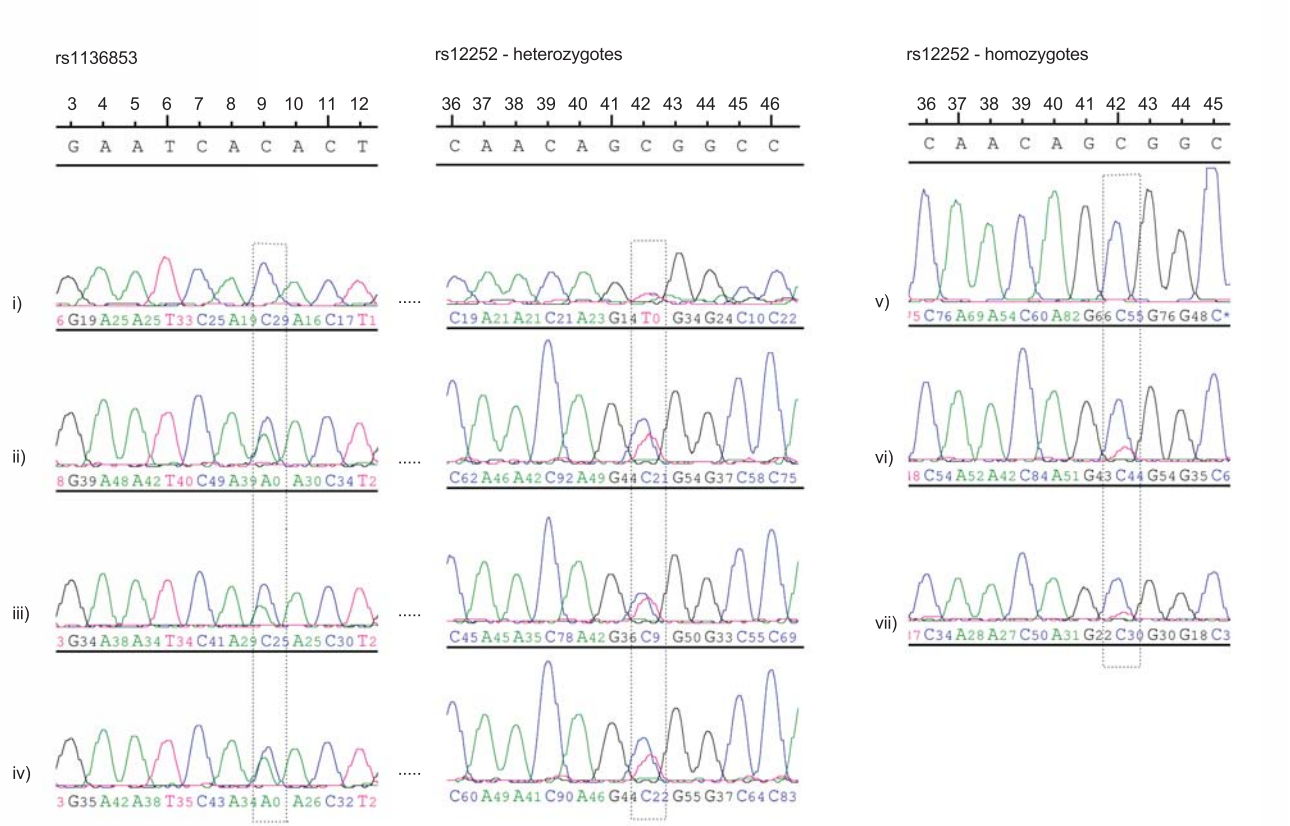

Figure 24: Allele frequencies for rs12252 and rs1136853 in human *IFITM3*

Seven patients possessed at least one known SNP in *IFITM3*; one was heterozygous for only rs12252 (i), three were heterozygous for both rs12252 and rs1136853 (ii – iv), and a further three were homozygous for the C allele at rs12252 (v-vii). Heterozygotes were called by low Phred scores and small peaks, compared to the surrounding bases. Numbers in black represent the nucleotide number, where the 'A' of the ATG start codon is 1.

Table 7: The allele frequency distribution for SNP rs1136853 in different populations

Population	Allele Frequency		Genotype Numbers			Total Samples	Proportion of AA	p-value[1]
	A	C	AA	AC	CC			
AFR[2]	0.067	0.933	2	29	215	246	0.81 %	0.293
ASN[2]	0	1	0	0	286	286	0 %	1
EUR[2]	0.024	0.976	0	18	361	379	0 %	1
A/H1N1/09 or influenza B[3]	**0.057**	**0.943**	**0**	**3**	**50**	**53**	**0 %**	**1**

[1]Probability that the observed genotype frequencies deviate from Hardy-Weinberg Equilibrium (Fisher's Exact test)

[2]Allele and genotype frequencies from 1000 Genomes sequence data (AFR, African ancestry [YRI, ASW, LWK]; ASN, Chinese and Japanese ancestry [CHB, JPT, CHS]; EUR, European ancestry [CEU, FIN, GBR, IBI, TSI]).

[3]Allele and genotype frequencies determined in this study

YRI (Yoruba in Ibadan, Nigera), ASW (Americans of African Ancestry in south west USA), LWK (Luhya in Webuye, Kenya) CHB (Han Chinese in Bejing, China), JPT (Japanese in Tokyo, Japan), CHS (Southern Han Chinese), CEU (Utah Residents with Northern and Western European ancestry), FIN (Finnish in Finland), GBR (British in England and Scotland), IBI (Iberian population in Spain), TSI (Toscani in Italia).

Analysis of SNP rs12252 in the HapMap dataset showed that the frequency of the C allele varies significantly between different ethnic populations (Figure 25). The C allele for this SNP is very rare in European populations (0.034), but more common in African and Asian populations (0.242 and 0.491, respectively) (Table 8). However, through directed sequencing, the C allele frequency in this study of hospitalised Caucasians was calculated to be 0.094, three times higher than in the ethnically-matched group derived from the 1000 Genomes project (Table 8 and Figure 25). This difference is even more distinct when comparing the proportion of CC homozygote individuals in this study (5.66 %) to the ethnically matched population (0 %). The genotype frequencies in this study also deviate from the Hardy-Weinberg equilibrium (unlike the control background population), suggesting an enrichment of the C allele. The frequency was also compared to a larger population (n=8892) of Caucasians from the Netherlands. The allele frequencies for rs12252 were imputed in this dataset against the June 2011 release of 1000 Genotypes phased haplotypes, and the frequency of the C allele was found to be 0.026. Therefore SNP rs12252 was over-represented in cases compared to Caucasian control groups.

Interestingly, the background population of Asian controls deviates significantly from Hardy-Weinberg equilibrium (p=0.00005, Table 8). Deviation in a control sample can be the result of poor sampling, however the excess of the ancestral C allele could mean that the locus is under selection in this population, *i.e.* that the allele has an unknown beneficial role that is being selected for[240].

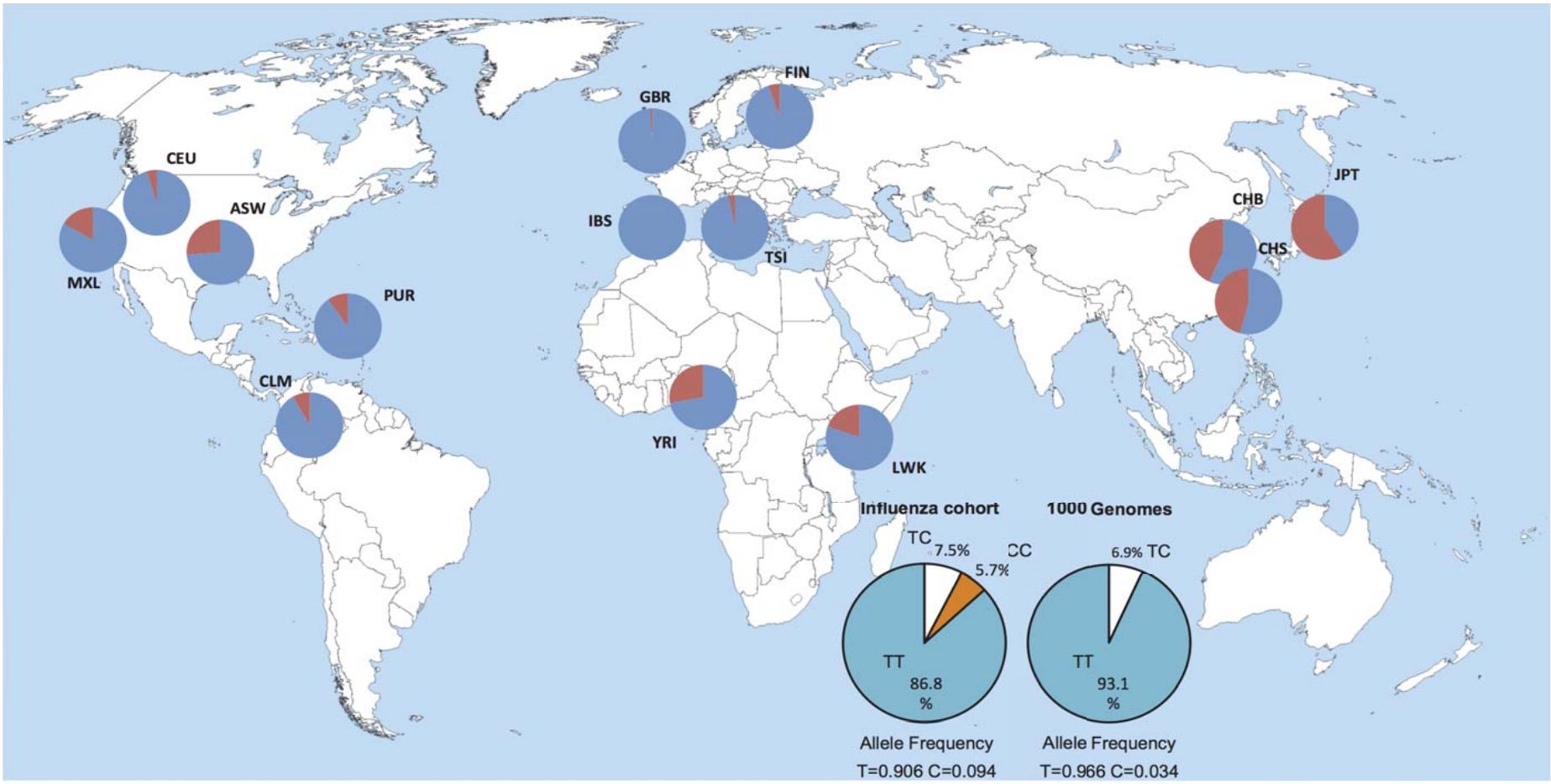

Figure 25: Global variation in the frequency of alleles at rs12252

The small pie charts show the frequency of the T (blue) and C (red) alleles for rs12252 in each population in the 1000 Genomes database. The ancestral C allele is much rarer in central Europe than people of Asian descent. The two larger pie charts show the genotype frequencies for rs12252 (TT, turquoise; CC, orange; TC, white) in the severe influenza cohort differed significantly from European (CEU,FIN,GBR,IBI,TSI) matched controls in the 1000 Genomes dataset.

