The Impact of Genetic Variation on *IFITM3* Related Influenza Virus Restriction



Henry Laurenson-Schafer
Kellogg College

Supervisor: Professor Tao Dong
Co-Supervisor: Professor Jan Rehwinkel

A Thesis Submitted for the Degree of
Doctor of Philosophy
in
Medical Sciences

Trinity Term 2018

The Impact of Genetic Variation on *IFITM3* Related Influenza Virus Restriction

Henry Laurenson-Schafer, Kellogg College
DPhil Thesis, Trinity Term 2018

The IFITM3 protein is a prominent restriction factor of influenza virus that has been shown to antagonise the cellular entry of influenza virus. In 2012, the minor allele of a synonymous SNP in *IFITM3*, known as rs12252 (major T, minor C), was shown to be associated with severe influenza virus infection. Numerous studies have corroborated this association, showing clear links between the C allele and increased risk of severe influenza virus infection in East Asian and European populations. It was hypothesised that the observed effects of rs12252-C were related to aberrant splicing of IFITM3, causing an isoform coding for an N-terminally truncated variant of the protein. More recently, additional IFITM3-associated genetic risk factors for severe influenza virus infection have also been uncovered.

In this thesis, we investigated the molecular mechanisms of rs12252. In chapter 3, we show that the *IFITM3* splicing profile is consistently dominated by the canonical isoform, and that truncated isoforms are produced at extremely low levels regardless of rs12252 genotype. In chapter 4, we investigated associations between rs12252 and expression of *IFITM3* at mRNA and protein level, using *in vitro* systems and healthy donors. We show that rs12252 is an expression Quantitative Trait Locus (eQTL) for *IFITM3*, and its activity is potentially tissue dependent. We also uncovered potential effects of this SNP on blood transcriptional signature. Next, in chapter 5, we identified two novel eQTLs for *IFITM3*, which may be risk variants for severe influenza virus infection. Finally, in chapter 6, we analysed the association between rs12252 and severe influenza infection using several model systems. These findings uncovered potential associations between rs12252 and differential interferon signalling, although further experimentation is required to confirm this.

Acknowledgements

Over the course of my PhD, I have had many fantastic experiences and opportunities. I attended conferences in Europe and North America, took part in an international collaboration between Oxford and China, and worked as an intern at the WHO Headquarters in Geneva. None of my achievements or these experiences would have been possible without the help of many people in and outside the WIMM who I would like to thank.

Firstly, I would like to express my gratitude towards the patients and donors who have generously donated samples for my studies, without which my PhD would not have been possible. I truly appreciate your selflessness and willingness to contribute our work. I would also like to thank Yonghong Zhang and Yan Zhao for being such generous hosts during my trips to Beijing, and for being such fantastic collaborators. Similarly, I would like to thank Ling, Guihai, and Hilary for all their help whilst in Beijing, both in the lab and in giving up your free time to show me around. My project absolutely revolved around this collaboration, and I could not have found better people to work with.

I would next like to acknowledge my supervisors, Tao Dong and Jan Rehwinkel. Tao, thank you for all your support over these four years. I am extremely fortunate to have had such an understanding and compassionate supervisor. Thank you for empowering and trusting me to make my own scientific choices, and helping me realise my ideas. Jan, I am endlessly appreciative of your helpful guidance; you have been indispensable as a teacher over these years.

My heartfelt thanks also go out to Shokouh Makvandi-Nejad. Thank you for being such a great mentor and friend to me when I was just starting in the lab, and for your continuously helpful advice. I would not have made it to this stage if it wasn't for you. To Dannielle Wellington and Lili Wang - thank you both for being the being the best teachers and friends I could ask for and for all the helpful advice and guidance. I would also like to acknowledge Mary Carrington and Alain Townsend for being approachable and insightful advisors throughout my PhD.

To Xi, Kerry, Ann, Liwei, Chi, Rhys, Megat, Peng, Xuan, Ushani, Shaza, Samara, Laura, Adi, and all other Dong lab members, thank you for making our a fun and welcoming environment. To Sophie Andrews, thank you for being my best PhD buddy and too many Toosdays! I also want to extend my

thanks to the brilliant WIMM facilities staff, especially Philip Hublitz, Kevin Clark, and Craig Waugh - you are the unsung heroes of the institute. Lastly, I want to thank Lorena Preciado-Llanes, Alice Mayer and Heidi Zinecker for their help with experiments throughout my PhD.

One of the most rewarding experiences of my DPhil was my internship at the WHO. To my supervisor Julia Fitzner - thank you so much for encouraging me to take on a challenging project, for your endless support, and for helping me get my coffee fix while I was an unpaid intern. To Kaat, Aspen, Bridget, Isabelle and the rest of the flu epidemiology team - thank you for giving me such a fun and welcoming working environment, and teaching me so much about epidemiology and public health. I also want to acknowledge all of the fantastic "future leaders in public health" I met during my time there; thank you for making my time in Geneva so enjoyable. While there are too many of you to list, I would like to especially thank Mirjam, Ajay, Bec, Fernando, Heather, Greg, Claudia, Emma, Priscilla, Christie and Katerina.

I also owe a great deal of thanks to friends in and out of Oxford for keeping me sane throughout my PhD. Firstly, thanks to Philip Staupe for being a constant and reliable source of friendship throughout my time here, and for *whisking* me up python scripts when needed. To Nameet Surana, Charmain Li, and Avtar Rekhi, thanks so much for putting up with me all these years. I have loved all the fun times we've had during my PhD, with the notable exception of the Butt Lake incident. A massive thanks also goes out to Emily, Ryan, Geoff, Janina, Lucy, Ed, Adrian, Hector, Ari, and everyone else for making these years such a fantastic time.

I would also like to acknowledge the MRC as the primary source of funding for my research. The MRC has supported me financially throughout my PhD, and has given me supplementary funding for conferences and training - I could not have had these experiences without them, and I am extremely grateful. I would also like to acknowledge the CTI for funding my trips to Beijing, and for supporting international collaboration in research.

Finally, I would like to thank my family for always supporting me in whatever I do. To my mum and dad, I cannot put into words how thankful I am to have such supportive and loving parents, who consistently empower me to choose my own path in life. To my sister Hannah, thank you for being such a kind and fun sibling and friend to me. To my grandparents, I am forever grateful for being such steadfast supporters of my PhD throughout, and for training me so well to give talks on my research. Finally, I would like to

thank my aunt Ann, for being a second mum to me, and for being such a helpful and loving person to me and my family throughout my life.

Statement of Originality

I declare that the work presented within this thesis entitled *"The Impact of Genetic Variation on IFITM3 Related Influenza Virus Restriction"* is my own. Contributions made by collaborators have been clearly acknowledged. I have not submitted this thesis in whole or in part for any other degree or qualification at The University of Oxford or elsewhere.

Henry Laurenson-Schafer
Trinity Term 2018

Authored Publications

- Makvandi-Nejad S, Laurenson-Schafer H, Wang L, Wellington D, Zhao Y, Jin B, Qin L, Kite K, Moghadam H, Song C, Clark K, Hublitz P, Townsend A, Wu H, McMichael A, Zhang Y, Dong T (2017) *"Lack of Truncated IFITM3 Transcripts in Cells Homozygous for the rs12252-C Variant Associated with Severe Influenza Infection."* The Journal of Infectious Diseases, 4 January 2018
- Hammond A, Laurenson-Schafer H, Marsland M, Besselaar T, Fitzner J, Vandemaele K, Zhang W (2018) *"Review of the 2017–2018 Influenza Season in the Northern Hemisphere"* Weekly Epidemiological Record, 24 August 2018

Grants and Awards

- WIMM Prize Studentship (Approximately £45,000 per anum, 4 years)
- WIMM Student Travel Award (2 awards, £1,500 total)
- Runner-up, Best Talk Prize, Influenza Update Meeting
- Best Student Poster Prize, Radcliffe Department of Medicine Annual Symposium (£100)
- MRC DTP Supplementary Funding for Exceptional Training Opportunities (£2,500)

Contents

1 Introduction **1**

1.1 Influenza Virus 1

1.1.1 The Influenza Virus Pathogen 1

1.1.2 Structure of the Influenza Virion 2

1.1.3 The Influenza Virus Genome 3

1.1.4 Influenza Virus Life Cycle 5

1.1.5 Influenza Virus Evolution 8

1.2 Influenza Virus Infection in Humans 9

1.2.1 Epidemiology of Influenza Virus 9

1.2.2 Determinants of Susceptibility to Influenza Virus Infection 11

1.3 The Immune Response to Influenza Infection 13

1.3.1 Innate and Intrinsic Immunity to Influenza Virus . . . 13

1.3.2 Adaptive Immunity to Influenza Virus 16

1.3.3 The Influenza Virus Vaccine 17

1.4 Innate Immune Signalling in Response to Influenza Infection . 18

1.4.1 Pattern Recognition of Influenza Virus 18

1.4.2 The Interferon Response 22
1.4.3 Interferon Stimulated Genes 29
1.4.4 ISG Restriction of Influenza Virus 30
1.4.5 Antagonism of Innate Immunity by Influenza Virus . . 32
1.5 *IFITM3* and other *IFITM* Genes 33
1.5.1 The *IFITM* Gene Family 33
1.5.2 The Antiviral Activity of IFITM Proteins 34
1.5.3 The *IFITM3* genetic locus 36
1.5.4 The IFITM3 Protein 41
1.5.5 IFITM3 Mechanism of Action 45
1.5.6 Evolution of *IFITM* Genes 47
1.6 IFITM3 Genetics and Infection 48
1.6.1 rs12225 Population Genetics 48
1.6.2 Association Between rs12252 and Viral Infection 52
1.6.3 The rs12252 Locus 55
1.6.4 Proposed rs12252 Mechanism of Action 57
1.6.5 Other Notable *IFITM3* Polymorphisms 58
1.7 Overall Aims 60

2 Methods and Materials **61**

2.1 General Methods and Materials 61
2.1.1 Extraction of gDNA from Cells 61
2.1.2 Amplification of *IFITM3* Exon 1 62

2.1.3 Purification of PCR Products 62
2.1.4 DNA Gel Electrophoresis 63
2.1.5 Gel Extraction of PCR Products 63
2.1.6 Genotyping of rs12252 63
2.1.7 TOPO-TA Cloning 63
2.1.8 Small Scale Plasmid Preparation (Miniprep) 65
2.1.9 Large Scale Plasmid Preparation (Maxiprep) 65
2.1.10 RNA extraction via RNeasy PLUS Mini Kit 65
2.1.11 Preparation of cDNA from RNA using High-Capacity RNA-to-cDNA Kit 65
2.1.12 Quantification of mRNA Transcripts via qPCR 66
2.1.13 Western Blotting For IFITM3 66
2.1.14 Culturing of HEK293 and A549 Cells 67
2.1.15 Collection and Counting of Cells in Culture 67
2.2 Methods and Materials for Chapter 4 68
2.2.1 Culturing Primary Human Fibroblasts 68
2.2.2 Preparation of cDNA via Superscript III 68
2.2.3 Targeted RNA Sequencing of rs12252 68
2.2.4 Analysis of rs12252 Allelic Expression 71
2.2.5 DNA Analysis via Bioanalyser 2100 System 72
2.2.6 Extraction of RNA via TRIzol 72
2.2.7 RNA Purification via RNeasy MinElute Cleanup Kit . 72
2.2.8 Nascent RNA Pulldown via 4SU Treatment 73

2.2.9 Nascent RNA pulldown via Click-iT Kit 74
2.2.10 Bisulfite Conversion of gDNA 74
2.2.11 Amplification of Bisulfite Treated rs12252 Locus 74
2.2.12 Detection of Methylation State of rs12252 74
2.2.13 Cloning sgRNA into px458 Plasmid Backbone 75
2.2.14 Transfection of HEK293 Cells 77
2.2.15 Transfection of A549 Cells 77
2.2.16 Single Cell Cloning of Transfected Cells 78
2.2.17 Single Cell Dilution of HEK293 Cells 78
2.2.18 Screening for *IFITM3* Knock-Out Cell Lines 80
2.2.19 Amplification of *IFITM3* Exon 2 80
2.2.20 Verification of *IFITM3* Knock-Out Status 81
2.2.21 Screening of rs12252 for Site Directed Mutagenesis . . 81
2.2.22 Revised Screening of rs12252 for Site Directed Mutagenesis 81
2.2.23 Amplification of *IFITM2* Exon 1 82
2.2.24 Verification of *IFITM2* Gene Integrity 83
2.2.25 Extraction of RNA from Whole Blood 83
2.3 Methods and Materials for Chapter 6 84
2.3.1 *In-Vitro* Infection of Adherent Cells with Virus 84
2.3.2 *In-Vitro* Infection of Suspended Cells with Virus 85
2.3.3 Separation of PBMCs from Whole Blood 85

2.3.4 Isolation of Monocytes and Generation of Monocyte-Derived DCs 85
2.3.5 Characterisation of DCs via FACS 86
2.3.6 Examination of S-Flu Infection via FACS 87
2.3.7 Flow Cytometry 87
2.3.8 Determination of IP10 Concentration using ELISA . . 88
2.3.9 Determination of Type-I IFN Concentration via 3C11 Bioassay 88
2.4 Buffers and Reagents 89
2.4.1 D10 Medium 89
2.4.2 R10 Medium 89
2.4.3 Viral Growth Medium 90
2.4.4 TAE Buffer 90
2.4.5 FACS Buffer 90
2.4.6 Freezing Medium 90
2.4.7 Biotinylation Buffer and Biotin Wash Buffer 91
2.4.8 S-Flu Virus 91
2.5 Bioinformatic Analyses 92
2.5.1 Quality Control with FastQC 92
2.5.2 Quality Control with cutadapt 92
2.5.3 Genome Alignment with STAR 94
2.5.4 Feature Counting with featureCounts 94
2.5.5 Production of Index Files for Alignment Visualisation . 94