Table 8: The allele frequency distribution for SNP rs12252 in different populations

Population	Allele Frequency		Genotype Numbers			Total Samples	Proportion of CC	p-value[1]
	C	T	CC	CT	TT			
AFR[2]	0.242	0.758	15	89	142	246	6.10 %	0.742
ASN[2]	0.491	0.509	86	109	91	286	30.07 %	0.00005
EUR[2]	0.034	0.966	0	26	353	379	0 %	1
1000 Genomes 06/11 (Netherlands)[3]	0.026	0.974	-	-	-	8892	-	-
A/H1N1/09 or influenza B[4]	**0.094**	**0.906**	**3**	**4**	**46**	**53**	**5.66 %**	**0.003**

[1]Probability that the observed genotype frequencies deviate from Hardy-Weinberg Equilibrium (Fisher's Exact test)

[2]Allele and genotype frequencies from 1000 Genomes sequence data (AFR, African ancestry [YRI, ASW, LWK]; ASN, Chinese and Japanese ancestry [CHB, JPT, CHS]; EUR, European ancestry [CEU, FIN, GBR, IBI, TSI]).

[3]Allele frequencies imputed against June 2011 release of 1000 Genomes phased haplotypes

[4]Allele and genotype frequencies determined in this study

3.5 The Functional Impact of rs1136853 and rs12252 on IFITM3 Expression

The minor A allele at SNP rs1136853 encodes a histidine to glutamine substitution at position 3 (H3Q_IFITM3). Although patients in this study were all heterozygous, this amino acid substitution was tested by John *et al. in vitro*. A549 cells were stably transduced to over-express IFITM3 or H3Q_IFITM3 and infected with influenza A (A/WSN/1933), but no difference in the percentage of infected cells was observed[5].

The functional consequences of SNP rs12252 was investigated to attempt to explain the apparent increase of the minority allele in the group of individuals hospitalised with influenza. Automatic *in silico* annotation of synonymous SNP rs12252 suggested that it is located next to the splice acceptor sequence in exon 1, which if functional could result in splicing of an alternative transcript of *IFITM3* (Figure 26).

Aside from the splice donor and acceptor sites for removal of the *IFITM3* intron, two additional splice donor sequences exist at position Chr11:321224 and Chr11:327251 (Figure 26). Used in combination with the splice acceptor adjacent to rs12252 (Chr11:320773), this would give rise to the predicted IFITM3_002 and IFITM3_004 transcripts (Figure 21). Therefore, SNP rs12252 could be associated with splicing and expression of the *IFITM3* splice variants IFITM3_002 or IFITM3_004, which are predicted to encode an IFITM3 protein lacking the first N-terminal 21 amino acids (Figure 26).

The strength of canonical splice sites depend on the base in the +1 position relative to the splice acceptor[241], with relative strength of splicing being T < C < A < G. SNP rs12252 is located at the +1 position of the putative splice acceptor. Therefore, the minority C allele may result in an increase in the proportion of spliced transcript IFITM3_002 or IFITM3_004. We hypothesise that having two C alleles at rs12252 increases the likelihood of splicing of these alternative transcripts, and shifts the balance in protein production from full-length protein, to a truncated and potentially reduced-function protein.

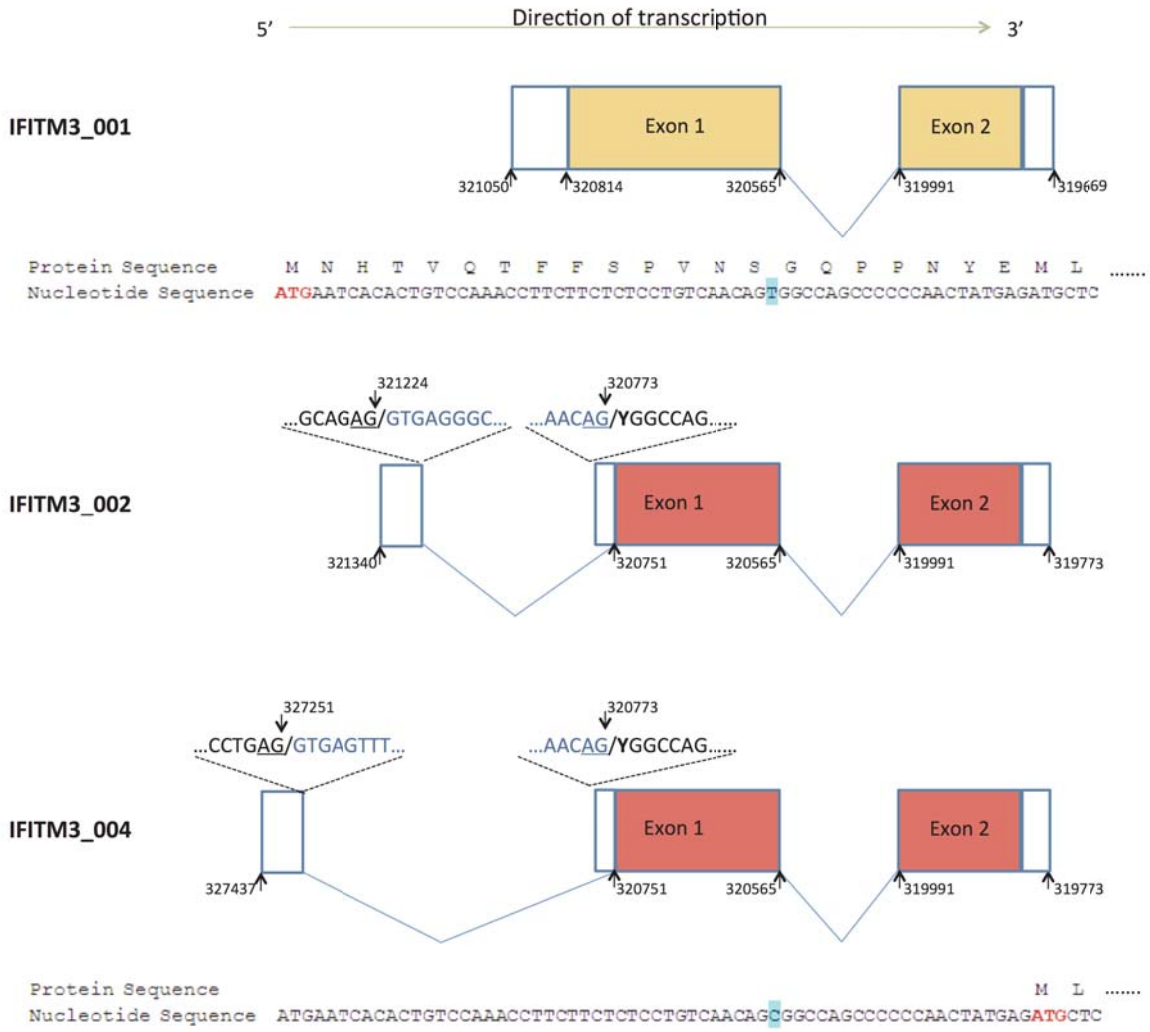

Figure 26: Alternative splicing of exon 1 of *IFITM3*

Full-length protein (IFITM3_001) is translated from an mRNA transcript consisting of two exons with a processed intron. Translation is initiated from the methionine at position Chr11:320814. A canonical splice site acceptor is present in exon 1, position Chr11:320773 (AG). Two alternative 5' UTRs are predicted for transcripts IFITM3_002 and IFITM3_004, which would use this splice acceptor and a donor sequence at position Chr11:321224 or Chr11:327251. Initiation of translation for these transcripts would start at the next available methionine (position Chr11:320751), encoding a truncated protein without the first 21 amino-acids. The strength of this splice site is theoretically dependent on the first base 3' to the splice acceptor sequence (Y, rs12252). Intronic nucleotides are in blue text and exonic nucleotides are in black text. Coloured boxes indicate coding regions. Splice acceptor and donor sites are underlined. The rs12252 allele is highlighted in turquoise.

3.6 Expression of IFITM1, 2, and 3 in Macrophages

Macrophages are important mediators of the innate immune response and produce proinflammatory cytokines in response to viral infection. As such, these cells were chosen to investigate the levels of endogenous and IFN-inducible IFITM1, 2, and 3 to determine if they are a suitable cell line for further investigation.

Expression of *IFITM1, 2,* and *3* was determined by PCR using primers designed to unique sequence stretches in each gene, allowing specific amplification, and therefore differentiation, of these similar genes (Figure 27). RNA was extracted from monocyte-derived macrophages (MDMs) (kind gift of Prof. Mark Marsh) that had been infected by HIV-1 BaL at an MOI of 3 with or without additional IFNβ treatment.

Without IFN stimulation or infection, MDMs expressed low levels of *IFITM2* and *3,* but no *IFITM1* was detected by RT-PCR. A dose response was not detected when an increasing amount of IFN was added to the cells, but saturation may have been reached at 2 ng. However, there was a substantial upregulation for all three genes when IFNβ was added compared to unstimulated, uninfected cells. Using ImageJ software, *IFITM* gene expression could be semi-quantified, allowing cross-comparisons. 2-fold less *IFITM2* was produced compared to *IFITM1* and *3.* Comparison of ‘uninfected +IFN’ cells to ‘infected –IFN’ cells showed that IFNβ induced 23 times as much *IFITM1* than did HIV-1 infection. A 6-fold difference and a 2-fold difference in induction between IFNβ and HIV-1 was detected for *IFITM2 and IFITM3* expression respectively (Figure 28). HIV-1 infection caused 15 times as much *IFITM3* expression as *IFITM1.* However, it is important to note that these calculations are based on terminal stage PCR quantification and therefore do not reflect rates of increase or potential saturation of rate-limiting reagents.

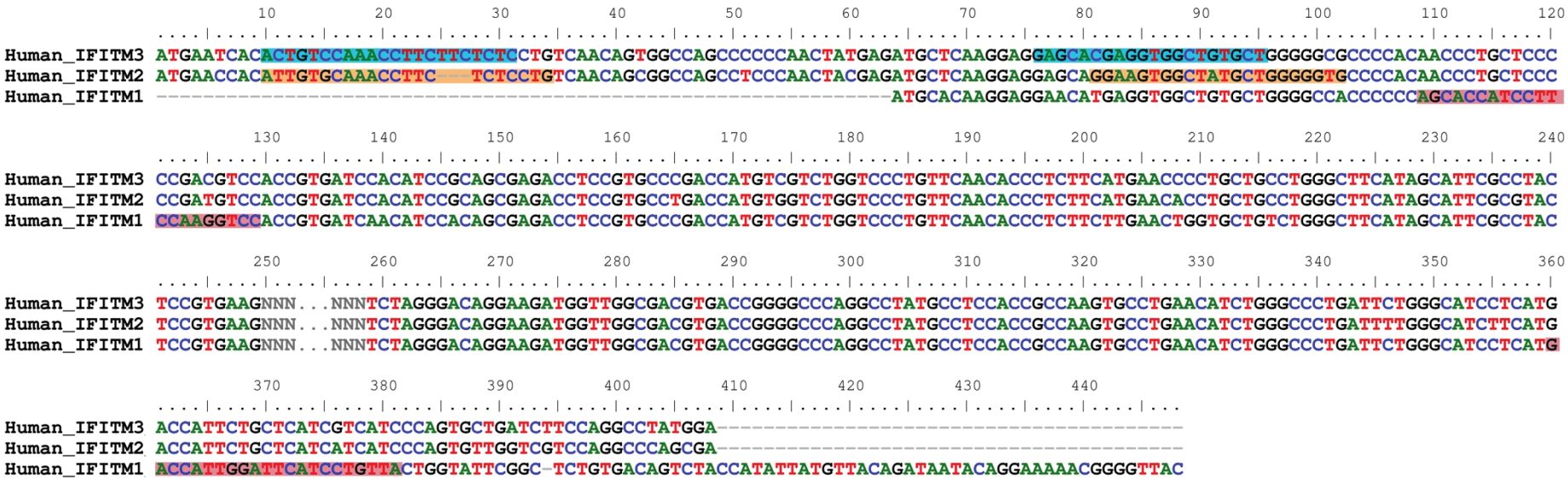

Figure 27: Primers used to distinguish between and amplify human *IFITM1*, *2*, and *3*

A multiple sequence alignment of *IFITM1, 2,* and *3* showing regions of nucleotide mismatches. Forward and reverse primer pairs were designed to cover these regions and specifically amplify each gene. Primer sets for *IFITM3* are shown as blue boxes, sets for *IFITM2* are shown as orange boxes, and sets for *IFITM1* are shown as pink boxes. Introns are denoted by a string of grey Ns.

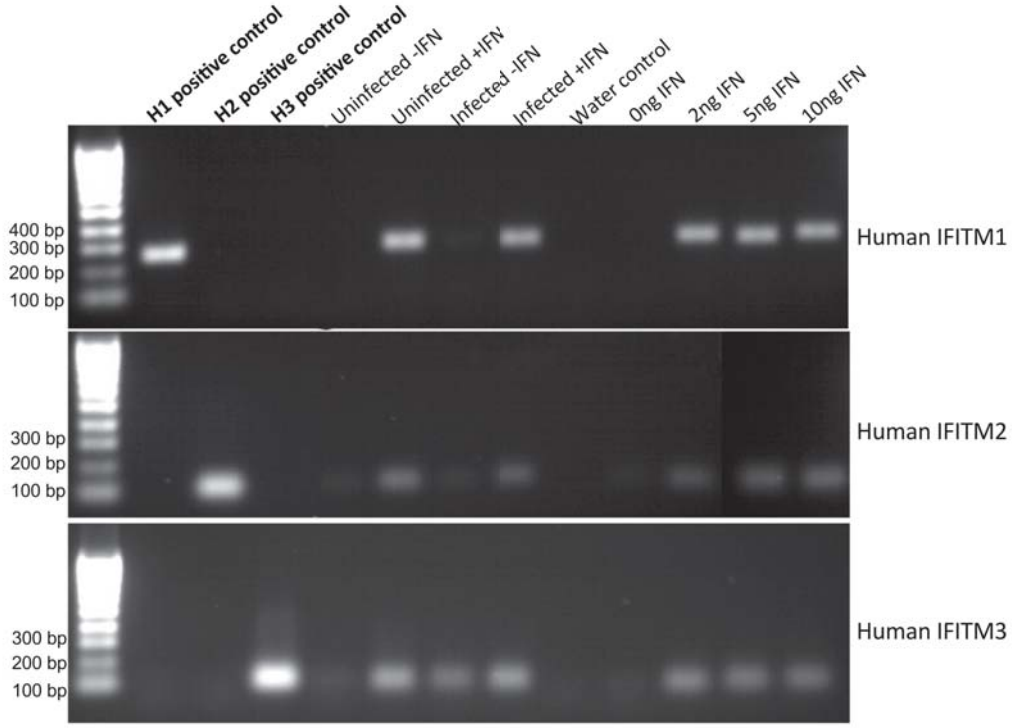

Figure 28: *IFITM* expression in macrophages

RT-PCR was carried out on RNA generated from macrophages under different treatment conditions. Macrophages were infected with HIV-1 or stimulated with IFNβ. Primers were designed to amplify, and differentiate between, *IFITM1, 2*, and *3*. Plasmids encoding *IFITM1, 2*, or *3* coding sequences were used as a positive control for each PCR reaction (H1, H2 or H3 positive control).

3.7 Detecting an Alternative IFITM3 Transcript in Macrophages

The existence of the IFITM3_004 transcript is supported by RNAseq reads from both adrenal and blood tissues (Figure 29). The position of a classical intron between exon 1 and 2 is well supported (as indicated by a large number of stacked 'reads'), however, as well as this intron, there are a number of long reads between exon 1 and the alternative 5' UTR. This suggests that splicing may occur downstream of this UTR. In addition, *in silico* analysis conducted by Ensembl indicates an additional promoter with transcriptional start site motifs around the alternative 5' UTR (Figure 29). These regions were identified by using two segmentation programs, ChromHMM and Segway[242,243], that detect motifs associated with open chromatin, transcription factors and histone modifications.

Using the same oligodT cDNA synthesised for the previous experiment, another PCR was designed to detect and amplify IFITM3_001 and IFITM3_004 (Figure 30) and a RT-PCR was performed (Figure 31). The no RT control shows that DNA was effectively removed from the samples before cDNA synthesis (Figure 31B). Consistent with Figure 28, full-length *IFITM3* was amplified from macrophages. Infection or IFNβ stimulation had a similar effect on the upregulation of *IFITM3*, and a dose response to IFNβ was undetectable. However, no bands were detected for the PCR using the alternative splice forward primer (Figure 31C).

We obtained an alternative source of monocyte RNA (THP-1 cells - a kind gift of Greg Towers) that had been treated with IFN-β, in order to repeat this PCR (Figure 32). Four amplicons between 450 bp and 1500 bp were amplified during the reaction. The remainder of the PCR reaction was separated by electrophoresis and the bands extracted, purified and sequenced by capillary sequencing. BLAST analysis of the sequencing reads showed that the bands represented random amplification of the genome, which had no sequence similarity to *IFITM3*. Therefore we were unable to detect IFITM3_004 in these samples.

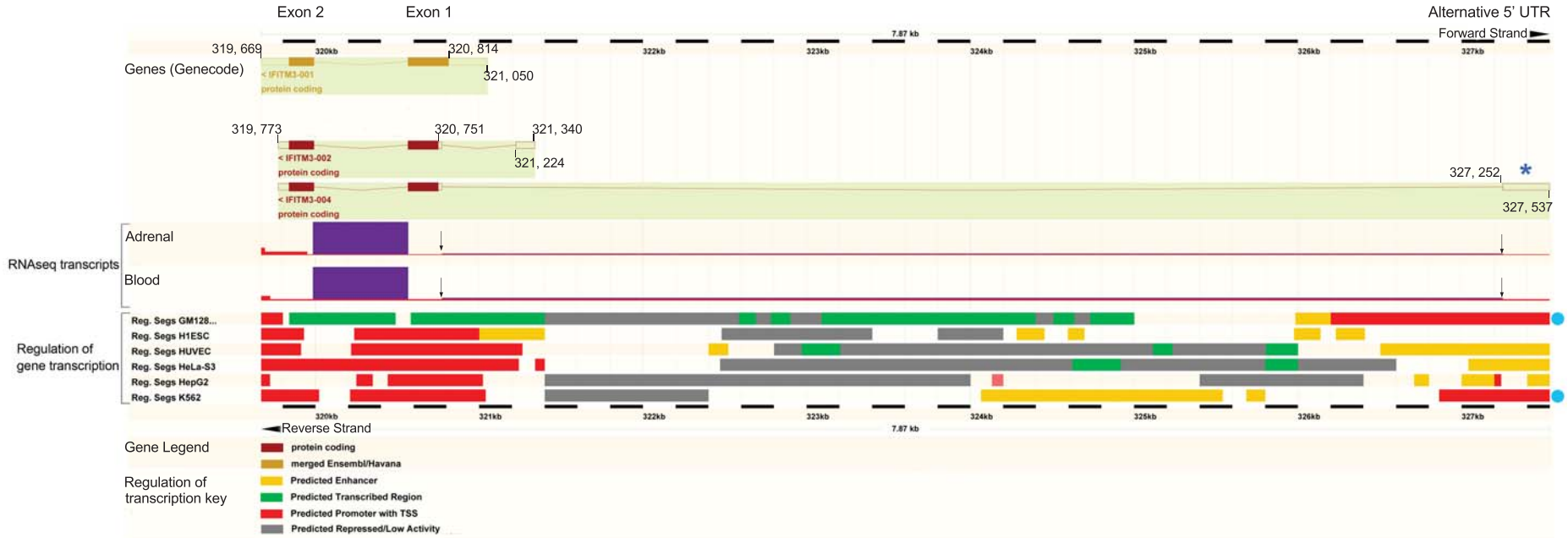

Figure 29: Evidence for alternative human *IFITM3* transcripts

Using the human genome build version 74.37, the full-length protein coding IFITM3 transcript is shown in gold (IFITM3_001) and alternative protein coding transcripts are shown in red. The existence of transcript IFITM3_004 is supported by RNAseq of transcripts from mRNA extracted from adrenal and blood samples. The height of the purple region is proportional to the likelihood of an intron being present. The position of the intron between exon 1 and 2 is well supported. As well as this intron, there are a number of long reads between exon 1 and the alternative UTR (*) denoted by black arrows. These reads suggest the presence of a further intron. The red bars represent the regulatory elements and show predicted promoter regions with transcriptional start sites. Some of these reads accumulate at the position of the alternative 5' UTR suggesting an alternative transcriptional start site, highlighted with blue circles.

Amplifying IFITM3_001

TCCTCAGGAATTTGTTCCGCCCTCATCTGGCCCCGGCCAAATCCCGATTTGACAAATGCCAGGAAAAG
GAAACTGTTGAGAAACCGAAACTACTGGGGAAAGGGAGGGCTCACTGAGAACCATCCCAGTAACCCGA
CCGCCGCTGGTCTTCGCTGGACACCATGAATCACACTGTCCAAACCTTCTTCTCTCCTGTCAACAGTG
GCCAGCCCCCCAACTATGAGATGCTCAAGGAGGAGCACGAGGTGGCTGTGCTGGGGGCGCCCCACAAC
CCTGCTCCCCCGACGTCCACCGTGATCCACATCCGCAGCGAGACCTCCGTGCCCGACCATGTCGTCTG
GTCCCTGTTCAACACCCTCTTCATGAACCCCTGCTGCCTGGGCTTCATAGCATTCGCCTACTCCGTGA
AGGTGCGTATGGCCCCAGGGAATGCTCAGAGGGTGCCGCTGAGCCTGGAGCTCCACCTGCCCACATGC

Predicted size: 204 bp

Amplifying IFITM3_004

TCCTCAGAGCGCAGCCAGGCCAGAGGCTGCACCGAGGTGCAGAATCAGAGGAGGCACCGGAGACCCCA
GAGTCCAGTCTGAGACGGCACAGGGAGCAGGTCTCTGGTGGCCTTGACAAGCTCCAGGATAGGGTGGG
GAGGGGACTGGACCCTGGGGACCTCAGAGCAGAGCAGGGGAAACAGGAGCCCCCACCTGGGGAGAGGG
GGCCTCCTCTCCAGGAACCCCAATCAAGACGAGCCTCACGTGACTCCCACTTCCTCTTGGAGGGTGCA
GGGGCCTCTCCTGAGGTGAGTTTT……TCAACAGTGGCCAGCCCCCCAACTATGAGATGCTCAAGGAGG
AGCACGAGGTGGCTGTGCTGGGGGCGCCCCACAACCCTGCTCCCCCGACGTCCACCGTGATCCACATC
CGCAGCGAGACCTCCGTGCCCGACCATGTCGTCTGGTCCCTGTTCAACACCCTCTTCATGAACCCCTG
CTGCCTGGGCTTCATAGCATTCGCCTACTCCGTGAAGGTGCGTATGGCCCCAGGGAATGCTCAGAGGG
TGCCGCTGAGCCTGGAGCTCCACCTGCCCACATGCTGCCTGGGGTGGGGACTTGTGTGTCCCTGTGAC

NNN = Exon1 F3

NNN = Alternative_transcript_1

NNN = Exon1 R2

NNN = Start codons

NNN = Intron

T = rs12252

Figure 30: Primers for amplifying IFITM3_001 and IFITM3_004

Primers to differentiate between IFITM3_001 and IFITM3_004 were designed to amplify part of exon 1. The forward primer for 001 (in yellow highlighter) begins 15 bp upstream of the start codon and the reverse primers binds upstream of the intron (blue text), amplifying cDNA of 204 bp. The forward primer to identify IFITM3_004 (in red text) binds to the DNA more than 6 kb upstream of the start codon and the same reverse primer (blue text) was used. Successful splicing of IFITM3_004 would result in a 374 bp amplicon.