2.5.6 Quantification of Isoform Level Expression Using Cufflinks . . . 94

2.5.7 Differential Gene Expression With DESeq2 and Accessory Packages . . . 95

2.5.8 Gene Ontology Enrichment with topGO . . . 95

2.5.9 eQTL Analysis with matrixEQTL . . . 96

2.5.10 Deconvolving IFN Signature from *IFITM3* Expression 96

3 Alternative Splicing of *IFITM3* in Relation to rs12252 98

3.1 Introduction . . . 98

3.1.1 Project Aims . . . 100

3.2 Results . . . 100

3.2.1 Examining the Splicing Profile of IFITM3 via western Blot . . . 100

3.2.2 Examining the Splicing Profile of *IFITM3* via Ribonucleic Acid (RNA)-Sequencing . . . 102

3.2.3 Analysis of *IFITM3* Splicing in Other Cohorts . . . 106

3.2.4 Examining the Splicing of *IFITM3* via the ExAC Database . . . 110

3.2.5 Analysis of SNPs in LD with rs12252 . . . 110

3.3 Discussion . . . 116

3.3.1 Full Length IFITM3 Protein is Expressed Regardless of rs12252 Genotype . . . 116

3.3.2 The Splicing Profile of *IFITM3* is Unaffected by rs12252 118

3.3.3 rs12252 is in LD with Three Proxy SNPs . . . 122

3.3.4 Concluding Remarks . . . 124

4 Expression of IFITM3 in Relation to rs12252 126

4.1 Introduction . . . 126

4.1.1 Project Aims . . . 128

4.2 Results . . . 128

4.2.1 Examination of *IFITM3* Expression Relative to rs12252 Using Primary Human Fibroblasts . . . 128

4.2.2 rs12252 eQTL Analysis Using Public Databases . . . 130

4.2.3 Further Investigation of rs12252 as an eQTL using Primary Human Fibroblasts . . . 131

4.2.4 Analysis of rs12252 as an eQTL using CRISPR-Cas9 . 141

4.2.5 Analysis of rs12252 In Healthy Donor Cohort . . . 148

4.3 Discussion . . . 176

4.3.1 *IFITM3* Expression is Modulated by rs12252 . . . 176

4.3.2 Mechanism of rs12252 Effect is Still Poorly Understood 182

4.3.3 ISG Signature is Related to rs12252 . . . 186

4.3.4 rs12252 is Associated with Local ncRNA Expression . . 190

4.3.5 Concluding Remarks . . . 194

5 Identification of Alternative eQTLs for *IFTIM3* 195

5.1 Introduction . . . 195

5.1.1 Project Aims . . . 196

5.2 Results . . . 197

5.2.1 eQTL Analysis for *IFITM3* . . . 197

5.2.2 Genomic Context of *IFITM3* eQTLs . . . 203

5.2.3 Population Genetics of *IFITM3* eQTLs 204

5.3 Discussion 206

5.3.1 *IFITM3* Expression is Associated With Two Novel SNPs 206

5.3.2 The rs28475163 SNP is Associated with Regulatory Features and Differential ncRNA Expression 209

5.3.3 Concluding Remarks 210

6 The Effects of rs12252 on Virus Infection 212

6.1 Introduction 212

6.1.1 Project Aims 214

6.2 Results 215

6.2.1 Examination of Blood Transcriptional Profile of Pandemic H1N1 Patients 215

6.2.2 Investigation of the Cytokine Profile of Infected DCs . 229

6.2.3 Infection of CRISPR-Edited HEK293 Cells 235

6.3 Discussion 238

6.3.1 RNA-Sequencing of Patients Reveals rs12252-C Associated IFN Signature 238

6.3.2 *In Vitro* Models of Infection Somewhat Recapitulate rs122252-C Associated Phenotypes 242

6.3.3 Concluding Remarks 243

7 Final Discussion and Future Directions 244

References 247

Appendix A Infection Cohorts used for Analysis 282

A.1 You'An Hospital Pandemic H1N1 Cohort 282
A.2 You'An Hospital Acute HIV-1 Cohort 283

Appendix B Primer and Guide RNA Sequences 285

Appendix C Biological Databases Used for Analysis 287

C.1 1000 Genomes Database 287
C.2 ExAC Database 289
C.3 GTEx Database 289
C.4 ensembl Database 290
C.4.1 Interferome Database 290

Appendix D Quality Control of RNA-Seq Cohorts 292

D.1 Quality Control of Primary Human Fibroblasts 292
D.1.1 Quality Control of PBMC and DC Datasets 292
D.2 Quality Control of Healthy Donor Cohort 295

Appendix E Additional Analysis of Healthy Donor Cohort 296

Appendix F Additional Analysis of 2009 H1N1 Cohort 301

List of Abbreviations

Acronyms

βME β-Mercaptoethanol

4SU 4-Thiouridine

aa Amino Acids

Ab Antibody

ActD Actinomycin D

AFR African

Ag Antigen

AMR American

ANO9 Anoctamin 9

APC Allophycocyanin

ATP Adenosine Triphosphate

B4GALNT4 Beta-1 4 N-acetylgalactosaminyltransferase

BGI Beijing Genomics Institute

BLAST Basic Local Alignment Search Tool

bp base pairs

BSA Bovine Serum Albumin

C25H Cholesterol 25-Hydroxylase

Cas9 CRISPR Associated Protein 9

CD14 Cluster of Differentiation 14

CD4+ Cluster of Differentiation 4 Positive

CD8+ Cluster of Differentiation 8 Positive

CD83 Cluster of Differentiation 83

CDC Center for Disease Control

cDNA Complementary DNA

CDX Chinese Dai in Xishuangbanna China

CEU Utah Residents with Northern and Western European Ancestry

cGAS Cyclic GMP-AMP Synthase

CHB Han Chinese in Beijing China

CHIP Chromatin Immunoprecipiation

CHS Southern Han Chinese China

CI Confidence Interval

CIL Conserved Intracellular Loop

cis-NAT Cis-Natural Antisense Transcript

CRISPR Cluster of Regularly Interspersed Short Palindromic Repeats

cRNA Complementary RNA

Ct Cycle Threshold

CTCF CCCTC-binding factor

CTD C-Terminal Domain

DAMP Danger Associated Molecular Pattern

DC Dendritic Cell

DC-SIGN Dendritic -Specific Intercellular Adhesion Molecule-3-Grabbing Non-Integrin

DMEM Dulbecco's Modified Eagle's Medium

DMF Dimethylformamide

DNase Deoxyribonuclease

DNase-I Deoxyribonuclease-I

DPBS Dulbecco's Phosphate Buffered Saline

dsDNA Double Stranded DNA

DTT Dithiothreitol

EAS East Asian

EBV Epstein-Barr Virus

E. coli *Escherichia coli*

EDTA Ethylenediaminetetraacetic Acid

eGFP Enhanced Green Fluorescent Protein

ELISA Enzyme-Linked Immunosorbent Assay

eQTL Expression Quantitative Trait Locus

EST Expressed Sequence Tag

EU 5-Ethynyl Uridine

EUR European

FACS Fluorescence Activated Cell Sorting

FCS Fetal Calf Serum

FIN Finnish in Finland

FITC Fluorescein Isothiocyanate

FN1 Fibronectin 1

FPKM Fragments Per Kilobase of transcript per Million mapped reads

GAPDH Glyceraldehyde 3-Phosphate Dehydrogenase

GAS Gamma Interferon Activation Site

GBR British in England and Scotland

gDNA Genomic DNA

GM-CSF Granulocyte-Macrophage Colony-Stimulating Factor

GO Gene Ontology

GWAS Genome-Wide Association Study

H3K27Ac Histone H3 Lysine 27 Acetylation

HA Haemagglutinin

HCV Hepatitis C Virus

HDR Homology Directed Repair

HEK293 Human Embryonic Kidney Cell 293

HIV Human Immunodeficency Virus

HRP Horseradish Peroxidase

IFITM1 Interferon Induced Transmembrane Protein 1

IFITM2 Interferon Induced Transmembrane Protein 3

IFITM3 Interferon Induced Transmembrane Protein 3

IAV Influenza A Virus

IBS Iberian populations in Spain

IC_{50} Half Maximum Inhibitory Concentration

IFI27 Interferon Alpha-Inducible Protein 27

IFN Interferon

IFN-α2 Interferon-α2

IFNAR Interferon α/β Receptor

IFN-β Interferon-β

IFN-γ Interferon-γ

IFNGR Interferon γ Receptor

IFN-λ Interferon-λ

IFNLR Interferon λ Receptor

Ig Immunoglobulin

IGV Integrated Genomics Viewer

IL-10 Interleukin-10

IL10Rβ Interleukin-10 Receptor Subunit β

IL-12 Interleukin-12

IL-18 Interleukin-18

IL-1α Interleukin-1α

IL-1β Interleukin-1β

IL-2 Interleukin-2

IL-4 Interleukin-4

IL-6 Interleukin-6

IMD Intramembrane Domain

indel Insertion or Deletion

IP10 Interferon γ-Induced Protein 10

IRF3 Interferon Regulatory Factor 3

IRF7 Interferon Regulatory Factor 7

IRF9 Interferon Regulatory Factor 9

IRG Interferon Regulated Gene

ISG Interferon Stimulated Gene

ISG15 Interferon Stimulated Gene 15

ISG20 Interferon Stimulated Gene 20

ISGF3 Interferon-Stimulated Gene Factor 3

ISRE Interferon-Sensitive Response Element

JAK1 Janus Kinase 1

JAK2 Janus Kinase 2

JPT Japanese in Tokyo Japan

kDa kilodaltons

KHV Kinh in Ho Chi Minh City Vietnam

KI Knock-In

KLH Keyhole Limpet Hemocyanin

KO Knock-Out

LB Lysogeny Broth

LCL Lymphoblastoid Cell Line

LD Linkage Disequilibrium

lincRNA Long Intergenic Non-Coding RNA

LPS Lipopolysaccharide

LRRC56 Leucine-rich repeat-containing protein 56

M1 Matrix-1

M2 Matrix-2

MAF Minor Allele Frequency

MAVS Mitochondrial Antiviral-Signaling Protein

MCP1 Monocyte Chemoattractant Protein 1

MDCK Madin-Darby Canine Kidney

MHC Major Histocompatibility Complex

MOI Multiplicity Of Infection

mRNA Messenger RNA

Mtb *Mycobacterium tuberculosis*

MX1 MX Dynamin Like GTPase 1

MyD88 Myeloid Differentiation Primary Response 88

NA Neuraminidase

NCBI National Center for Biotechnology Information

ncRNA Non-Coding RNA

NFκB Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B Cells

NK Natural Killer

NKT Natural Killer T

NLRP3 NOD- LRR- and pyrin domain-containing 3

NMD Nonsense-Mediated Decay

NP NucleoProtein

NS1 Non-Structural 1

NS2 Non-Structural 2

NTD N-Terminal Domain

OASL 2'-5'-Oligoadenylate Synthetase

oligo Oligonucleotide

OR Odds Ratio

PAM Protospacer Adjacent Motif

PAMP Pathogen Associated Molecular Pattern

PBMC Peripheral Blood Mononuclear Cell

PCA Principle Component Analysis

PCR Polymerase Chain Reaction

PE Phycoerythrin

PGGHG protein-glucosylgalactosylhydroxylysine glucosidase

PKP3 Plakophilin-3

PKR Protein Kinase R

PNK Polynucleotide 5'-Hydroxyl-Kinase

PR8 Influenza A/Puerto Rico 8/1934

PRAG1 PEAK1 Related Kinase-Activating Pseudokinase 1

PRR Pattern Recognition Receptor

PTC Premature Termination Codon

qPCR Quantitative PCR

RASSF7 Ras association domain family member 7

RIG-I Retinoic Acid-Inducible Gene I

RNA Ribonucleic Acid

RNA18S 18S rRNA

RNAP-II RNA Polymerase II

RNase-L Ribonuclease-L

RPMI Roswell Park Memorial Institute Medium

RT Room Temperature

SARI Severe Acute Respiratory Infection

SAS South Asian

SERPINE1 Serpin Family E Member 1

SeV Sendai Virus

SFTPA2 Surfactant Protein A2

sgRNA Single Guide RNA

SIGIRR Single Ig IL-1-related receptor

siRNA Small Interfering RNA

SIRT3 Sirtuin 3

SIV Simian Immunodeficency Virus

SNP Single Nucleotide Polymorphism

Sp1 Specificity Protein 1

ssDNA Single Stranded DNA

ssODN Single Stranded Donor Oligonucleotide

ssRNA Single Stranded RNA

STAT1 Signal Transducer and Activator of Transcription 1

STAT2 Signal Transducer and Activator of Transcription 2

STAT3 Signal Transducer and Activator of Transcription 3

TAE Tris-Acetate-EDTA

TB Tuberculosis

TBST Tris-Buffered Saline with Tween 20

TE Tris EDTA Buffer

TEF-1 Transcription Enhancer Factor 1

TF Transcription Factor

TLR3 Toll Like Receptor 3

TLR7 Toll Like Receptor 7

TMPRSS2 Transmembrane Protease Serine 2

TNFα Tumour Necrosis Factor α

TRIF TIR-Domain-Containing Adapter-Inducing Interferon-β

TRIM22 Tripartite Motif Containing 22

TRIM25 Tripartite Motif Containing 25

TRIM56 Tripartite Motif Containing 56

TSI Toscani in Italy

TSL Transcript Support Level

TKY2 Tyrosine Kinase 2

USF1 Upstream Stimulatory Factor 1

UTR Untranslated Region

VAPA Vesicle-Membrane-Protein Associated Protein A

v-ATPase Vacuolar ATPase

vRNP viral Ribonucleoprotein

dH_2O Distilled Water

WHO World Health Organisation

WT Wild-Type

ZAP Zinc Finger Antiviral Protein

ZMPSTE24 Zinc Metallopeptidase STE24

Chapter 1

Introduction

1.1 Influenza Virus

1.1.1 The Influenza Virus Pathogen

Influenza viruses are members of the *Orthomyxoviridae* family of viruses - these are a family of enveloped, negative-sense RNA viruses. Influenza viruses are most well known for infecting humans, other mammals, and birds. However, there is evidence that these viruses share evolutionary relationships to fish and amphibian viruses and even, more distally, arthropod viruses [1, 2].