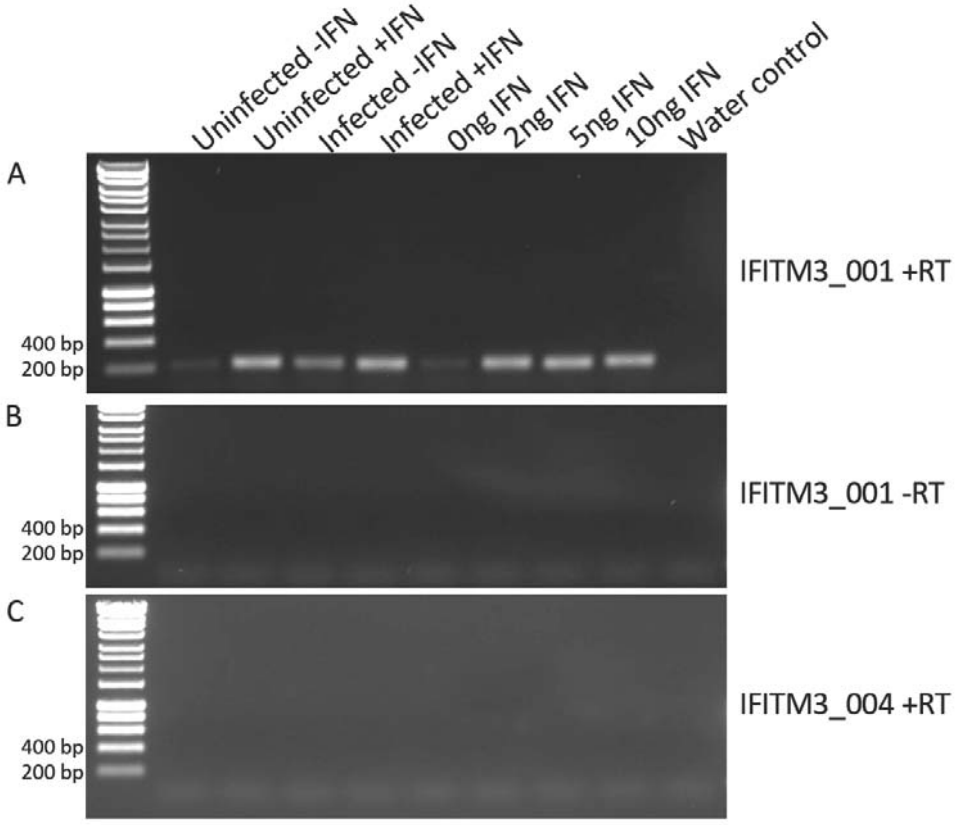

Figure 31: PCR of IFITM3_004 on macrophage cDNA

RNA was extracted from macrophages treated with varying levels of IFN or HIV-1 infection (MOI 3). cDNA was synthesised using oligodTs and a PCR carried out with primers specific for full-length IFITM3 +RT (A), without RT (B), or the alternative transcript (C).

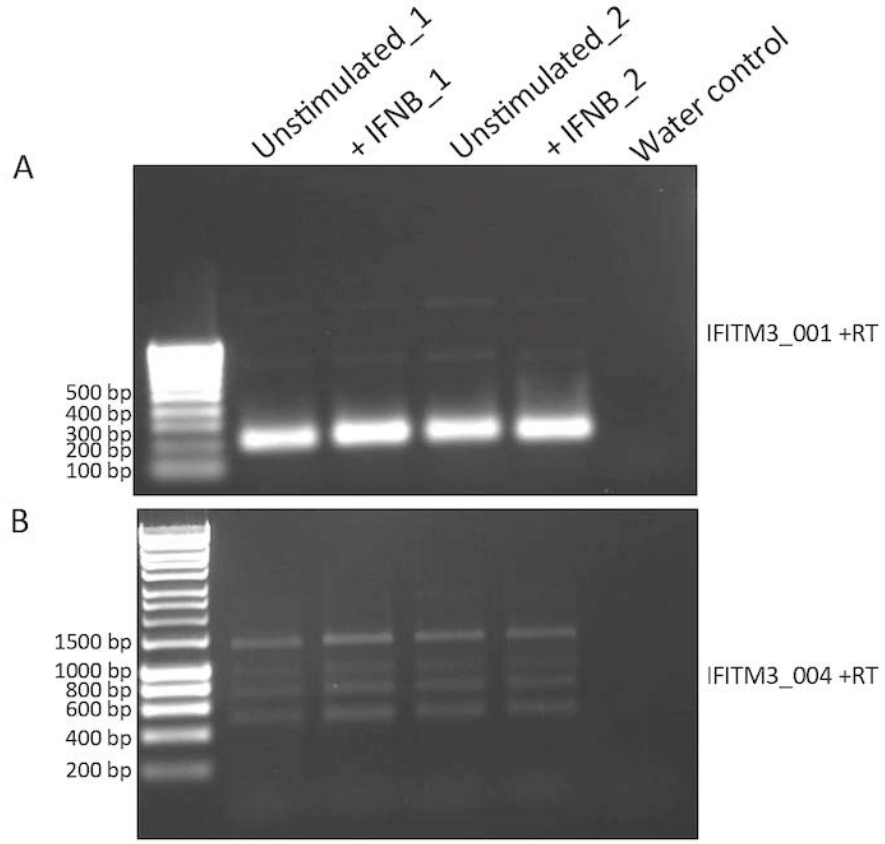

Figure 32: PCR of IFITM3_004 on monocyte cDNA

RNA was extracted from THP-1 cells treated with IFNβ or unstimulated. cDNA was synthesised and a PCR carried out using primers specific for IFITM3_001 (A) or the alternative transcript IFITM3_004 (B). Biological duplicates (1 & 2) are shown.

3.8 Detecting an Alternative Transcript of IFITM3 in Primary Airway Epithelial Cells

Splicing of alternative transcript IFITM3_004 was not detected in macrophages using this assay, however in the absence of positive controls, it is difficult to assess the significance. We hypothesised that splicing would be more apparent in primary cells of the airway epithelium, as this is more consistent with cells of the lung.

Primary airway epithelial (PAE) cells were obtained from LGC Standards, and genotyped using a set of specific primers (Figure 33), revealing that the PAEs were homozygous TT for rs12252. PAEs were treated with IFNα or PBS for 24 h prior to RNA and protein extraction. qRT-PCR was carried out on 100 ng of RNA using the same primers to amplify the full-length IFITM3_001 transcript and the alternatively-spliced transcript IFITM3_004 (Figure 34).

Amplification of IFITM3_001 produced a cycle threshold (Ct) of 20 in unstimulated cells and this reduced to a Ct of 18 upon IFNα stimulation, indicating that the PAEs are IFNα sensitive and can upregulate full-length *IFITM3* (Figure 34A). This is supported by a Western blot showing that IFITM3 is present in unstimulated cells and increases by approximately 3-fold following IFNα stimulation (Figure 34C). The primers to amplify the alternative transcript IFITM3_004 also generated a product at Ct 33, which reduced to Ct 30 after addition of IFNα. Although the high Ct suggests that the transcript has low abundance, the non-template control for these primers produced a Ct of 37.

These Ct values were used to estimate the number of copies of each transcript per 100 ng of input RNA, using the standards for full-length IFITM3 (Figure 35). However it is important to note that standards were not available for the alternative transcript and thus these calculations are based on the assumption of equal PCR efficiency. Unstimulated PAEs transcribed 9.32×10^3 copies of IFITM3_004 compared to 1.42×10^6 copies of IFITM3_001. The abundance of IFITM3_004 increased 3-fold to 3.16×10^4 copies after IFN stimulation, whereas the abundance IFITM3_001 increased by a more modest 2-fold after IFN stimulation (Figure 35).

SYBR green assays cannot differentiate between specific and non-specific amplification, so the IFITM3_004 PCR products were separated on an agarose gel (Figure 34B), showing a single product of the predicted size (374 bp). The remaining

Genotyping IFITM3

AACTGTTGAGAAACCGAAACTACTGGGGAAAGGGAGGGCTCACTGAGAACCATCCCAGTAACCCGACC
GCCGCTGGTCTTCGCTGGACACCATGAATCACACTGTCCAAACCTTCTTCTCTCCTGTCAACAGTGGC
CAGCCCCCCAACTATGAGATGCTCAAGGAGGAGCACGAGGTGGCTGTGCTGGGGGCGCCCCACAACCC
TGCTCCCCCGACGTCCACCGTGATCCACATCCGCAGCGAGACCTCCGTGCCCGACCATGTCGTCTGGT
CCCTGTTCAACACCCTCTTCATGAACCCCTGCTGCCTGGGCTTCATAGCATTCGCCTACTCCGTGAAG
GTGCGTATGGCCCCAGGGAATGCTCAGAGGGTGCCGCTGAGCCTGGAGCTCCACCTGCCCACATGCTG
CCTGGGGTGGGGACTTGTGTGTCCCTGTGACTGTGAGTTTGTGTGCACCTCTGTCCCGTGTGTGCCCA
CGTCAGTGGCTTTGTCTGTGTGATCTGTGTGTGTGTGTGGCTTGGGGAATCTGCCCAGTGCAGGTTTA

Predicted size: 282 bp

NNN = Forward primer

NNN = Reverse primer

NNN = Start codon

NNN = Intron

T = rs12252

Figure 33: Primers for genotyping *IFITM3* at rs12252

Primers were designed to specifically amplify *IFITM3* and allow identification of the allele at rs12252. Both forward and reverse primers had 8 mismatches with human *IFITM2*, to minimise the likelihood of non-specific amplification.

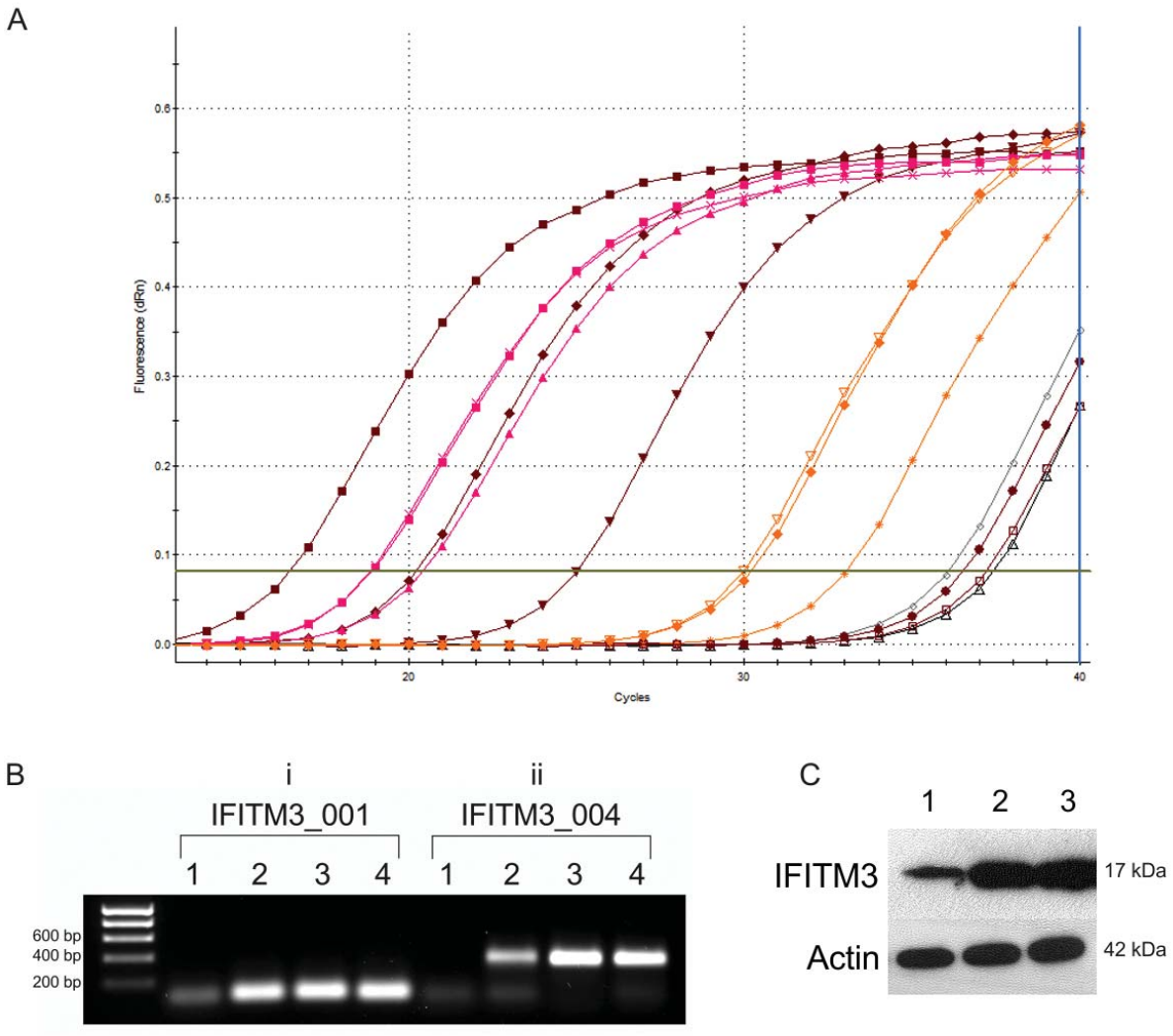

Figure 34: qRT-PCR of RNA from primary airway epithelial cells stimulated with IFNα

Primary airway epithelial (PAE) cells were treated with IFNα or left unstimulated and the RNA harvested 24 h later. (A) The RNA was reverse transcribed and a SYBR green assay performed with primers specific for IFITM3_001 (pink) or the alternative transcript, IFITM3_004 (orange). Standards showing copy numbers of IFITM3_001 are shown in red (■,10^7; ◆, 10^6; ▼, 10^5; ●, 10^4; □, 10^3). No template controls (○) were used in both cases . X, full-length + 2000 units IFN; ■, full-length + 200 units IFN; ▲, full-length untreated; ▽alternative transcript + 2000 units IFN; ◆, alternative transcript + 200 units IFN; *, alternative transcript untreated. (B) The full-length (i) and alternative transcript (ii) PCR products were separated on an agarose gel. 1= no template control, 2= untreated PAE, 3= PAE with 200 units IFN, 4= PAE with 2000 units IFN. Protein was also extracted from PAEs (C) – untreated (1), with 200 units IFN (2) and with 2000 units IFN (3). Cell lysates were probed for IFITM3_001 and β-actin. Predicted sizes: full-length transcript = 204 bp, alternative transcript = 374 bp.

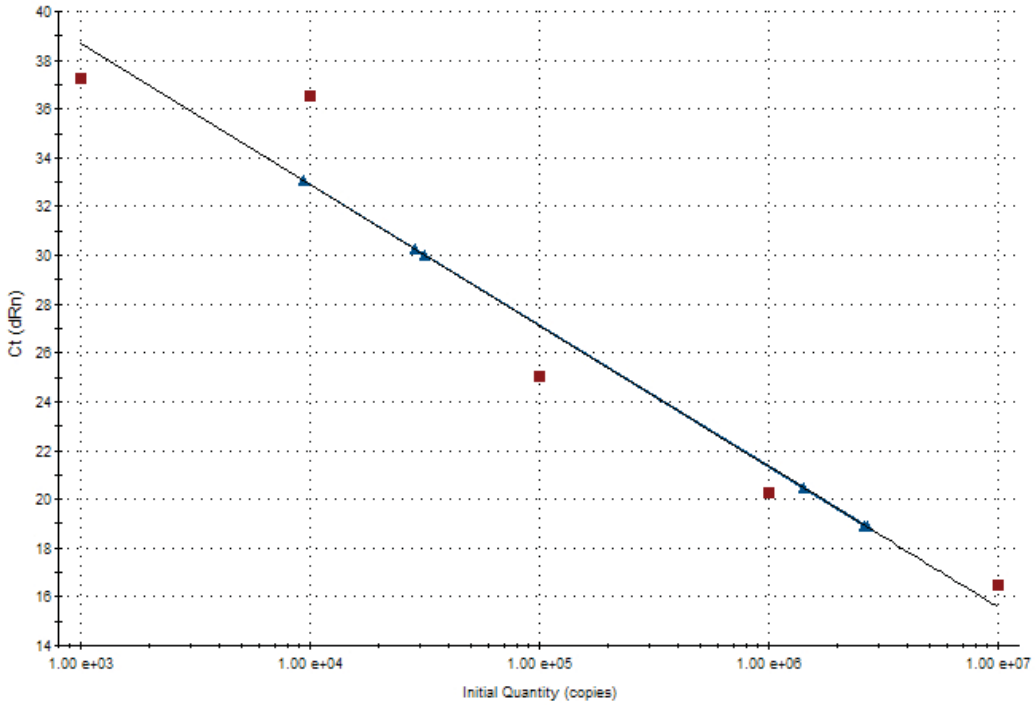

Figure 35: Standard curve to calculate the quantity of human IFITM3 transcripts

Five standards for IFITM3_001 (10^7 to 10^3 copies; ■) were analysed by SYBR green qRT-PCR alongside RNA extracted from unstimulated PAEs and PAEs stimulated with 200 units of IFN or 2000 units of IFN(▲). R^2=0.939.

PCR product was purified and the sequence determined by capillary sequencing (Figure 36). Analysis confirmed that the alternative transcript was authentic (Figure 36B) and that rs12252 is adjacent to an active splice acceptor site, whose splice donor is over 6 kb upstream.

The region a further 10 kb upstream of the splice donor (Chr11:327530 – 337000) was analysed for open reading frames (ORFs) in all six transcriptional frames using the NCBI software ORF Finder. Seven open reading frames encoding proteins of greater than 100 amino acids were identified and the sequences used in multiple BLASTp searches against the human reference (taxid 9606). Other than the truncated IFITM3 protein no other ORFs encode proteins with significant similarity to human proteins, or have conserved structural domains. Furthermore, analysis of the sequence upstream and in frame with the putative Met start site shows no alternative start sites and no possibility of an N-terminal extension of IFITM3.

As discussed previously, automatic regulatory analysis software included in Ensembl (ChromHMM and Segway) (Figure 29) predicts that the region around Chr11:3272500 has promoter activity. Further *in silico* analysis was performed on the 10 kb upstream region before to the splice donor of IFITM3_004 (Chr11: 327252 – 337000) using an online bioinformatic tool called TSSW[244], which applies motif recognition algorithms to detect human pol II promoter regions. This software identified three TATA box motifs at positions -6670 from start site (TATAAAA), -7810 (ATATAAA) and -8197 (ATATAAA). Interestingly a TATA box and CAAT motif were identified at was position -1901 and -2141 respectively in full length IFITM3_001, however a classic CAAT motif (GGCAATCT) was not present in the 10 kb upstream of IFITM3_001, therefore it wasn't appropriate to expect this for IFITM3_004.

Since this transcript increases in abundance after IFN stimulation, ISRE motifs were also used as search terms. The consensus sequence for an ISRE is GAAANNGAAAG/CT/C[245] or its reverse complement. Two ISREs are centered around position -77 and -94 from the start site of IFITM3_001 (Figure 22). However no sites matching this motif are identifiable in the -10 kb region in IFITM3_004, but many copies of the core region (TTTNNNTTT or AAANNNAAA)[246] are present around the TATA box at position -6670. It is difficult to ascertain the confidence of these binding sites without carrying out Chipseq experiments.

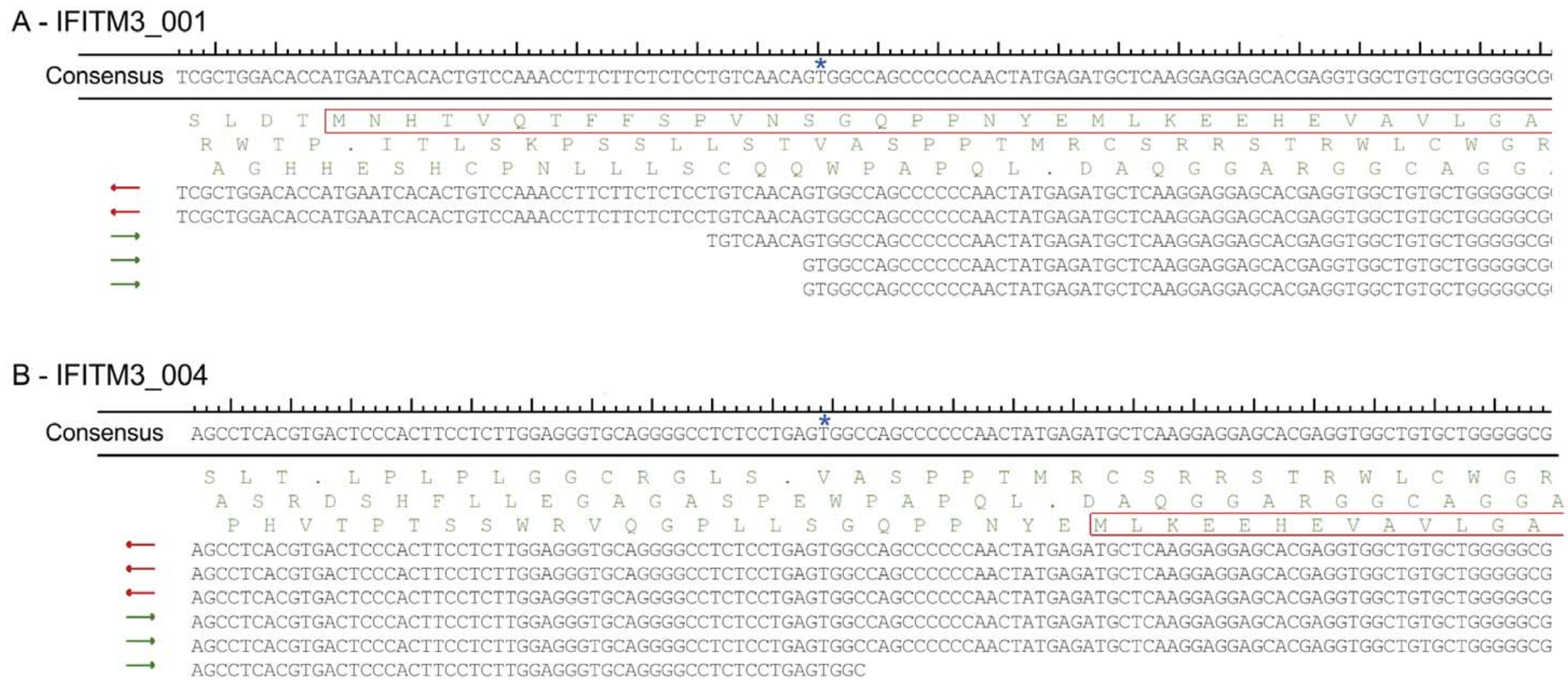

Figure 36: cDNA Sequence alignment of IFITM3 transcripts in primary airway epithelial cells

RNA was extracted from PAE cells, reverse transcribed, and a PCR carried out using primers specific for ‘full-length’ IFITM3_001 (A) or the alternative IFITM3_004 transcript (B). Rs12252 is highlighted by *, which is adjacent to the splice donor (AG) and marks the transition from the alternative 5’ UTR and exon 1 in the splice variant. Coding regions are shown by a red box. Red and green arrows indicate individual sequencing reads.