Influenza viruses are of particular concern due to their heavy disease burden around the world. In humans, influenza infection can result in disease ranging from asymptomatic infection to death. In most cases, influenza morbidity and mortality is associated with respiratory symptoms, including coughing, fever, headache and malaise [3]. Incubation time for influenza viruses ranges from 1 to 4 days; shedding can occur roughly one day before the onset of symptoms. The illness itself generally resolves in a week or less after onset of symptoms. There is also evidence that influenza virus-related deaths can be associated with cardiovascular symptoms as well as respiratory ones [4].

In severe instances of influenza virus infection, patients can be exposed to coinfection by other bacteria or viruses. It is difficult to define severe influenza virus infection, due to the heterogeneity of more severe symptoms

- studies often use a positive test for influenza virus infection alongside admission to hospital or intensive care units as case definitions [5]. On a molecular level, severe influenza is characterised by a prolonged state of inflammation in the lung [6].

Influenza virus can spread relatively easily in the general population - its modes of transmission include:

- Aerosol droplets, generally generated through coughing or sneezing [7].
- Fomites, or direct contact with objects contaminated by the virus [8].
- Direct contact with an infected individual.

Human influenza viruses are split into three types, known as A, B and C. Virus types A and B are sister groups, while type C is more extant [9]. The majority of human infections are caused by types A and B, meaning the bulk of research is focused on these viruses. Influenza A Virus (IAV) is further classified into subtypes, with names based on the antigenic profile of the virus. Influenza B viruses are grouped into two main subtypes, known as Influenza B (Victoria) and Influenza B(Yamagata).

1.1.2 Structure of the Influenza Virion

Influenza virions are enveloped, negative-sense, single-stranded RNA particles. These can be long and filamentous, or small and round [10]. When in spherical form, these virions are roughly 100nm in diameter [11]. An individual virion has two primary surface markers known as Haemagglutinin (HA) and Neuraminidase (NA) (fig. 1.1). The HA protein is found in greater abundance, with a HA to NA ratio of roughly four to one. HA proteins have two subunits, known as HA1 and HA2. On the surface, HA exists as a homotrimer, with three HA proteins forming a full surface Antigen (Ag). NA proteins form homotetramers. The HA and NA genes are key antigenic sites in the Antibody (Ab) response to influenza virus. Matrix-2 (M2) ion channels can also be found at the surface of the virus - these are the least abundant of the three surface proteins.

Beneath the viral envelope, the Matrix-1 (M1) protein forms a protein coat, encasing the viral genome. This protein simultaneously binds the viral RNA

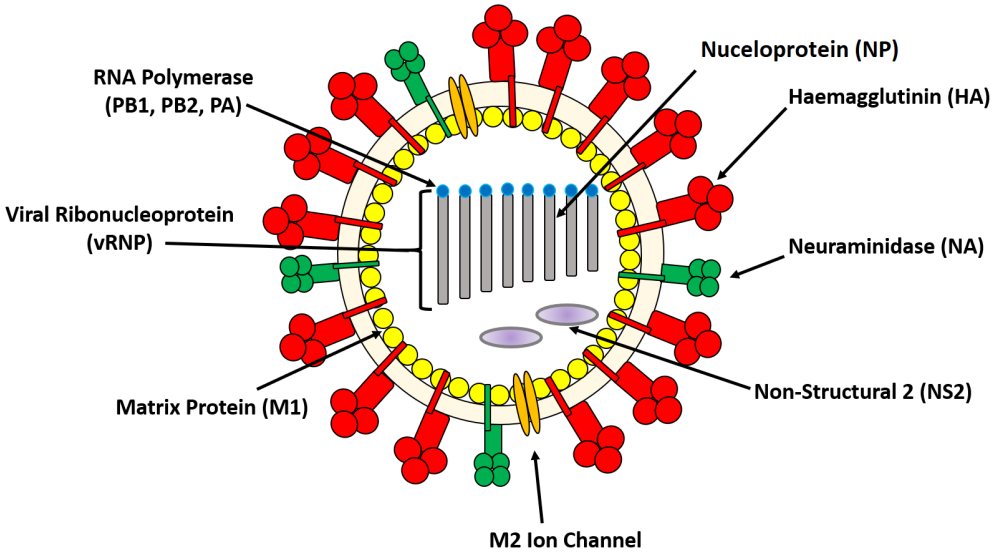


Figure 1.1: Cartoon showing cross section through a spherical influenza (A or B) virion, showing structural features. Figure adapted from Encyclopaedia of Virology, 3rd Edition [12]

and membrane [13]. Inside the virion, the genome is packaged into a viral Ribonucleoprotein (vRNP), composed of a polymerase and a NucleoProtein (NP) coated strand of genomic RNA. The viral polymerase is composed of three protein subunits, known as PB1, PB2, and PA. Viral RNA is coated in the RNA-binding viral protein, NP. In Influenza A and B viruses, the genome is divided into eight different RNA segments; in influenza C viruses, there are only seven segments. The viral protein Non-Structural 2 (NS2) is found inside influenza virions.

1.1.3 The Influenza Virus Genome

As mentioned in section 1.1.2, the genome of Influenza viruses A and B is split into eight RNA segments, contained in the virion as vRNPs. This complex has roles in replication, production of viral mRNA, and intracellular trafficking throughout the cell [14]. Given that the influenza genome is negative-sense RNA, delivery of a polymerase molecule alongside the viral genome is necessary for these viruses, to create positive-sense viral RNA for the purpose of translation.

Each viral RNA molecule contains one or more viral genes, alongside

segment specific packaging signals. These packaging signals are found within conserved non-coding regions of the genome, at both the 5' and 3' ends of each segment. Each sequence is specific to its segment, meaning one copy of the genome segment is incorporated into each particle [15]. Complementary regions on each termini bind together, causing each segment to form a hairpin structure, which acts as a promoter for RNA replication and transcription [11].

The segments of the influenza A viral genome are described below, in order from largest to smallest, with core genes shown:

1. **PB2** - The PB2 gene found in the first segment makes up a component of the viral RNA polymerase, and is involved in recognition of 5' mRNA caps
2. **PB1 & PB1-F2** - The second segment encodes PB1, an additional component of the viral polymerase involved in elongation. In some virus strains, the PB1-F2 gene is present - this is a pro-apoptotic gene, and its presence is considered a possible determinant of virus pathogenicity [16].
3. **PA** - The third segment encodes PA, a component of the viral RNA polymerase.
4. **HA** - The fourth segment encodes HA, the fusion protein and major surface Ag for influenza.
5. **NP** - The fifth segment encodes NP, involved in genome trafficking and replication of the virus.
6. **NA** - The sixth segment encodes NA, the other major surface Ag for influenza. NA releases influenza virus from the cell surface during budding
7. **M1 & M2** - The seventh segment encodes M1, which creates the viral structural coat. Also encoded is M2, the H+ ion channel present on the surface of influenza virions.
8. **NS1 & NS2** - The eighth segment encodes Non-Structural 1 (NS1), a viral protein with a wide range of functions related to suppression of the innate immune response of the host cell. NS2 is also encoded - this is involved in export of vRNPs from the cell nucleus. NS1 function is reviewed in greater detail in section 1.4.5

1.1.4 Influenza Virus Life Cycle

In humans, influenza replication is largely limited to the cells in the respiratory tract. Viral tropism is defined by proteases native to the respiratory tract that specifically cleave HA, ensuring that the protein is in an active form [17]. The full replication cycle of an influenza virus infecting an airway epithelial cell is explained below and in fig. 1.2.

Viral Entry

Entry of influenza virus is mediated by binding of HA to host sialic acid at the terminus of host glycans. Human viruses bind sialic acid attached to galactose by an α-2,3 linkage, the predominant form of this bond. Conversely, avian influenza viruses bind to sialic acid attached to galactose via a α-2,6 linkage [18]. α-2,6 linkages are only present deeper within the human lung, making the difference in receptor binding activity a major barrier to host-switching for avian influenza viruses [19].

Upon binding between HA and sialic acid, influenza virus is endocytosed by the host cell. While it is well established that clathrin-mediated endocytosis is the primary means by which influenza enters the cell, other host endocytosis pathways, such as macropinocytosis, have been implicated in entry [20, 21]. There are a range of additional host factors important for triggering and facilitating influenza virus entry [22]. For example, a recent study has highlighted the potential role of C-type lectins in enhancing influenza virus entry [23].

Upon endocytosis, influenza virions are trafficked through the host cell endosomal pathway. Regulation of the eukaryotic endocytic network is highly complex and poorly understood. With respect to influenza, it is thought that particles are trafficked from early to late endosomes before eventual fusion. This trafficking was demonstrated to comprise of three distinct steps [24]. The virion is transported by actin at the periphery, followed by rapid dynein-mediated transport. Finally, microtubules traffic the virion towards the nucleus, where the endosome will mature from an early to late endosome.

The key event in triggering influenza virus entry into the cell is low pH within the endosome containing the influenza virus. Reduction in endosomal pH is a natural aspect of endosomal maturation and is driven by Vacular

ATPases (v-ATPases). Low pH levels trigger a conformational change in the HA protein, causing a hydrophobic fusion peptide to become exposed to the endosomal membrane. Insertion of these fusion peptides into the endosomal membrane causes viral and endosomal membrane leaflets to be brought into close proximity. This results in hemifusion of the membranes, before full fusion and creation of a fusion pore [25].

Upon fusion, influenza virions disassemble, releasing vRNPs into the cytosol. This is mediated by acidification of virus interior via the M2 ion channel. As a result of the reduced pH, M1 protein interactions with vRNPs are destabilised, allowing vRNPs to enter the cell [26]. Upon release into the cytosol, vRNPs are transported to the nucleus, as mediated by a nuclear localisation signal on the NP protein [27]. Nuclear import itself occurs via the karyopherin pathway [28].

Genome Replication and Protein Synthesis

In the nucleus, influenza virus vRNPs undergo active replication and transcription. The negative sense RNA genome is used to generate both Messenger RNA (mRNA), used to make viral proteins, and Complementary RNA (cRNA), used as an intermediate in the synthesis of novel viral genomes. It is likely that these RNA species are both created simultaneously, but the cRNA requires newly synthesised polymerase and NP to stabilise it [29]. In synthesising mRNA, the viral polymerase cleaves the 5' cap from host mRNA - this process disguises viral mRNA as host mRNA and is known as "cap-snatching" [30].

All viral proteins are synthesised from viral mRNA produced inside the host cell. Upon synthesis, HA, NA, and M2 are sent to the cell surface via the golgi. vRNPs are assembled in the cell nucleus. The NS2 protein mediates the export of vRNP complexes to the cytosol [31]. It has been demonstrated that an interaction between M1 and NS2 is required for nuclear export to take place [32].

Viral Assembly and Budding

Viral progeny are produced via budding at the cell surface. This takes place at lipid rafts, where HA and NA proteins accumulate [33]. Assembly of the

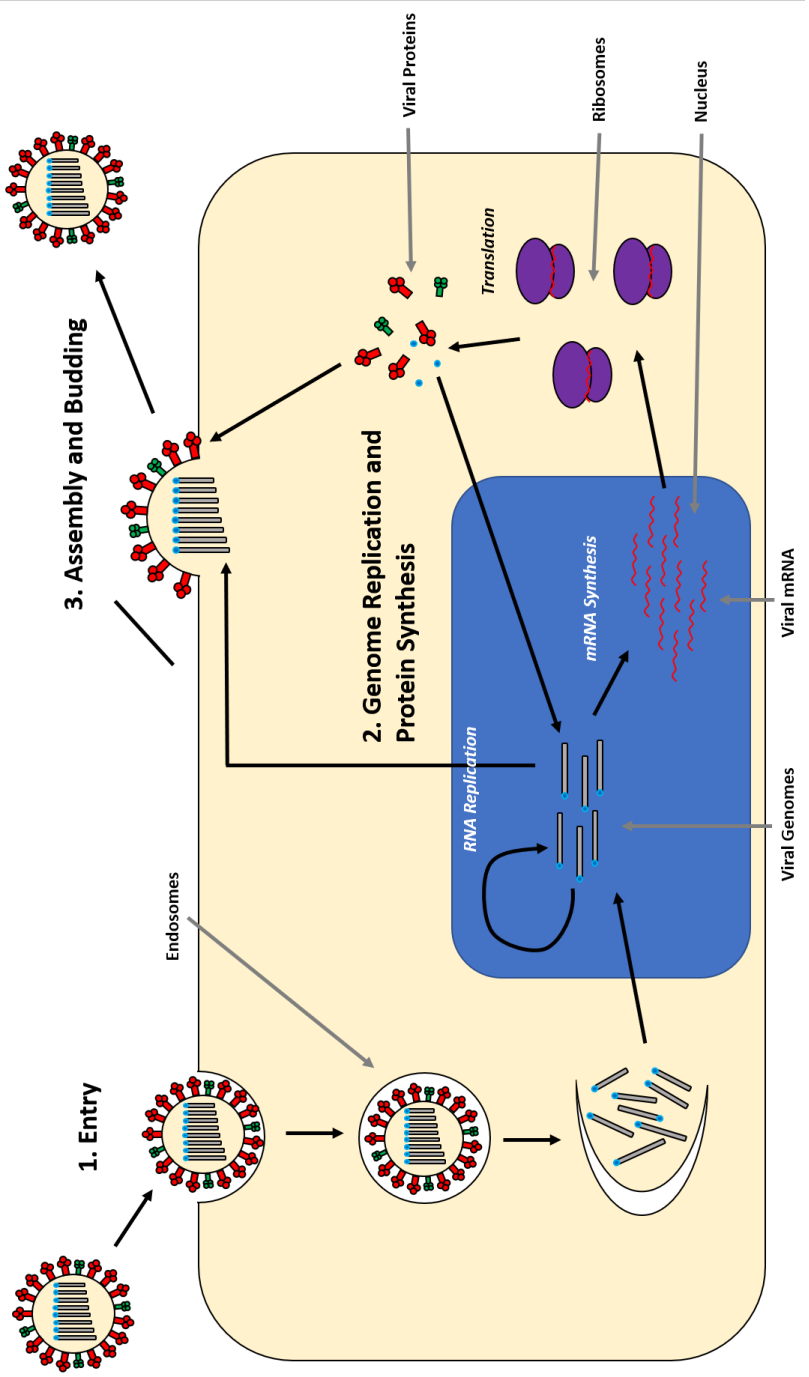


Figure 1.2: Cartoon representing the life cycle of an influenza virus infecting a host cell (simplified). 1. An influenza virion enters the host cell via the endosomal pathway, before fusion and uncoating. 2. vRNPs are trafficked to the cell nucleus, where they are used to produce mRNA and replicate to form novel virus genomes. 3. Progeny influenza virions are assembled, and bud at the cell surface

virus is thought to be mediated by accumulation of the M1 protein. M1 protein aggregates at the cell surface, interacting with HA, NA, vRNPs, and the membrane itself. Self-polymerisation of M1 at the cell surface may also be able to introduce the membrane curvature that results in budding [26]. Membrane scission is mediated by the M2 protein, independently of host cell factors [34]. As HA molecules bind to sialic acid residues, NA acts as a sialidase, cleaving terminal sialic acids, and releasing the virus particles from the cell.

1.1.5 Influenza Virus Evolution

In both the pandemic and seasonal context, virus evolution underpins the constant threat of infection with respect to influenza virus. Due to the changing nature of the virus, there is no known way by which consistent immunity to influenza virus can be established - adults are assumed to have encountered influenza virus multiple times in their life.

There are two main mechanisms by which influenza virus evolution occurs, known as drift and shift. Drift refers to accumulation of naturally occurring mutations [35]. This is underpinned by the error-prone viral RNA polymerase, and the large quantity of progeny virions produced from a single infected cell. Interestingly, recent evidence has suggested that only single, specific mutations are sufficient to change the antigenic profile of influenza virus and generate escape mutants [36, 37].

Evolutionary shift derives from the fact that influenza virions contain segmented genomes. When coninfection between two virus particles in a single cell occurs, progeny virus can incorporate genome segments derived from both parent viruses. This form of evolution is particularly notable with regards to pandemic influenza viruses - completely novel strains of virus are generated in this fashion. For example, the virus causing the 2009 H1N1 pandemic was thought to derive from influenza virus reassortment occurring in swine [38].

1.2 Influenza Virus Infection in Humans

1.2.1 Epidemiology of Influenza Virus

Worldwide, influenza virus is one of the largest contributors to respiratory infection [9]. Given the significant effects of influenza on healthcare systems, the World Health Organisation (WHO) and individual countries, have established influenza surveillance systems to monitor viral spread, and anticipate effects of spreading influenza virus [39]. Influenza is also of concern because of its potential to cause widespread pandemics.

Seasonal Influenza Virus

Most human influenza virus infections are attributable to regularly circulating influenza virus, where infection generally occurs in seasonal patterns. Seasonal influenza spreads relatively easily; the basic reproduction number of seasonal influenza is approximately 1.28 [40]. In temperate countries with regular seasonality, circulating influenza virus primarily circulates during winter months, causing a seasonal epidemic. Seasonal influenza becomes regionally extinct in temperate zones during summer months of the year. The global source of human influenza virus is thought to be East and South East Asia, where endemic levels of seasonal influenza subsist over the course of the year [41].

Currently, there are four influenza virus types circulating seasonally. These are Influenza A(H3N2), Influenza A (H1N1)pdm2009, and two Influenza B strains known as B(Yamagata) and B(Victoria). While these are all found within a single season, different strains will show dominance in different transmission zones.

On a yearly basis, it is estimated that seasonal influenza virus infection results in three to five million cases of hospitalisation, and 290,000 to 600,000 deaths [42, 43]. Rates will vary year-to-year depending on seasonal severity. It is important to recognise that death rate does not account for the effects of community influenza infection. In developed countries, influenza related deaths are mostly limited to elderly people [4]. Additionally, young children, pregnant women, and immunosuppressed individuals are at higher risk of severe infection and death. A 2018 Cassini *et al* study found that, for

all communicable diseases in Europe, influenza incurred the greatest loss of disability-adjusted life years in individuals greater than 65 years of age, and was a significant factor in those younger [44].

There are several counter-measures that can be used to control influenza virus. The first and most effective of these is vaccination, as discussed in section 1.3.3. Active influenza virus infection can also be treated with antiviral drugs such as oseltamivir [45].

Pandemic Influenza Virus

Perhaps the greatest danger associated with influenza virus is its pandemic potential. Pandemics are caused by emergent strains of influenza virus, separate from those circulating before the onset of the pandemic. These are often the result of viral reassortment in non-human hosts.

Since the advent of the 20th century, there have been five influenza pandemics of variable severity occurring in 1918, 1957, 1968, 1977, and 2009. The pandemic in 1977 was mild compared to the others mentioned, and is often not considered a pandemic [46]. The most severe influenza pandemic on record was the 1918 pandemic, known colloquially as Spanish flu. The virus spread in three epidemic waves across the world between 1918 and 1919 [47], culminating in the deaths of roughly 50 million people [48]. Epidemiological modelling has estimated that the same pandemic occurring in 2018 would result in 32 million deaths [49]. There is evidence that the virus that caused this pandemic, a subtype of Influenza A(H1N1) was entirely avian in origin [50], and was novel to humans. To date, reassorted genetic elements from the 1918 virus exist in all circulating strains of influenza A virus in humans [51].

The most recent influenza virus pandemic occurred between 2009 and 2010. This pandemic was mild in comparison to the 1918 pandemic causing roughly 300,000 deaths [52]. This virus was of an Influenza A(H1N1) subtype, and originated in Mexico and the United States. Genome segments in this virus came from a combination of swine, avian and human viruses [53]. This virus incurred higher risk of hospitalisation in young people than seasonal influenza, alongside a higher risk of admission to intensive care units [54].

Zoonotic Influenza Outbreaks

In addition to causing epidemics and pandemics, small scale outbreaks of zoonotic influenza infection occur in humans. Outbreaks of avian influenza are usually associated with close human to avian contact at live poultry markets, most of which are found in China [55]. In the United States, agricultural fairs are also a source of swine to human influenza transmission [56]. The most common virus strains associated with human cases of avian influenza are Influenza A(H5N1) and Influenza A(H7N9). Human infections are associated with high mortality rates, but do not transmit well between humans, making their spread self-limiting [57]. However, there is concern that these viruses could develop faster rates of transmission in humans, and become a source of a novel influenza pandemic.

1.2.2 Determinants of Susceptibility to Influenza Virus Infection

In the general population, there are some well understood risk factors for developing severe influenza virus infection. In a 2013 meta-analysis by Mertz *et al*, 234 articles were used to evaluate evidence for a number of risk factors [58]. This study pointed to age as a risk factor for both seasonal and pandemic influenza. For both, elderly people had the highest risk of hospital admission and death. For seasonal influenza, young children were at higher risk of developing pneumonia. For pandemic influenza, pregnancy was also identified as a risk factor. In addition to these findings, a number of medical conditions were associated with either death or hospital admission. These include:

- Lung Disease
- Asthma
- Cardiovascular Disease
- Diabetes
- Immunosuppression
- Obesity

In a South African study conducted by Tempia *et al* in 2017, the above results were largely replicated, but an additional set of risk factors for influenza-associated Severe Acute Respiratory Infection (SARI) was identified [59]. These included prematurity, active Human Immunodeficency Virus (HIV) infection, and malnutrition. These results were applicable to seasonal influenza only.

In modern times, the role of human genetics in disease is becoming more apparent, with a number of human genetic risk factors for severe influenza identified. The WHO has suggested that understanding genetic links to severe and pandemic influenza should be a research priority [60]. Currently, polymorphisms, mainly Single Nucleotide Polymorphisms (SNPs), in a number of genes related to the host cell response have been implicated in risk for severe influenza.

One example of genetic susceptibility to severe influenza virus infection comes from a 2015 study, where whole-exome sequencing identified a patient with homozygous null mutations in the *Interferon Regulatory Factor 7 (IRF7)* gene (discussed in section 1.4.2) [61]. The ability to induce an antiviral response in this patient was severely impaired. Similarly, polymorphisms impairing the function of the *Toll Like Receptor 3 (TLR3)* gene (discussed in section 1.4.1), have been associated with severe outcomes to infection in the 2009 influenza pandemic [62]. There is some evidence that SNPs found in the *Interleukin-10 (IL-10)*, Interleukin-1α (IL-1α), Interleukin-1β (IL-1β) genes, which encode cytokines important to the antiviral response, are associated with severe influenza virus infection [63, 64]. The preceding examples highlight how disruption of antiviral immunological responses can incur susceptibility to influenza virus infection

SNPs affecting genes with direct influenza interactions have also been identified as risk variants for influenza virus infection. For example, SNPs in the antimicrobial genes *Surfactant Protein A2 (SFTPA2)* and *Galectin-1* have been implicated in risk for severe influenza [65, 66]. These proteins are found in the airways, and can directly interact with influenza virus, limiting its replication. Interestingly, a SNP affecting the expression of *Transmembrane Protease Serine 2 (TMPRSS2)*, a host protease involved in HA cleavage, has also been associated with severe influenza infection [67].

Currently, the most prolific set of genetic variants related to severe influenza are associated with the *Interferon Induced Transmembrane Protein 3 (IFITM3)* gene. This gene, and associated polymorphisms are discussed

further in section 1.5 and section 1.6.

Despite these findings, only one Genome-Wide Association Study (GWAS) investigating influenza severity has been performed to date. This study used a small cohort, with a total of 156 patients stratified into mild and severely infected individuals [68]. This study was performed on a Spanish population. In the Everitt *et al* study, a single SNP, rs28454025, was identified as a risk variant - this SNP is located inside an intron of the *PEAK1 Related Kinase-Activating Pseudokinase 1 (PRAG1)* gene. Given this gene's role in neural development, the authors concluded that this was a false positive. Despite a lack of results, this GWAS does not effectively rule out the idea that host genetics are partial determinants of influenza severity given its small sample size and focused population.

1.3 The Immune Response to Influenza Infection

The immune response to influenza infection can be divided into the innate and adaptive immune responses. The former is a more generalised term, which refers to early-acting, relatively non-specific immunity. The latter refers to the virus specific, anti-influenza immunity mediated by B and T cells. In most cases, the innate immune response curtails influenza virus replication and pathogenesis, until it can be eliminated by the adaptive immune response (fig. 1.3).

1.3.1 Innate and Intrinsic Immunity to Influenza Virus

The forefront of the innate immune response to influenza virus is the intrinsic immune response. This primarily takes the form of the mucosal layer in the upper respiratory tract. Influenza virions can become trapped in this layer. In addition, human mucus contains decoy sialic acid receptors, which can occupy binding domains of influenza HA [70].