3.9 Testing the Functional Impact of rs12252 in LCLs

Although splicing was detected in PAEs, we were only able to obtain cells of the homozygous TT genotype. Therefore we cannot associate the allele at rs12252 with any change in the proportion of full-length or alternative transcripts.

In order to test our hypothesis that alternative splicing would occur more in CC homozygotes than in TT homozygotes, we used LCLs from the HapMap project. These cell lines come from a broad range of ethnically diverse donors and have all been extensively genotyped. Nine cell lines that were assigned as either homozygous TT or CC, or heterozygous for SNP rs12252 were identified for further study: GM11994 (TT), GM12154 (TT), GM12155 (TT), HG00524 (TC), HG01108 (TC), HG00478 (TC), HG00533 (CC), HG00530 (CC), HG00557 (CC).

The LCLs were re-sequenced using the methods described previously (Figure 33) and their genotype at rs12252 was confirmed. The cells were stimulated with IFNα2 and the level of IFITM3 produced by each cell line compared by Western blot. We hypothesised that the level of expression of IFITM3 would be lower in the CC homozygous cell lines, because of the predicted lower proportion of full-length transcripts.

The expression of IFITM3 was significantly induced in all cell lines 24 h after IFNα stimulation (Figure 37A), regardless of the genotype for rs12252. There was variation in the amount of IFITM3 produced by each cell line, but this did not correspond to the rs12252 genotype. For instance when comparing all three CC homozygous cell lines (Figure 37 6-8), constitutively-expressed levels of IFITM3 are different, as are the IFN-induced levels. There are numerous other genetic differences between the cell lines, aside from rs12252, which makes such a comparison difficult. The β-actin loading controls are not as consistent as they could be, however control Western blots of A549s over-expressing IFITM1, 2, and 3 show that IFITM2 is also detected by the N-terminal anti-IFITM3 antibody (Abgent) (Figure 37B). This is because of shared sequence identity at the N-termini of IFITM2 and IFITM3. Therefore without an IFITM3 specific antibody, this experiment cannot be interpreted. Moreover the N-terminal antibody cannot distinguish between full-length IFITM3 and an N-terminal truncated protein (Figure 45).

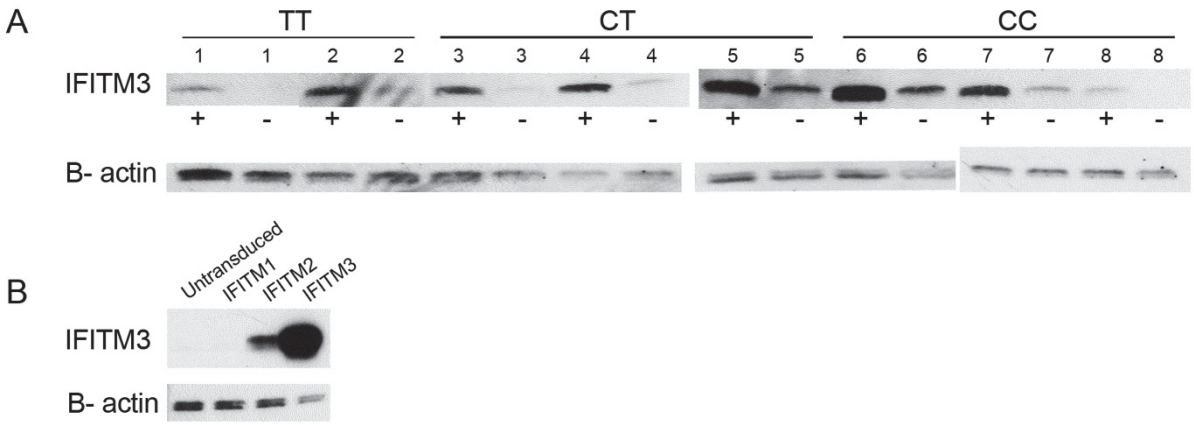

Figure 37: Induction of *IFITM3* expression in LCLs stimulated with IFN

A) LCLs were treated for 24 h with IFNα (+) or with control PBS (-), cell lysates harvested and IFITM3 detected by Western blot (Abgent antibody). Cells were genotyped for SNP rs12252 and were homozygous TT (1 and 2), heterozygous (3-5), or homozygous CC (6-8). β-actin detection was used as a loading control. LCL numbers are as follows: 1= GM11994; 2 = GM12155; 3 = HG00524; 4= HG01108; 5 = HG00478; 6 = HG00533; 7 = HG00530; 8 = HG00557. B) The anti-IFITM3 antibody was tested for specificity on untransduced A549s or A549s over-expressing IFITM1, 2, or 3.

3.10 Detecting an Alternative IFITM3 Transcript in LCLs

Since confounding factors made detecting splicing at the protein level difficult, we sought evidence of the alternative IFITM3_004 transcript expression in the these LCLs. RNA was extracted from LCLs grown in culture, treated for 24 h with IFNα2b or PBS. One-step qRT-PCR was carried out using primers to amplify the full-length transcript (IFITM3_001) and the alternative transcript IFITM3_004 (Figure 38).

The full-length and alternative transcripts were amplified in all cell lines tested, with a moderate induction after IFNα stimulation. IFN stimulation had a larger effect on the expression of transcript IFITM3_001 (average Ct decrease of 2.28) compared to IFITM3_004 (average Ct decrease of 1.22).

In all cell lines, IFITM3_001 was expressed at higher levels than the alternative transcript, on average 7.3 Cts different. However, there was no significant difference in the expression of endogenous IFITM3_001 transcripts between cells with a CC or TT genotype, and in general LCLs homozygous for the C allele had a small increase in expression of the IFITM3_001 transcript (Ct=26.9 compared to Ct=27.6).

A similar pattern was detected for the alternative transcript IFITM3_004; cells homozygous for the T allele averaged higher expression of this transcript than the cells homozygous for the C allele. There was also no significant difference in the amount of upregulation of IFITM3_004 after IFN induction between both groups.

Therefore although we see variation in the level of IFITM3 expression in different LCLs, we cannot ascribe these differences to the allele at rs12252 as there is as much variation within the groups as between them.

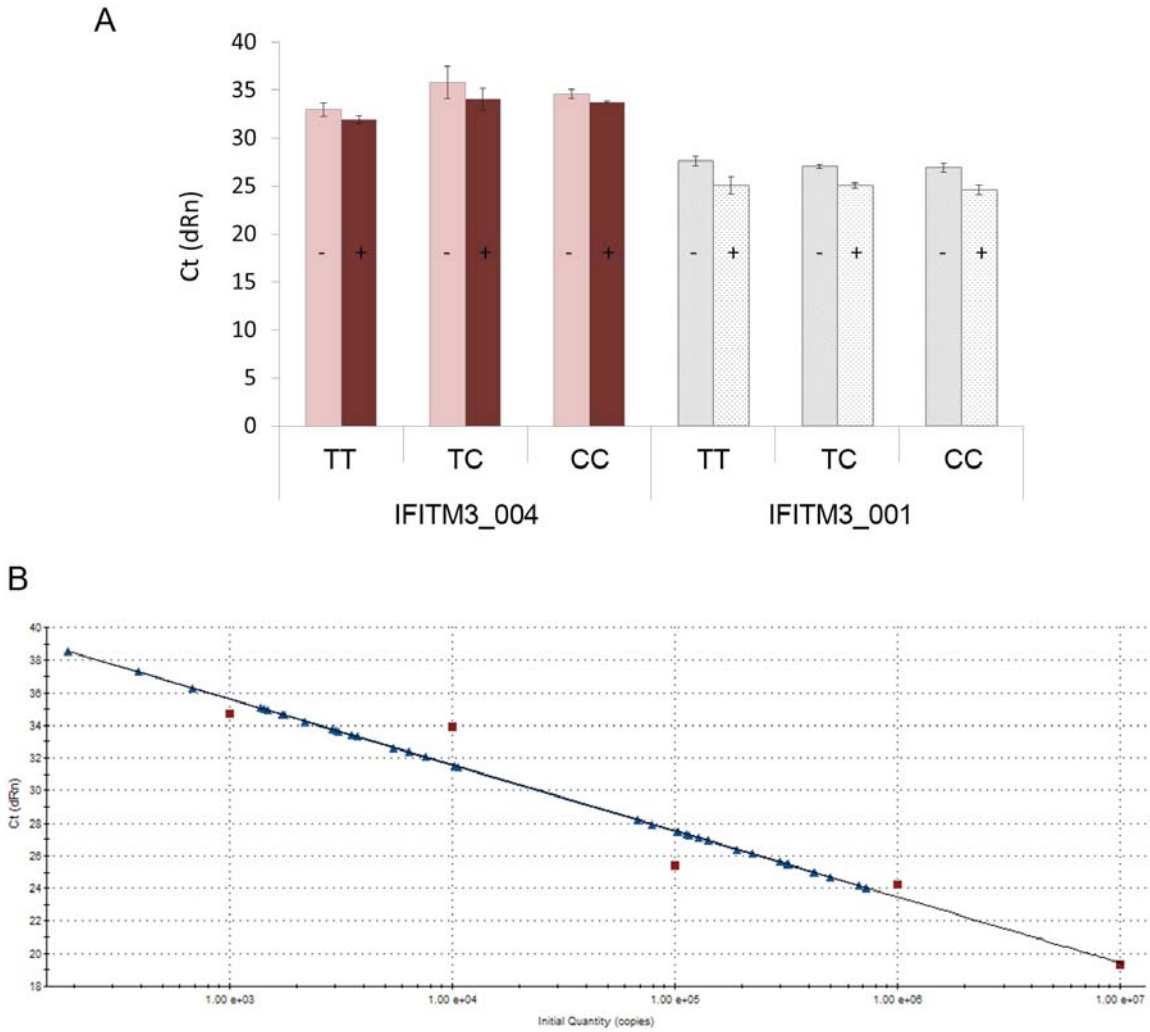

Figure 38: qRT-PCR of *IFITM3* transcripts in LCLS after treatment with IFNα

A) RNA was obtained from nine LCLs treated with IFNα (+) or PBS (-) for 24 h. Primers designed to amplify full-length IFITM3 (IFITM3_001 [grey]) and an alternatively-spliced transcript (IFITM3_004 [pink]) were used. Cts represent the number of cycles at which the fluorescence intensity for each primer pair breached an arbitrary threshold. Error bars represent standard deviation about the mean (n=3). B) A standard curve was derived from five standards encoding IFITM3_001 (10^7 to 10^3 copies [■]) and indicates the inferred copy numbers per 50 ng of input RNA for all LCLs (▲). R^2=0.936.

3.11 Alternative Transcripts of IFITM3 in Human Lung Tissue Sections

As no commercial PAEs homozygous for the rs12252 C allele were available, we established a collaboration with Professor John Nicholls from the University of Hong Kong to examine IFITM3 expression in lung sections from 22 lung cancer patients undergoing lung lobe resection.