When influenza virus productively infects host cells, early innate immune signalling is triggered inside infected cells. This signalling results in the release of type-I and III interferons and proinflammatory cyokines (e.g. Interleukin-6 (IL-6), Tumour Necrosis Factor α (TNFα), and IL-1β) into

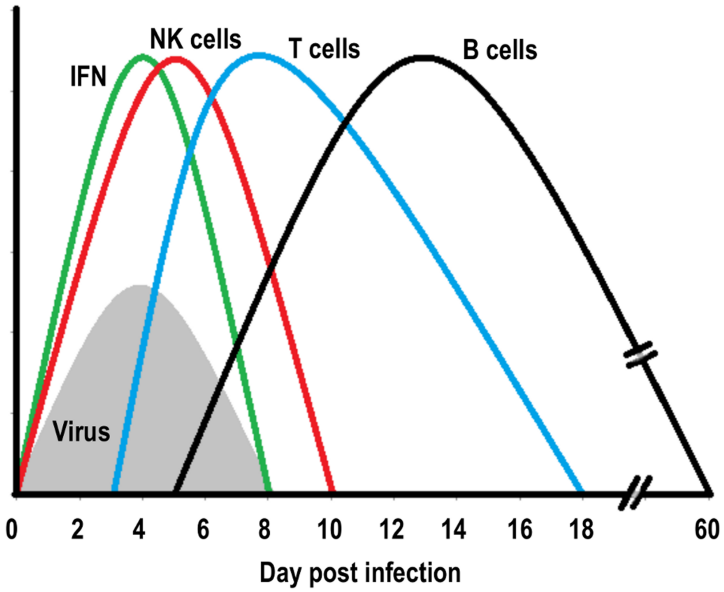


Figure 1.3: Schematic diagram showing the immune response to influenza infection. Different aspects of the immune response are invoked under different time frames in the typical response to influenza. Virus replication peaks at roughly four days. The innate immune response occurs early in the course of infection, as exemplified in the figure by IFN (Interferon) and NK (Natural Killer) cells. The adaptive immune response occurs later in the course of infection, as shown via the T and B cell curves. Taken from [69].

the surrounding environment [71]. In addition, this will result in infected cell death, often via pyroptosis [72]. Later recruitment of immune cells will result in release of a number of additional cytokines, including type-II IFNs. These cytokines perform a wide range of functions; interferons (reviewed in section 1.4.2) tend to invoke direct antiviral defenses within cells, whereas proinflammatory cytokines will induce a change in the immune environment which facilitates pathogen clearance and immune cell recruitment.

The innate immune response in the respiratory epithelial cells of the lung results in the recruitment of a wide range of innate immune effector cells. These cells include monocytes, lung resident macrophages, neutrophils and Dendritic Cells (DCs):

- **Neutrophils** are recruited early to the site of infection, and mainly phagocytose dead or infected cells. Phagocytosis of viral material often results in neutrophil cell death and release of inflammatory cytokines.

On a molecular level, severe influenza virus infection is characterised by high numbers of neutrophils at early stages of infection, accompanied by high levels of inflammation in the infected tissue [73].

- **Macrophages** in influenza virus infection are either resident in the tissue (alveolar macrophages), or derived from recruited circulating monocytes (monocyte-derived macrophages) [74]. Like neutrophils, macrophages function to phagocytose dead or infected epithelial cells, curtailing the replication of influenza virus. In addition to this, alveolar macrophages are heavy producers of type-I Interferon (IFN), which leads to increased antiviral activity in the tissue environment [75].

- **Monocytes** migrate to the site of infection from the circulatory system, where they infiltrate the lung and nasal mucosa. In the infected tissue, monocytes can differentiate into macrophages (as mentioned previously), or monocyte-derived DCs [76]. Some studies have reported roles for monocytes in the production of inflammatory cytokines, namely TNFα. However, mouse studies where mice depleted of monocytes were infected with influenza virus did not find that clearance of virus was augmented by monocyte recruitment [77].

- **Dendritic Cells** also accumulate in the lung following influenza virus infection. DCs are seen as critical mediators of the adaptive immune response, due to their roles in surveillance of the local environment, and their interface with adaptive immune cells [78]. Similarly to macrophages, DCs involved in the anti-influenza immune response are divided into two subsets, the plasmacytoid DCs, and myeloid-lineage DCs. The former are generally tissue resident; they make up a small fraction of Peripheral Blood Mononuclear Cells (PBMCs) [79]. Like macrophages, plasmacytoid DCs are heavy producers of type-I IFN, especially in the context of infection with highly pathogenic strains of influenza virus [80]. DCs are most well known for their role in Ag presentation; DCs acquire foreign Ag either by direct infection with influenza virus, or via macropinocytosis. Foreign Ag can be presented on Major Histocompatibility Complex (MHC) class I and II molecules. The former occurs via direct presentation (via infection) or cross presentation, while the latter occurs via processing of macropinocytosis. Ag presented on MHC class I molecules are presented to Cluster of Differentiation 8 Positive (CD8+) T cells, while Ag presented on MHC class II molecules are presented to Cluster of Differentiation 4 Positive (CD4+) T cells. This Ag presentation forms the critical bridge between the innate and adaptive responses to influenza virus infection [81].

- **Natural Killer (NK) Cells** migrate from the periphery to the site of influenza infection. These are specialised cells that recognise influenza infected cells via expression of natural cytotoxicity receptors. These receptors function to recognise various virus-related biomarkers [82]. For example, HA on the surface of infected cells can be recognised by one these receptors [83]. In addition, NK cells can recognise a number of stress-associated ligands upregulated on cells infected by influenza. An example of this is the NKG2D-MICA interaction. Upon activation, NK cells contribute to the antiviral response via release of type-II IFN, and direct lysis of infected cells.

1.3.2 Adaptive Immunity to Influenza Virus

As mentioned previously, the adaptive immune response to influenza virus infection is delayed relative to the innate immune response, with specific activity mediated by T and B cells. Adaptive immunity is based on specific recognition of viral Ag by highly variable T and B cell receptors. The role of these cells will be discussed only briefly, as it is not fully relevant to this thesis.

T and B cells reside in the lymphatic system before active infection, usually within lymph nodes. During the course of infection, DCs and other Ag presenting cells infected with influenza virus will migrate to lymph nodes, presenting foreign peptide on the surface of MHC molecules [84]. Specific recognition of this peptide via T cell receptors leads to T cell activation, causing clonal expansion and migration to the site of infection. T cells are divided into CD8+ and CD4+ T cells; the former directly lyse infected cells upon recognition of viral peptide presented on the cell surface via MHC molecules [85]. CD4+ T cells perform a wide range of effector functions, including cytokine production at the site of infection and B cell activation in lymphoid tissues [86].

B cells, are activated by a combination of free Ag and co-stimulation from CD4+ T cells. These cells produce virus-specific Abs, which neutralise virus infection via a range of mechanisms. These include direct neutralisation of virus particles, and activation of the complement system. Anti-influenza Abs are actively improved by B cells via a process known as affinity maturation. The most important antibody classes involved in clearance of influenza virus infection are Immunoglobulin (Ig) A and IgG [87].

The adaptive immune response, in the majority of cases, will fully clear influenza virus infection. Some T and B cells will differentiate into long lived memory cells, which will become reactivated in a repeat infection. In addition, long lived plasma B cells will produce low serum levels of anti-influenza Ab at steady-state. This memory response can be avoided by influenza viruses via Ag shift and drift as discussed in section 1.1.5.

1.3.3 The Influenza Virus Vaccine

Influenza virus infection can be controlled, to some extent, via vaccination. This is the most cost effective way to prevent influenza virus infection. Given the constant change in the antigenic profile of circulating influenza viruses, vaccines are redesigned and manufactured before each influenza season. This causes vaccine efficacy to vary season-to-season. Vaccines are either trivalent or quadrivalent; the latter contains Influenza A(H3N2), A(H1N1)pdm2009, B(Victoria), and B(Yamagata) strains, while the trivalent vaccine only contains one Influenza B strain. The majority of influenza vaccines used today are prepared using inactivated virus.

Vaccine efficacy varies between strains; vaccines against Influenza A(H3N2) generally have the worst efficacy. This is related to the fact that Influenza A(H3N2) has a faster rate of evolution than other circulating influenza strains [88]. On top of this, H3N2 virus adaptation to egg-based growth, where the majority of vaccine production occurs, incurs altered antigenicity [89]. While cell culture based virus production is possible, current infrastructure dictates that seasonal vaccines will be produced in eggs for years to come [90]. Despite these limitations, vaccination remains the most effective method to curtail seasonal influenza infection; the United States Center for Disease Control (CDC) estimates for national vaccine efficacy range from 10 to 60% effectiveness [91].

The current seasonal influenza vaccine is generally only capable of generating humoral immunity. While influenza vaccination is effective at eliciting these responses, they are generally only directed against the HA protein [92]. While some of the Abs induced by this response will cross react between strains of influenza virus, the fact that this memory response is directed against surface Ags contributes to low vaccine efficacy and potential for immune escape.

In response to the limitations of B cell mediated immunity, a number of groups have suggested that a CD8+ T cell response to influenza, alongside

this B cell response would result in more effective protection. T cells can be directed against internal influenza virus Ags which are more evolutionarily conserved. This has been observed in a number of studies, including a 2008 Lee *et al* study, where memory T cells established against seasonal influenza were seen to cross-react with influenza A (H5N1) [93]. It is generally accepted that induction of T cell response would require targeting DCs for activation [94].

In recent years, a live attenuated version of the influenza vaccine has been used in the USA and Europe [95]. This quadrivalent vaccine was shown to induce T and B cell immunity in clinical trials [96]. However, recent vaccine efficacy results have been mixed with this vaccine. While it has been shown to be effective in European trials, this is not the case for the USA [97].

In addition to the live attenuated vaccine, a novel influenza vaccine was developed in 2011, with the goal of universal protection between influenza strains. This vaccine is based on a recombinant modified vaccinia Ankara virus encoding the NP and M1 proteins of IAV [98]. This vaccine has shown promising results in early studies, with more effective and high quality T cell responses against influenza virus [99, 100]. This vaccine has now moved to the clinical trial stage.

1.4 Innate Immune Signalling in Response to Influenza Infection

1.4.1 Pattern Recognition of Influenza Virus

Pattern recognition is the means by which influenza virus is detected inside cells, allowing an early innate immune response to the virus to take place. Viral infections are detected via broad recognition of molecular signatures specifically present during infection or cell damage, known as Pathogen Associated Molecular Patterns (PAMPs) and Danger Associated Molecular Patterns (DAMPs). These signatures are recognised by Pattern Recognition Receptors (PRRs); these are specific human receptors that recognise viral and other pathogenic infection [101].

PRR recognition of PAMPs and DAMPs triggers intracellular signalling cascades. These generally alter cell transcription, or in the case or

NOD- LRR- and pyrin domain-containing 3 (NLRP3), have direct antiviral effects. PRR activation generally is augmented by the action of signal transducing adaptor proteins, that aid the formation of signalling complexes [102]. These complexes recruit kinases which induce nuclear import of Transcription Factors (TFs). The main transcription factors involved in PRR signalling are Interferon Regulatory Factor 3 (IRF3), IRF7 and Nuclear Factor Kappa-Light-Chain-Enhancer of Activated B Cells (NFκB). Nuclear translocation of the aforementioned TFs begins the induction of a cellular antiviral transcriptional state. Chiefly, these TFs incur the production of IFN (IRF3/7, reviewed in section 1.4.2) and proinflammatory cytokines (NFκB) [103].

Only a small number of PRRs have been implicated in the recognition of influenza virus. These are described below:

TLR7

Toll Like Receptor 7 (TLR7) is a member of the toll like receptor family found primarily in endosomes of plasmacytoid DCs [104] (fig. 1.4). It was shown recently that GU-rich Single Stranded RNA (ssRNA) is the ligand for TLR7, which is consistent with its recognition of influenza virus [105].

Mouse studies have given conflicting evidence as to the role of TLR7 in influenza infection. In a 2010 study, TLR7, in tandem with TLR3 was shown to be indispensable for control of influenza replication [106]. However, a 2012 study found that TLR7 was not required for primary control of the virus, but was needed for a full antibody response, suggesting that this interaction is important for the interface between innate and adaptive immune responses [107].

TLR7 activation engages the Myeloid Differentiation Primary Response 88 (MyD88) adaptor protein. In DCs, this pathway culminates in nuclear import of IRF7 and NKκB transcription factors [108]. As a result, TLR7 signalling is associated with both IFN and inflammatory cytokine signalling, although this pathway is more classically associated with production of IFN in the context of DCs.

TLR3

The second TLR involved in detection of influenza is TLR3. TLR3 is expressed throughout epithelial cells of the respiratory tract and localises to endosomes [109]. While it is established that TLR3 plays a role in the recognition of influenza [110], it is unclear how this occurs. TLR3 is known to recognise dsRNA; it was initially suggested that RNA duplexes generated during replication were recognised [109]. However, it is now known that influenza recruits host factors to prevent formation of dsRNA species as a means of innate immune evasion [111]. TLR3-inducing dsRNA is found in dying infected cells that are phagocytosed by DCs and macrophages [112]. It is possible that this is the main avenue by which TLR3 activation takes place (fig. 1.4).

TLR3 signalling is mediated through the adaptor protein TIR-Domain-Containing Adapter-Inducing Interferon-β (TRIF). This induces the import of NFκB and IRF3 into the nucleus , inducing the transcription of IFN (IRF3) and proinflammatory cytokines (NFκB) [113].

RIG-I

Retinoic Acid-Inducible Gene I (RIG-I) is a cytosolic PRR expressed in epithelial and immune cells [114]. The viral ligand for RIG-I is 5' triphosphate RNA; this is the form that influenza viral genomes take [115]. The binding of 5' triphosphate RNA to RIG-I triggers a conformational change in the protein, facilitating downstream signalling [116]. While the majority of influenza genomic RNA is found within the nucleus, it has been shown that RIG-I and viral RNA colocalise to viral stress granules within the cytosol [117].

In a 2015 study by Wu *et al*, it was demonstrated that RIG-I was the most important PRR for interferon induction in lung epithelial cells, highlighting its importance in early innate immunity [110]. However, a 2007 study demonstrated that in mice deficient for RIG-I signalling, there was no impairment of the adaptive immune response to influenza [118].

RIG-I signalling is facilitated by its adaptor protein, Mitochondrial Antiviral-Signaling Protein (MAVS). The signalling complex associated with MAVS causes nuclear import of IRF3 and NFκB, similar to TL3

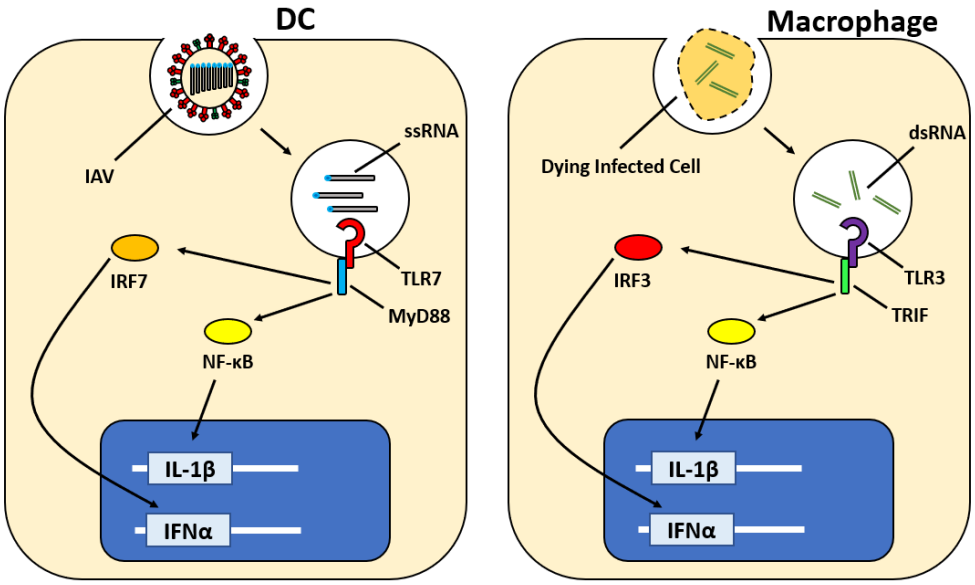


Figure 1.4: Cartoon showing PRR signalling via TLR7 (left) and TLR3 (right). In pDCs, ssRNA is recognised by TLR7 within endosomes. This triggers a MyD88-mediated signalling pathway, where IRF7 and NF-κB are trafficked to the nucleus. These TFs drive the expression of IFNs, such as IFN-α and proinflammatory cytokines such as IL-1β. In macrophages, dying cells infected by IAV are consumed via phagocytosis. dsRNA from IAV within these cells is recognised by TLR3, triggering a TRIF-mediated signalling pathway, where IRF7 and NF-κB are trafficked to the nucleus. Like with TLR7, this drives the expression of IFNs and proinflammatory cytokines. Adapted from [101].

signalling [119]. These TFs are associated with the transcription of IFN and proinflammatory cytokines.

NLRP3/Inflammasome

NLRP3 is a cytosolic PRR that associates with other proteins to form inflammasome complexes. Inflammasomes have roles in production of proinflammatory cytokines and pyroptosis. NLRP3 is expressed universally in airway epithelial cells and infiltrating immune cells. While molecular interactions between NLRP3 and influenza-associated PAMPs and DAMPs are poorly understood, it has been shown that influenza virus ssRNA, proton influx into the golgi via M2, and PB1-F2 aggregates activate NLRP3.

Unlike the previously discussed PRRs, activation of the NLRP3 inflammasome is not associated with the transcription of IFN or other cytokines. Instead, its activation is associated with the cleavage of the pro-caspase 1 protein into its active form, which in turn is associated with cleavage of pro-IL-1β into its active, and proinflammatory form. In addition, activation is associated with cell death via pyroptosis [120].

In a 2016 mouse study, inflammasome activity associated with NLRP3 was found to be beneficial for viral clearance at early stages of infection. However, at later stages of infection, suppressing NLRP3 activity reduced the severity of influenza virus pathogenesis [121]. In a 2013 study, these findings were largely echoed, with a finding that NLRP3 was specifically activated by PB1-F2. The authors postulated that the increase in influenza pathogenesis was associated with an observed increase in IL-1β production [122].

1.4.2 The Interferon Response

IFNs are PRR inducible intercellular signalling molecules, and are hallmarks of the innate immune response to virus infection [123]. Broadly speaking, IFN signalling induces a cellular antiviral state, strengthening host defense to an ongoing infection. They also have wide ranging effects om immune cells, and can induce both cell survival and death. Genes that are regulated under IFN signalling are known as Interferon Stimulated Genes (ISGs), or Interferon Regulated Genes (IRGs). IFNs are divided into three types, outlined below.

Type-I IFN

Type-I IFNs are comprised of the IFN-α and Interferon-β (IFN-β) genes, alongside some several little-studied members of the subgroup, IFN-κ and IFN-*omega*. In humans, there are 13 members of the IFN-α family. The effects of IFN-β are similar to the IFN-α family, with some differences pertaining to the genes induced by its binding. A hallmark of the type-I IFNs is the fact that they can be produced by the majority of cell types, while also having effects on nearly all cell types.

In nearly all cell types, the transcription of type-I IFN can be driven by PRR signalling. Promoters of IFN-α family members contain consensus sequences that interact with both IRF3 and IRF7 - these TFs drive the transcription

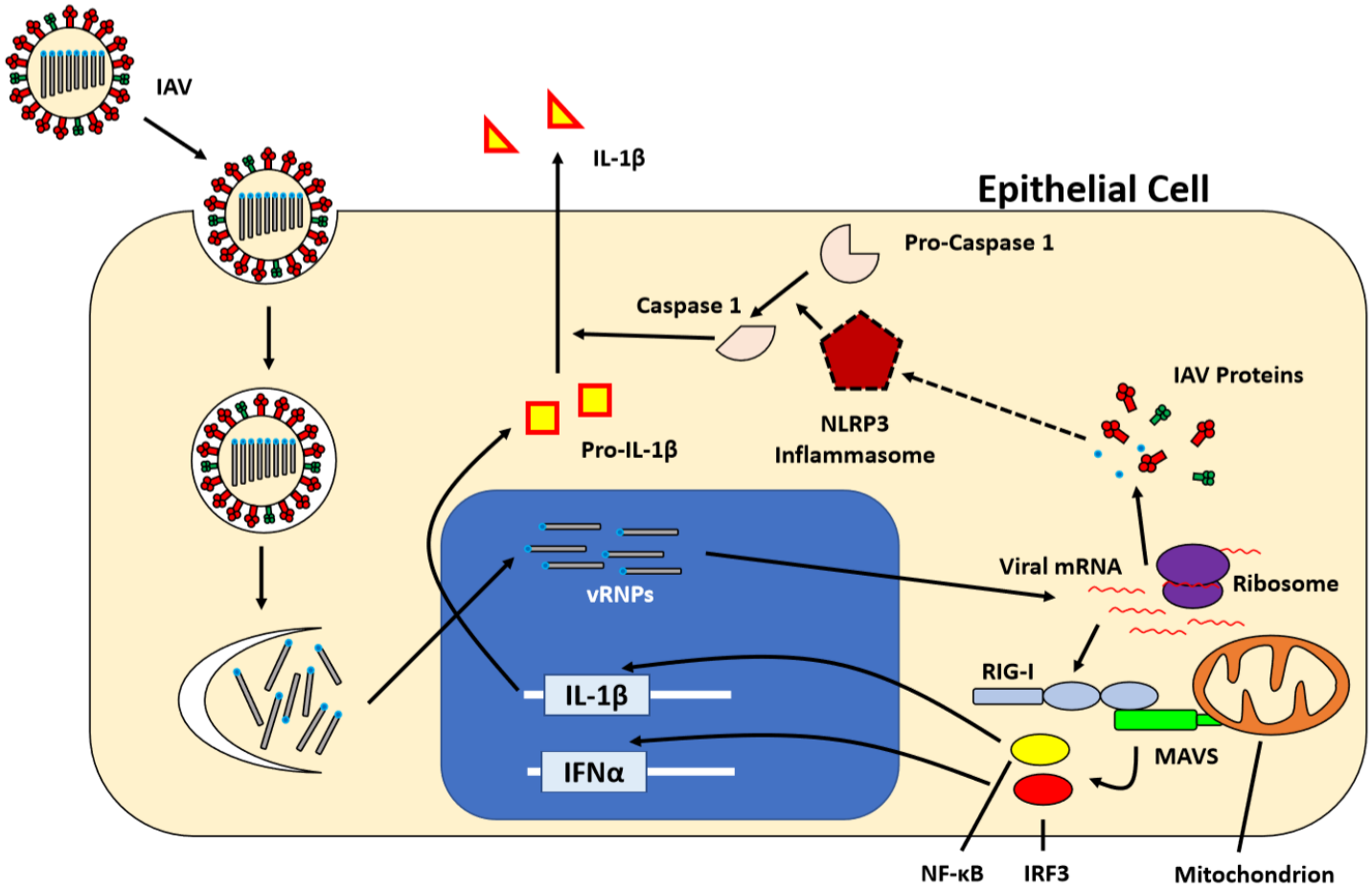


Figure 1.5: Cartoon showing PRR signalling within infected epithelial cells via RIG-I and NLRP3. IAV enters the cell via the endosomal pathway, and delivers vRNPs to the nuclueus. Viral mRNA is produced. RIG-I recognises viral mRNA without 5' caps, triggering a signalling cascade mediated by the MAVS protein. This protein is associated with the mitochondrial membrane. MAVS signalling occurs via NF-κB and IRF3. This induces expression of IFN and proinflammatory cytokines. Viral mRNA is translated into viral proteins, which can activate the NLRP3 inflammasome via various mechanisms. The inflammasome cleaves pro-caspase 1 to its active form. In turn, caspase 1 cleaves pro-IL-1β to its active form. Adapted from [101].

Figure 1.6: Cartoon representing types I, II and III IFN signalling. IFN signalling is initiated by binding of IFN molecules to cognate receptors. Type-I IFN signals via IFNAR, which works to phosphorylate STAT1 homodimers, STAT3 homodimers and ISGF3. Type-II IFN signals via IFNGR, which works to phosphorylate STAT1 homodimers and STAT3 homodimers. Type-III IFN signals via IFNLR, which works to phosphorylate ISGF3. Phosphorylated TFs are translocated to the nucleus, where they will interact with GAS (STAT1 homodimers) and ISRE (ISGF3) elements.

of IFN-α, and link PRR signalling and type-I IFN production. Similarly, the IFN-β transcriptional activity is driven by IRF3 and IRF7. However, unlike IFN-α, IFN-β transcription is also driven by NFκB signalling. Expression of IFN-β is induced preferentially by IRF3. Conversely the transcription of IFN-α members is preferentially driven by IRF7 [124]. IRF3 has higher constitutive expression then IRF7, causing IFN-β to be produced at earlier time points.

The broad activity of type-I IFNs is driven by the wide expression of their constitutive receptor complex. This receptor, known as Interferon α/β Receptor (IFNAR), is composed of the IFNAR1 and IFNAR2 proteins (fig. 1.6) and is used by both IFN-α and IFN-β. Canonically, type-I IFN binding activates kinases known as Janus Kinase 1 (JAK1) and Tyrosine Kinase 2 (TKY2). These kinases phosphorylate Signal Transducer and Activator of Transcription 1 (STAT1) and Signal Transducer and Activator of Transcription 2 (STAT2) TFs, which form a heterotrimeric complex with Interferon Regulatory Factor 9 (IRF9) known as Interferon-Stimulated Gene Factor 3 (ISGF3). ISGF3 is translocated to the nucleus, where it binds to promoter elements known as Interferon-Sensitive Response Elements (ISREs), and influences the expression of associated genes. Type-I IFN signalling is also associated with formation of STAT1 homodimers, which interact with promoter elements known as Gamma Interferon Activation Sites (GASs). Finally, IFNAR-associated kinase activity has been associated with formation of Signal Transducer and Activator of Transcription 3 (STAT3) homodimers - these repress the expression of inflammatory genes via an unknown mechanism. In a 2006 study, it was shown that the balance between STAT1/2 and STAT3 responses was important in coordinating the degree of inflammation associated with the type-I IFN response [125].

It is well understood that the binding affinity of IFN-β to the IFNAR receptor complex is higher than that of IFN-α. This difference is thought to underlie differential gene expression programs driven by binding of the two different IFNs. This is demonstrable in the context of antiproliferative genes, whose expression is driven by IFN-β but not IFN-α under physiological conditions [126].

Type-II IFN

In contrast to Type-I IFN, Type-II IFNs production is limited to specific immune cells. These cells include, but may not be limited to NK cells,

Natural Killer T (NKT) cells, T cells, macrophages, DCs, and B cells [127]. There is a single member of the type-II IFNs known as Interferon-γ (IFN-γ).

Regulation of type-II IFN production is not intrinsically linked to PRR signalling as with the type-I IFNs in the majority of cells. Except in the case of $\alpha\beta$-T cells, production of type-II IFN is controlled by cytokine stimulation in the immune cell microenvironment. While the finer points of this regulation are specific to the cells producing IFN-γ, the most prolific cytokines involved in stimulation of IFN-γ production are Interleukin-12 (IL-12) and Interleukin-18 (IL-18). It is thought that IFN-γ, IL-10, and IL-18 are produced in a paracrine positive feedback loop [128]. There is some sparse evidence that, in Ag-presenting cells, Lipopolysaccharide (LPS) can induce the production of IFN-γ, suggesting that PRR signalling may be involved in IFN-γ production in a limited capacity [129, 130].