DNA and RNA from fixed sections of tissue from the normal regions of resected lung lobes were extracted and purified. The patient samples were genotyped at rs12252 using the primers to amplify exon 1 (Figure 33). The amplicons were sequenced by capillary sequencing: 41 % of the samples were homozygous for the C allele, 45 % were heterozygous, and 14 % were homozygous for the T allele. These numbers do not deviate from Hardy-Weinberg equilibrium ($p=0.829$) and are comparable to the genotype frequencies observed in the Japanese population in 1000 Genomes phase 1 data (38 %:42 %:20 %, respectively, $p=0.215$).

Representatives from each genotype were analysed for IFITM expression and RNA extracted from LCLs was used as a positive control. As the tissue sections had been fixed, a random priming method was used to maximise the amplification of small, degraded fragments of RNA. However IFITM3_001, IFITM3_004, and GAPDH could not be amplified from the RNA extracted from the fixed tissue (data not shown), suggesting the RNA was too degraded by the fixation process for amplification.

3.12 Immunohistochemistry on Human Lung Tissue Sections

The tissue sections were suitable for immunohistochemistry, therefore we examined if patients homozygous for the T allele at rs12252 express more IFITM3 in their lung tissue than patients homozygous for the C allele. Fixed tissue sections were stained for IFITM3 using a monoclonal anti-IFITM3 antibody (Abnova, H00010410-M01) suitable for immunohistochemistry (Figure 39). This showed significant staining for macrophages (marked with a black arrow) and some light staining for epithelial cells (marked with a red arrow) (Figure 39). However, there was no obvious difference in the amount of IFITM3 detected in homozygous CC (Figure 39 A) or TT (Figure 39 C and D) tissue types.

Two additional anti-IFITM3 antibodies were tested for immunohistochemistry (Abgent and LifeSpan Biosciences) but no staining was observed (data not shown), which was difficult to interpret because of the lack of a positive control. However, the antibody from Abnova not been previously used in the literature and its cross-reactivity with IFITM2 or other proteins was unknown. The Abnova antibody was tested by immunofluorescence on A549 cells over-expressing HA-tagged IFITM1, 2, or 3. Apparent expression of IFITM3 was detected throughout the cytoplasm of all three cell lines (Figure 40). However, co-staining with an anti-HA antibody (AbCam) showed little co-localisation. The pattern of protein expression detected by the Abnova antibody suggests it may have cross-reactivity to a common cellular protein. This is supported by a Western blot using the Abonva antibody, which shows cross-reactivity to a protein of approximately 54 kDa (Figure 40E). Therefore the immunohistochemistry data cannot be interpreted and we cannot determine if the expression of IFITM3 is higher in patients carrying the TT or the CC allele at rs12252.

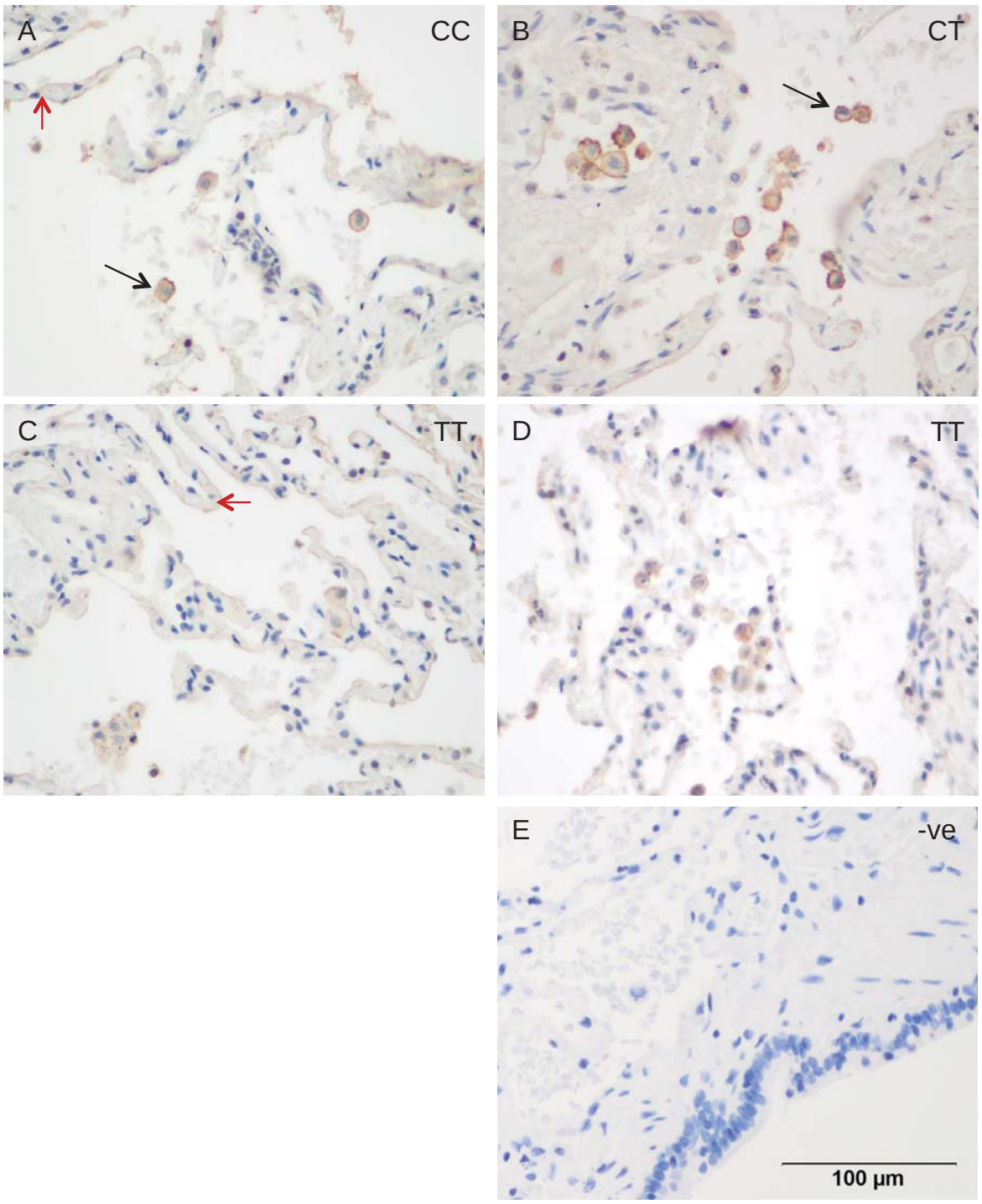

Figure 39: Immunochemistry of human lung tissue sections for IFITM3

Lung sections are from surgical specimens taken from the normal part of the lung when patients had a lobectomy for lung cancer. The code in the right hand corner of each panel represents the genotype, and '–ve' is a representative section stained with secondary antibody only. Nuclei are stained blue and brown cell membranes are positive for IFITM3. Black arrows show IFITM3-positive macrophages and red arrows show faint surface staining of epithelial cells. Images courtesy of Kevin Fung.

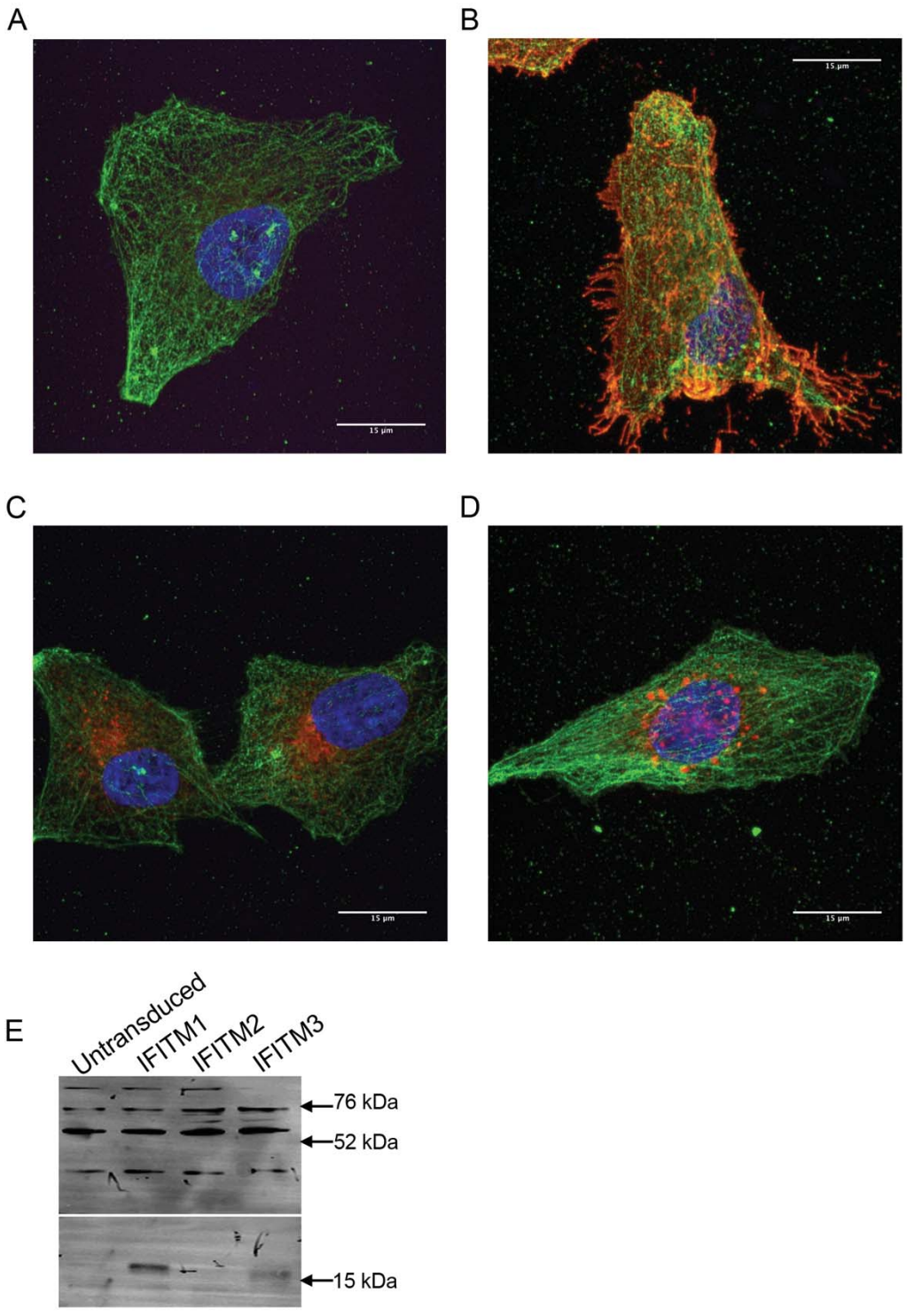

Figure 40: Testing the Abnova antibody on A549 cells stably expressing human IFITM1, 2, or 3

Untransduced A549s (A) or A549s over-expressing human IFITM1 (B), 2 (C), or 3 (D) were fixed and probed for IFITM expression using an anti-HA antibody (red) and an anti-IFITM3 antibody (green, Abnova). The Abnova antibody was used to probe cell lysates of cells over-expressing IFITM1, 2, and 3 and untransduced cells by Western blot (E).

3.13 Discussion of Results

3.13.1 Variation in Human IFITM3

53 Caucasian individuals infected with influenza during the 2009 pandemic in England and Scotland were genotyped at rs12252 to discover if any alleles for known SNPs were over-represented in this cohort, compared to an ethnically-matched general population. The results of this small study showed that the minority C allele at SNP rs12252 is over represented in this cohort of hospitalised influenza patients.