The activity of Type-II IFN is relatively broad, due to widespread expression of its constitutive receptor, Interferon γ Receptor (IFNGR). Classical IFN-γ signalling shares some signalling pathway components with type-I IFN signalling. The IFNGR receptor is composed of IFNGR1 and IFNGR2 proteins (fig. 1.6), and like with type-I IFN, signals via associated kinase molecules. These are primarily JAK1 and Janus Kinase 2 (JAK2) [131], which polyphosphate STAT1 and STAT3. Like type-I IFN, STAT1 homodimers are trafficked to the nucleus, where they interact with GAS elements in gene promoters [132].

Type-III IFN

Type-III IFNs are poorly studied relative to IFN types I and II. Type-III IFNs are similar to to type-I IFNs in terms of induction and molecular function. While type-III IFNs can be produced by a wide range of cell types, their activity is mostly limited to epithelial cells due to receptor expression [133]. There are four members of the type-III IFN family - these are the members of the Interferon-λ (IFN-λ) family.

Regulation of IFN-λ production is, like with type-I IFN, linked to PRR signalling. IFN-λ genes are induced by the IRF3, IRF7 and NFκB TFs, albeit with a greater dependence on NFκB [134]. For this reason, in the majority of cells that produce type-I IFN, type-III IFN will be produced in tandem.

The signalling of IFN-λ is also similar to that of type-I IFN (fig. 1.6), with the largest difference in signalling being the receptor complex. Type-III IFNs signal via the Interferon λ Receptor (IFNLR) receptor [135]. This is composed of the IFNLR1 and Interleukin-10 Receptor Subunit β (IL10Rβ) proteins. Binding to IFNLR activates JAK1 and TYK2 molecules, which in turn induce formation of ISGF3. This binds to ISRE elements within promoters, controlling gene expression. While it is hard to disentangle type-I and III IFN responses in active infection, it is thought that type-III IFNs produce a subdued and longer-lasting IFN response when compared to type-I IFNs [136].

Tonic Interferon Signalling

It has has also been observed that low levels of type-I IFN are produced without any connection to PRR signalling. This is commonly known as tonic IFN production. Tonic IFN has been recognised for its role in maintaining immune homoeostasis and regulating cytokine cross-talk [137]. High throughput transcriptomics has revealed roles for tonic IFN signalling in regulating baseline expression of genes sensitive to type-I IFN (ISGs) [138]. It is thought that IFN-β is the master regulator of tonic IFN signalling, and is essential for all IFN signalling in general. This expression is likely dependent on NFκB and the Activator Protein 1 TF.

Maintenance of tonic IFN is crucial for IFN responses to virus infection; the STAT and IRF proteins (with the exception of IRF3) are dependent on tonic IFN for their baseline expression. Maintenance of STAT expression is crucial for non-IFN cytokines as well, like IL-6 and IL-10 [139]. Below average levels of tonic IFN have been associated, with poor outcomes to viral infection [140]. On top of this, overexpression of tonic IFN has been associated with autoimmunity, highlighting a need for nominal levels of tonic IFN in healthy individuals [141].

Interferon Cross-Talk

When considering the role of IFN in the innate immune response, it is important to recognise that different IFN types do not act independently of one another. There is well documented cross-talk between type-I and type-II IFN in the context of infection, with complex, cell specific interactions

occurring between the two interferon types. In the context of NK cells, it has been shown that type-I IFN stimulation can, for a limited time, directly induce the expression of IFN-γ. This induction is antagonised by STAT1 in a classical negative feedback loop [142, 143].

Type I and II IFN can also negatively regulate the effects of the other IFNs. Perhaps the most prominent example of this occurs in macrophages, where type-I IFN signalling downregulates surface expression of IFNGR [144, 145]. It is notable that both type-I IFN stimulation and PRR signalling are associated with downregulation of IFNAR surface expression [146, 147]. Overall, these interactions highlight the fact that type-I and type-II IFN regulation is highly interlinked. It should be noted however, that the extent of these interactions are not fully appreciable without consideration of the wider cytokine network.

Interferons in Influenza Infection

In influenza infection, IFNs play diverse roles in the control of influenza virus replication. Type-I IFNs are perhaps the most studied in this context; these studies have largely occurred in mouse models, and are often performed with *Ifnar* Knock-Out (KO) mice that cannot undergo IFN signalling. A 2013 study by Arimori *et al* showed, using *Ifnar*-KO mice, that a lack of type-I IFN signalling increased inflammation associated with IAV infection. However, a 2014 Davidson *et al* study showed the opposite; in this case, a loss of type-I IFN was protective in the context of infection [148]. In addition, a 2012 study showed that *Ifnar*-KO mice cleared bacterial infection secondary to IAV more effectively than Wild-Type (WT) mice [149]. Overall, the level of type-I IFN signalling appears to be important in determining whether it has beneficial or harmful effects in infection.

While type-I and type-III IFNs share similar signalling mechanism, recent findings have reported differential roles of these cytokines in infection. In a 2017 study by Galani *et al*, it was found that, in mice, IAV replication was initially controlled by IFN-λ, and type-I IFN activity was only engaged in the event of uncontrolled replication [150]. Type-I IFN led to the induction of proinflammatory cytokines, and the engagement of immune cell subsets, causing host damage. In early infection, this was attributable to type-I IFN action on invading neutrophils. This is supported by the findings of a 2016 Davidson *et al* study, where, in mice, therapeutic IFN-λ was found to be effective in the context of IAV infection, while IFN-α was

found to be detrimental [151]. These findings be partially explained by the fact that type-I, and not type-III IFNs are associated with formation of STAT1 homodimers, which are generally associated with the transcription of inflammatory cytokines [152].

In the context of IAV infection, IFN-γ expression occurs much later in the infection process. A 2016 mouse study found that type-II IFN was not significantly induced until seven days after infection [153]. In this study, it was also observed that a lack of IFN-γ promoted infiltration of neutrophils to the site of infection, and a higher levels of inflammation at the site of infection. Interestingly, a contrasting 2018 study found that depletion of IFN-γ in mice incurred increased survival rates in mice, and IFN-γ expression was appreciable at six days after infection [154]. Overall, the distinct role of IFN-γ in influenza infection is still poorly understood.

Generally, IFN regulation in the context of IAV infection is important in determining of the disease severity. It is likely that IFNs, with the possible exception of type-III IFN, are protective at low levels, but become immunopathogenic at higher concentrations. In fact, one of the molecular hallmarks of severe IAV infection is known as "cytokine storm", or overproduction of proinflammatory cytokines. It is thought that this symptom derives from excessive PRR signalling within respiratory epithelial cells, and that it may be coordinated in part by IFN responses [155].

1.4.3 Interferon Stimulated Genes

The primary function of IFN signalling is triggering transcriptional changes in affected cells. Collectively, genes which change in expression after IFN stimulation are known as IRGs; however, most literature related to gene regulated by IFN is in relation to stimulated genes (ISGs). ISGs tend to have antiviral, antiproliferative, or proinflammatory effects. Recently, efforts have been made to identify more ISGs, and obtain a broader perspective with regard to the effects of IFN on different cell types. This has been spearheaded by the Interferome database (discussed in section C.4.1) which aims to understand detailed effects of IFN types on different cell types using RNA microarrays and RNA-Sequencing [156].

As mentioned in section 1.4.2, ISG induction can be driven by TF binding to ISRE and GAS promoter elements. Genes with ISREs tend to have direct antiviral effects, while GAS-associated genes tend contribute more

to the inflammatory response [152]. It is common for ISGs to contain both ISRE and GAS elements, highlighting the fact that effects of stimulation by different IFNs overlap heavily. The concentration of environmental IFN is also highly related to the transcription that occurs in cells. This has been best characterised with type-I IFN, where it has been observed that low concentrations of type-I IFN are enough to stimulate antiviral gene expression, but more immunopathogenic genes, such as those involved in cell antiproliferation, are induced under higher interferon concentrations [157]. Finally, there is a great deal of stochasticity between cell types, and even within homogeneous cell populations with regards to IFN responsiveness [158].

Despite the fact that holistic understanding of IRG expression is currently limited, a number of targeted, reductionist studies on ISGs have been helpful in understanding the direct antiviral properties of ISGs. A large portion of these antiviral ISGs have been discovered due to discovery based screens, where ISGs are screened for antiviral effects. Perhaps the best example of this was a 2011 Schoggins *et al* study, where 380 different ISGs were tested, via an overexpression system, for their ability to restrict the replication of six different viruses [159]. ISGs which have been shown to antagonise influenza virus replication are reviewed in section 1.4.4. It is worth mentioning that a large number of the genes involved in PRR recognition, IFN induction and IFN signalling are themselves ISGs, highlighting the ability of IFN to self-regulate.

1.4.4 ISG Restriction of Influenza Virus

Currently, a handful of ISGs are known to restrict IAV via a range of different mechanisms. These are summarised in table 1.1.

The ISGs outlined in table 1.1 are likely not an exhaustive list. In addition, there are also restriction factors for IAV that are not ISGs [160].

- **MX Dynamin Like GTPase 1 (MX1)** is one of the most prolific ISGs involved in influenza restriction. It is thought that MX1 works to both retain vRNPs in the host cytosol, and attenuate replication via retention of viral proteins in the cytosol [161, 162]. Murine MX1 localises to the nucleus, not the cytosol, and is not present in the majority of laboratory mouse strains [163].

Table 1.1 ISGs Implicated in Influenza Virus Control (Adapted from Iwasaki et al [101])

ISG	Subcellular Localisation	Stage Targeted [1]
MX1	Cytosol	Nuclear Import & Replication
OAS1-3	Cytosol	Replication
RNase-L	Cytosol	Replication
PKR	Cytosol	Translation
IFITM1-3	Endosomes/Cell Surface	Fusion
Viperin	Endoplasmic Reticulum	Budding
CH25H	Cytosol	Fusion
TRIM22	Nucleus	Replication
ISG15	Cytosol	Replication
ZAP	Cytosol	Replication
SERPINE1	Extracellular Environment	HA Cleavage
ISG20[2]	Nucleus	Replication
TRIM56[2]	Nucleus	Replication

[1] Of viral life cycle [2] Conflicting evidence

- **2'-5'-Oligoadenylate Synthetase (OASL) 1-3** and **Ribonuclease-L (RNase-L)** act as part of the same molecular pathway. OASL genes generate a second messenger from Adenosine Triphosphate (ATP), which activates RNase-L. RNase-L is a broadly active RNase, which degrades both viral and cellular ssRNA, often causing apoptosis [164].

- **Protein Kinase R (PKR)** is a multi-functional ISG with numerous anti-influenza functions. For example, PKR phosphorlylates host factors that inhibit viral translation and stabilise IFN [165]. PKR has also been implicated in production of viral stress granules [117].

- **IFITM3**, alongside **Interferon Induced Transmembrane Protein 1 (IFITM1)** and **Interferon Induced Transmembrane Protein 3 (IFITM2)**, block viral fusion. These proteins are reviewed in section 1.5.

- **Viperin** is another multi-functional ISG implicated in IAV restriction. Current evidence points toward a role for viperin in inhibition of lipid raft formation, and hence viral budding [166].

- **Cholesterol 25-Hydroxylase (C25H)** is an ISG that alters cholesterol metabolism, inducing a cellular state less favourable to viral entry [167].

- **Tripartite Motif Containing 22 (TRIM22)** is an ISG that targets NP. This protein causes ubiquitination of NP, leading to its proteasomal degradation [168].
- **Interferon Stimulated Gene 15 (ISG15)** is an ISG that, like TRIM22, targets viral proteins. In this case, NS1 is bound to by ISG15, inhibiting NS1 protein function [169].
- **Zinc Finger Antiviral Protein (ZAP)** is an ISG with dual anti-influenza functions. The gene encodes two protein isoforms; the first of these induces ubiquitination and subsequent proteasomal degradation of PB2 and PA proteins [170]. The second isoform reduces levels of PB2 mRNA [171].
- **Serpin Family E Member 1 (SERPINE1)** is a recently described ISG, unique for its extracellular localisation. SERPINE1 is an inhibitor of extracellular proteases; this helps to limit HA cleavage in newly formed virus particles outside the cell [172].
- **Interferon Stimulated Gene 20 (ISG20)** is an exonuclease with anti-influenza activity *in vitro*. It was shown that overexpression of ISG20 limited the replication of IAV, and that ISG20 interacted with viral NP [173].
- **Tripartite Motif Containing 56 (TRIM56)** has recently been described as a host restriction factor for influenza virus. In a 2016 study, overexpression of TRIM56 in cell culture limited IAV replication, likely via impairment of RNA synthesis [155].

1.4.5 Antagonism of Innate Immunity by Influenza Virus

Importantly, despite numerous host resistance mechanisms, IAV still effectively replicates in human cells. This is thanks, in part, to virus-associated countermeasures to the innate immune response, which allow influenza to evade and attenuate the host innate immune response. These countermeasures are mostly associated with NS1, which targets the innate immune response via a variety of mechanisms.

NS1 primarily targets stages of the RIG-I signalling pathway. In a 2009 study, it was shown that NS1 inhibits Tripartite Motif Containing 25 (TRIM25);

this protein ubiquitinates RIG-I, a post-translational modification necessary for RIG-I signal transduction [174]. NS1 mediated suppression of another positive regulator of RIG-I signalling, known as Riplet, was also identified in 2012 [175]. In a 2015 study, a specific structural interaction was shown between NS1 and RIG-I in some influenza strains [176]. Finally, NS1 has also been described as an inhibitor of various members of the downstream RIG-I signalling pathway, namely IRF3 and NFκB [177, 178] Overall, these interactions work to attenuate a great deal of the early IFN response to influenza infection.