Independent studies have now corroborated these findings in different patient cohorts, and a summary of genotype frequencies are shown in Table 9. Zhang *et al.* (2013)[247] looked at the genotype of 83 Han Chinese individuals with mild or severe disease associated with A/H1N1/09 infection. The authors found that the CC genotype was present in 69 % of patients with severe disease symptoms compared with only 25 % in those with mild infection. This equates to a six-fold greater risk for severe infection in CC than the CT and TT individuals. Since the C allele is more prevalent in the Chinese population, this translates to a population-attributable risk of 54.3 %, meaning that 54.3 % of individuals could develop severe infection due to their CC genotype. This is a much larger fraction than for those of Northern European decent, which is 5.39 %.

Furthermore, Wang *et al.* (2013)[248] showed that of 16 patients hospitalised with influenza H7N9 virus, fatal outcomes were recorded in 33.3 % of CC, 28 % of CT, and none of the TT individuals. In addition to this, CC genotype patients were less able to control their infections; four of six patients had viral titres greater than $1x10^4$ pfu/ml, whereas this was only reported in one of the five heterozygous patients and again, none of the TT patients had such high viral titres. This ability to control infection was also reflected in the time to first methylprednisolone steroid treatment; CC individuals had the first dose after an average of 5.5 days, whereas TT patients took an average of 12 days. Mechanical ventilation was required by two thirds of the CC patients, but only a third of the TT individuals.

However, Mills *et al*. genotyped 34 individuals (self-reported Caucasians) with severe pneumonia associated with H1N1 (recruited to the Genomic Advances in Sepsis [GAinS] study) and compared these to 2730 community-acquired (mild) respiratory

Table 9: The allele frequency distribution for SNP rs12252 in different studies

Study		Allele Frequency		Genotype Numbers			Total Samples	Proportion of CC	p-value[1]
		C	T	CC	CT	TT			
Zhang	Severe influenza	0.813	0.187	22	8	2	32	69 %	$5x10^{-6}$
	Mild influenza	0.559	0.441	13	31	7	51	25.5 %	0.2
Wang		0.483	0.517	6	16	7	16	37 %	0.719
Bowles	CAL and KD	0.625	0.375	21	13	10	44	47 %	0.0004
Everitt		**0.094**	**0.906**	**3**	**4**	**46**	**53**	**5.66 %**	**0.003**

[1]Probabilitiy that the observed genotype frequencies deviate from Hardy-Weinberg Equilibrium (Fisher's Exact Test or Chi-squared test).

cases recruited across Europe in the Genomics to combat Resistance against Antibiotics in Community acquired LRTI in Europe (GRACE) study. Contrary to previous studies, the authors found an association between rs12252C and mild influenza symptoms[249], not severe symptoms. One difference between this study and previous studies is that a large reference set of individuals were genotyped directly for rs12252 (2730 mild cases and 2623 healthy matched controls). In previous studies rs12252 was imputed in the control groups. This means it was not sequenced directly, but the allele was predicted with high accuracy based on inherited haplotypes[250]. Mills *et al.* suggest this is a reason for the differences seen between these cohorts. No CC patients were seen in this cohort, but that is likely to be due to the smaller sample size (n=34) compared to the study in this thesis (n=53). It is also unclear what the clinical details of the 37 GAinS patients were; where possible patients with high body mass indexes and co-morbidities were ruled out of the study in this thesis. Furthermore with a disease like influenza, symptoms fall on a spectrum from asymptomatic to fatal, it is therefore difficult to establish whether or not those in the control 'healthy' population were actually infected but asymptomatic or pre-clinical in symptoms.

More recently, Bowles *et al.* show an association between the C allele of rs12252 and the onset of coronary artery lesions (CAL) in children suffering with Kawasaki Disease (KD)[251] p=0.0004 (Table 9). KD is a systemic vasculitis disease, which is particularly prevalent in Japan, affecting 218 children under 5 years old per 100,000[251]. Although the cause of Kawasaki disease is still unclear, Okano *et al.* have shown that prior infection with human herpes virus 6 or adenovirus is associated with an increased risk of developing the disease[252,253]. Significantly more patients homozygous for the SNP developed CAL than patients with the other genotypes (51.2% vs. 23.2%: p=0.001). The author tested several models for CAL in KD and found the recessive model was supported by the data. Therefore, the authors suggest that IFITM-susceptible viruses may play an etiological role in the development of CAL associated with KD.

3.13.2 Alternative Transcripts of Human IFITM3

As rs12252 is positioned next to a splice acceptor site it was hypothesised that the allele at this position could affect splicing of mRNA transcripts, and result in the

production of a truncated IFITM3 protein. Alternative splicing of innate immunity genes has been reported previously for the zinc-finger antiviral protein (ZAP) and TRIM5[59,133]. ZAP has two isoforms, the longer of which has greater antiviral activity against retroviruses than the shorter isoform. The α-isoform of TRIM5 in rhesus macaques has a strong antiviral effect on HIV-1, but the shorter γ-isoform does not. However it is important to note that this is alternative splicing of the same transcript Similarly, it has been shown by several groups that an artificially N-terminally truncated form of IFITM3 (ΔN-21) is significantly less effective at restricting replication of influenza virus and Vesicular Stomatitis Virus (VSV)[3,254], suggesting that the N-terminus is involved in anti-viral function. However this form of the protein had not been reported as occurring naturally. We aimed to find evidence at the transcript or the protein level for splicing of alternative IFITM3 transcripts.

Support for the IFITM3_004 transcript with an alternative 5' UTR was found in the RNAseq datasets and the regulation of gene transcription data, which were accessed via Ensembl (Figure 29). Splicing of this transcript could produce mRNA capable of encoding an N-terminally truncated protein that initiates translation at the second methionine (M22). As discussed previously, *in vitro* studies suggested that the N-terminus was essential for viral restriction[3,254], although the mechanism of IFITM3's action is still unclear. IFITM3 encodes a YEML motif in the N-terminal 21 amino acids, directly proximal to the second methionine. This motif is known to enable proteins to localise in endosomes, via the AP-2–clathrin-associated pathway[124]. IFITM3_004 would lose the YEML motif through use of Met22, potentially altering its subcellular localisation. Interestingly, IFITM1 does not have this motif and has been shown to be predominately expressed on the cell surface (section 4.7).

From RNAseq data, primers were designed to try and capture expression of the alternative IFITM3_004 transcript in several cell types. Macrophages, PAEs, and LCLs were infected or treated with type I IFN to promote ISG expression. Although IFITM3_001 was induced in all three cell types, IFITM3_004 could not be detected in macrophages. However, IFITM3_004 was detected by qRT-PCR in PAEs, and it was also found to be IFNα-inducible. Sequencing of the PCR products confirmed the presence of an alternative 5' UTR, which uses the splice acceptor 5' to rs12252. The basal level of the full-length transcript was much greater than that of the alternative transcript, but IFNα treatment only caused a 1.6 Ct decrease for IFITM3_001,

whereas IFNα treatment caused a 3 Ct decrease for IFITM3_004. The PAEs used in this study were TT homozygous for rs12252, which suggests that a low level of this transcript is transcribed, regardless of the alleles at rs12552. However, lack of availability of CC homozygote PAEs meant that we could not test whether or not increased transcription and splicing of IFITM3_004 occur in CC individuals. Further *in silico* analysis of the alternative 5' UTR showed several potential TATA box motifs 6 kb upstream of the potential start site, but no perfect ISRE binding sites were detected in a 10 kb region. ORF analysis did not identify any other potentially protein-encoding transcripts.

Large numbers of LCLs had already been genotyped as part of the HapMap project, so obtaining cells with the rare CC allele at rs12252 was possible. Although variation in the level of IFITM3 protein expressed was detected between different cell types, these differences could not be associated with SNP rs12252. Detecting differences in the abundance of IFITM3 at the protein level is difficult because of epitope overlap between IFITM2 and IFITM3. The Epstein-barr virus (EBV) used to immortalise the cells may well have an impact on the LCL transcriptome; EBNA3 proteins are known to impinge on host gene expression through recruitment of chromatin modifying proteins, such as histone deacetylases[255]. Since the EBV viral load of these cell lines is not determined it is difficult to establish the impact it could have on these experiments. Furthermore, although permissive to influenza infection, LCLs are derived from peripheral blood mononucleocytes, which are not naturally infected by respiratory viruses. Thus, we concluded that although LCLs were a convenient cell line, and had been used previously to investigate IFITM3 expression, they were not the most suitable *in vitro* model.

A cohort of 20 lung cancer patients in Hong Kong was genotyped and probed for expression of IFITM3_001 and IFITM3_004. Sequencing the *IFITM3* gene from these individuals showed that the spread of genotypes were 14:45:41 (TT:TC:CC). The genotype frequencies of rs12252 are not known in the Hong Kong population, but these ratios are in line with the known genotype frequencies in the Japanese population[247]. However, the RNA was too degraded to allow identification of IFITM3_004, and comparison of IFITM3 protein expression in these samples was prevented by the lack of discriminating immune reagents for immunohistochemistry.

Our current data suggests that although splicing of an alternative IFITM3 transcript can occur during transcription in PAE cells, it is unclear whether or not the rs12252 C allele has an impact on the control of splicing. Nevertheless, there is a clear association between this SNP and poor control of influenza during infection, resulting in severe symptoms. Since this synonymous SNP is not causing an amino acid change in the protein, perhaps it functions in a different way. Polymorphisms may also effect gene expression notably through control of DNA methylation by CpG islands or by affecting transcription factor binding sites. Rs12252 T – C change results in the formation of a CpG dinucleotide. Scott *et al.* described the methylation state of IFITM3 in two different human melanoma cell lines: D10 and ME15[256]. Although they both have identical core promoter regions for IFITM3[256], the former is IFNα insensitive (IFITM3 is constitutively expressed), whereas the latter is IFNα sensitive. The authors showed that application of 5'-aza-2'-deoxycytidine (a demethylating agent) onto ME15 cells causes an upregulation of IFITM3 after IFNα treatment. More specifically, the authors show that CpG motifs in the promoter of *IFITM3* are demethylated, promoting transcription. Scott *et al.* show that IFNα alone results in demethylation of the promoter region. Rs12252 could contribute to the methylation of this region of *IFITM3* or an alternative promoter region, or be in linkage disequilibrium with a true causal SNP, yet to be characterised.

Alternatively, the C allele of rs12252 may control ribosome movement along the mRNA transcript, or be tagging a SNP that does this. MAVS, an adapter protein involved in inducing the expression of anti-viral molecules, is known to encode a bicistronic transcript, which uses an alternative translation initiation site to produce a different protein[257]. The authors used ribosomal profiling to identify regions of ribosomal-protected RNA, in order to predict ribosomal start sites. Two peaks were detected for *MAVS*, suggesting two functional translational start sites. IFITM2 was also identified in this study as using an alternative ribosome binding site, which could produce an N-terminally truncated protein. Although this study did not identify IFITM3, it is possible that the C allele at rs12252 causes ribosomal stalling and a preferential use of the downstream start codon for IFITM3 also.

3.14 Conclusions

This thesis has shown that the CC allele at SNP rs12252 in *IFITM3* is associated with an increase in the severity of influenza infection in humans. Splicing of an alternative *IFITM3* transcript was detected in LCLs and PAE cells, but no association with the allele at rs12252 could be determined. To further research into changes in protein abundance in different cell types, antibodies that can differentiate between IFITM1, 2, and 3 are required. Unfortunately PAEs homozygous for the C allele at rs12252 were unavailable, but engineered cell types, either from induced pluripotent stem cells derived from LCLs, or using DNA editing techniques such as CRISPR, are also required to determine the significance of the C allele at this locus. Once these cell types are available splicing could be further investigated along with the methylation status of the IFITM3 promoter and ribosomal profiling studies to determine start site usage.