NS1 has also been shown to specifically bind Double Stranded DNA (dsDNA) as a mechanism of transcriptional inhibition. In a 2016 study, it was shown that the promoter and exonic regions of the *IFNB* gene are targeted by NS1, suppressing its transcription via antagonising RNA Polymerase II (RNAP-II) assembly on chromatin [179]. It is unclear if this interaction is specific to *IFNB*, or if this is part of a more generalised mechanism of transcriptional suppression. Due to the actions of NS1, most cultured cells do not mount an IFN response to WT IAV infection.

There are also non IFN-related mechanisms by which NS1 inhibits the host response. The best described of these is an interaction between NS1 and host cell polyadenylation machinery. By cleaving a critical host polyadenylation factor, the ability of the host cell to produce mature mRNA is attenuated [180]. In addition, NS1 inhibits PKR and OAS action [181, 182], demonstrating an ability to antagonise the later ISG response. Irrespective of the NS1 protein, influenza viruses also limit productive host mRNA production during cap-snatching. While the aforementioned mechanisms of host innate immune antagonism by influenza virus are not extensive, they demonstrate the myriad levels of interplay between the host response and viral countermeasures.

1.5 *IFITM3* and other *IFITM* Genes

1.5.1 The *IFITM* Gene Family

As discussed in section 1.4.4, *IFITM1*, *IFITM2*, and *IFITM3* are ISGs implicated in the restriction of influenza virus. These genes are part of the *IFITM* gene family. There are five functional members of this gene family;

IFITM1, *2*, and *3* are IFN-inducible, and have antiviral properties, while *IFITM5* and *10* are not IFN inducible or antiviral. *IFITM5* has roles in bone mineralisation; there is no known function for *IFITM10* [183].

The *IFITM* genes are found widely throughout metazoan species. Phylogenetic analysis places IFITM genes in a larger family of genes known as Dispanins; which likely arose in eukaryotes via horizontal gene transfer [184]. Dispanins are characterised by their dual transmembrane domains, which anchor the proteins within cell membranes. The number of *IFITM* gene copies varies widely between species. In contrast to the five functional *IFITM* genes found in humans, there are seven in mice, and four in chimpanzees. Overall, antiviral *IFITM* genes share close evolutionary relationships, even between species [185].

Surprisingly, alongside their antiviral effects, *IFITM* genes have been implicated as markers in various human cancers, with expression levels of *IFITM3* showing direct and potentially causal relationships with degree of malignancy [186, 187]. In mice, *ifitm1* and *ifitm3* genes have also been associated with germ cell homing and repulsion [188].

In this section, IFITM3 (and IFITM1 and 2) will be reviewed specifically in the context of its antiviral capacities.

1.5.2 The Antiviral Activity of IFITM Proteins

The IFN-inducible *IFITM* genes were first described in a 1984 Cell report by Friedman *et al* [189]. However, these genes were not well researched until 25 years after their discovery; the majority of studies focused on an apparent interaction between IFITM1 and B cell receptor components [190]. Widespread interest in *IFITM* genes did not occur until 2009, when a role for IFITM proteins as antiviral restriction factors was discovered in a Small Interfering RNA (siRNA) screen for host cell restriction factors for IAV [191]. In this study, IFITM3 also restricted Dengue and West Nile viruses.

In the past ten years, the list of viruses known to be restricted by IFITM proteins has been expanded (table 1.2). With the exception of Reovirus, all restricted viruses are enveloped - this is consistent with the idea that IFITM genes block membrane fusion events.

One of the ways in which IFITM-mediated viral restriction has been

Table 1.2 Viruses Restricted by Human IFITM Genes

Virus Family	**Virus**	**Restrictive IFITM**	**References**
Flaviviridae	West Nile Virus	1, 2, 3	*Brass et al, 2009* [191]
	Dengue Virus	1, 2, 3	*Brass et al, 2009* [191]
	Hepatitis C Virus	1, 2, 3	*Narayana et al, 2015* [192]
	Zika Virus	1, 3	*Savidis et al, 2016* [193]
	Japaense Encephalitis Virus	3	*Zhang et al, 2013* [194]
Filoviridae	Ebola Virus	1, 2, 3	*Huang et al, 2011* [195]
	Marburg Virus	1, 2, 3	*Huang et al, 2011* [195]
Coronaviridae	SARS Coronavirus	1, 2, 3	*Huang et al, 2011* [195]
	MERS Coronavirus	1, 2, 3	*Wrensch et al, 2014* [196]
Retroviridae	HIV-1	1, 2, 3	*Lu et al, 2011* [197]
	SIV	1, 2, 3	*Qian et al, 2015* [198]
Bunyaviridae	Hantaan Virus	1, 2, 3	*Mudhasani et al, 2013* [199]
	La Crosse Virus	1, 2, 3	*Mudhasani et al, 2013* [199]
	Andes Virus	1, 2, 3	*Mudhasani et al, 2013* [199]
	Rift Valley Fever Virus	1, 2, 3	*Mudhasani et al, 2013* [199]
	La Crosse Virus	2, 3	*Mudhasani et al, 2013* [199]
Orthomyxoviridae	Influenza A Virus	1, 2, 3	*Brass et al, 2009* [191]
	Influenza B Virus	1, 2, 3	*Everitt et al, 2012* [200]
Rhabdoviridae	Vesicular Stomatitis Virus	1, 2, 3	*Weidner et al, 2010* [201]
	Rabies Virus	2, 3	*Smith et al, 2013* [202]
	Lagos Bat Virus	2, 3	*Smith et al, 2013* [202]
Reoviridae	Reovirus	3	*Anafu et al, 2013* [203]
Pneumoviridae	Respiratory Syncytial Virus	1, 2, 3	*Zhang et al, 2015* [204]
Alphaviridae	Semliki Forest Virus	2, 3	*Weston et al, 2016* [205]
Asfarviridae	African Swine Fever Virus	1, 2, 3	*Munõz-Moreno et al, 2016* [206]
Herpesviridae	Cytomegalovirus	3	*Stacey et al, 2017* [207]

highlighted is via *ifitm3*-KO mice. Using this model, studies have shown associations between a lack of Ifitm3 and higher virus replication, higher death rates in infected mice, and differing cytokine and leukocyte responses to infection. This was shown with IAV in the 2012 Everitt *et al* study, but has also been shown with respect to West Nile virus and murine cytomegalovirus, among others [200, 207, 208]. In these studies, the phenotypes of *ifitm3*-KO mice are different in the context of different viral infections. For example, in IAV infection, a lack of Ifitm3 is consistent with an increase in viral replication and inflammatory cytokine production, while in murine cytomegalovirus infection, only cytokine production is affected.

There is some evidence linking *IFITM* genes to restriction of *Mycobacterium tuberculosis* (Mtb). An study by Everitt *et al* in 2013 found, that *ifitm3*-KO mice and WT mice were similarly susceptible to several bacterial and protozoan pathogens, including Mtb [209]. However, a 2016 Ranjbar *et al* study established a role for human IFITM proteins in Mtb clearance [210]. The authors postulated that the Everitt *et al* study did not take into account later infection time points, where Ifitm3 may play a more apparent role. This experiment was not performed. The notion that IFITM genes contribute to Mtb control is also supported by two studies linking *IFITM3* genetic variation to susceptibility to TB [211, 212].

1.5.3 The *IFITM3* genetic locus

In humans, the *IFITM* genes are clustered on the short arm of chromosome 11, roughly 300,000bp from the end (fig. 1.7). The antiviral *IFITM* genes are closer together still, with roughly 5kbase pairs (bp) of intergenic distance between each gene. *IFITM5* is located closest to the chromosome end, while *IFITM10* is roughly 1.4Mbp from the other *IFITM* genes, towards the centromere. *IFITM1*, *2*, and *10* are transcribed from the forward strand, while *IFITM3* and *IFITM5* are transcribed from the reverse. All canonical transcripts for *IFITM* genes have two exons.

In the current annotation of the genome (fig. 1.8), there are multiple proposed transcripts for each *IFITM* gene. These annotations do not confirm their existence, nor does it suggest that there is a protein product associated with them. There are predicted to be four different *IFITM1* isoforms, six *IFITM2* isoforms, and four *IFITM3* isoforms (table 1.3). For each gene, these isoforms use differing 5' Untranslated Regions (UTRs), and all are predicted to contain

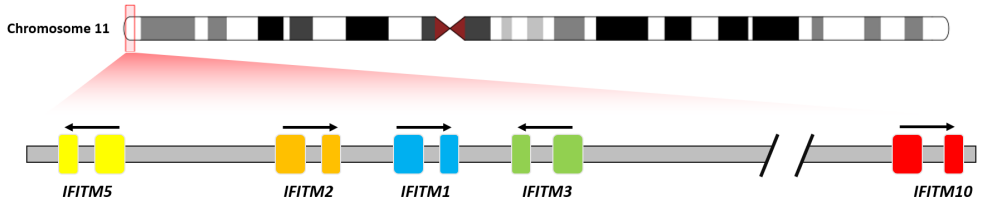


Figure 1.7: Cartoon of the IFITM locus in the human genome. The IFITM genes are densely clustered on the end of the short arm of chromosome 11. Notably, IFITM10 is farther in distance from the other IFITM genes, with roughly 1.4Mbp between it and IFITM3. Genes are transcribed from different strands. Adapted from Bailey et al, 2014 [213]

between 2 and 3 exons. There is one predicted IFITM1 protein product, three IFITM2 protein products, and two IFITM3 protein products.

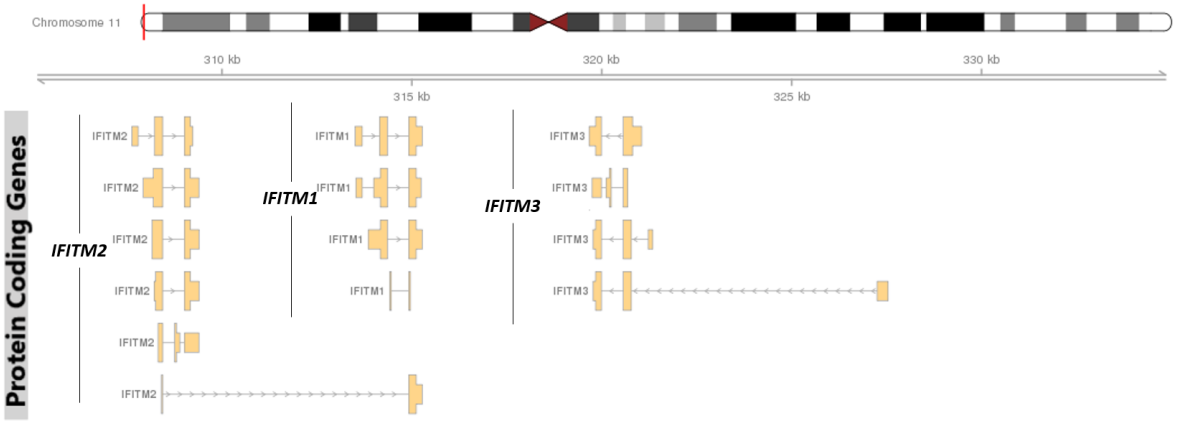
Two of the four *IFITM3* transcripts (ENST00000602735.2 and ENST00000526811.4) are derived from alternative 5' UTRs from the canonical transcript (ENST00000399808.4). One transcript has been associated with Nonsense-Mediated Decay (NMD) (ENST00000531688.1) (Explained in section 3.3). Shown in table 1.3 are Transcript Support Level (TSL) levels for each transcript. These are annotated by GENCODE 27 and refer to the amount of evidence that is available for each transcript [214]. Transcripts with the most evidence for their existence are TSL1, and ones with the lowest are TSL5. In reference to *IFITM1* and *IFITM2*, there is some evidence from transcript assemblies obtained on the GTEx browser, for alternative, tissue-specific, splicing of both genes [215].

Table 1.3 IFITM3 Predicted Transcripts

Transcript ID	**Base Pairs**	**Protein**[1]	**Biotype**[2]	**TSL**[3]
ENST00000399808.4	808	133	Protein Coding[4]	TSL1
ENST00000602735.2	713	112	Protein Coding	TSL5
ENST00000526811.4	543	112	Protein Coding	TSL5
ENST00000531688.1	500	55	NMD	TSL3

[1] Size in Amino Acids (aa) [2] Predicted [3] Transcript support level
[4] Canonical transcript

Despite the fact that the *IFITM3* splicing profile suggests the use of multiple 5' UTRs, there is only one characterised promoter for *IFITM3* (fig. 1.9). This promoter contains no TATA or CCAAT boxes - *IFITM1* however, does



(a) Transcripts for antiviral *IFITM* genes



(b) Annotated *IFITM3* transcripts

Figure 1.8: ***a.*** *Antiviral IFITM transcripts shown with respect to their genomic alignments. Direction of transcription of IFITM1 and IFITM2 is left to right, and right to left for IFITM3.* ***b.*** *IFITM3 transcripts shown in greater detail with respect to their genomic alignments, and annotated with their ensembl transcript IDs. The canonical transcript is shown in yellow. The alternative protein coding transcripts are shown in red, and the NMD transcript is shown in blue. Direction of transcription for IFITM3 is right to left*

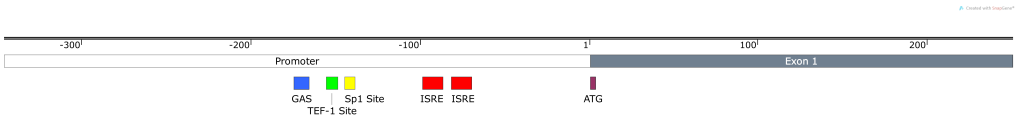
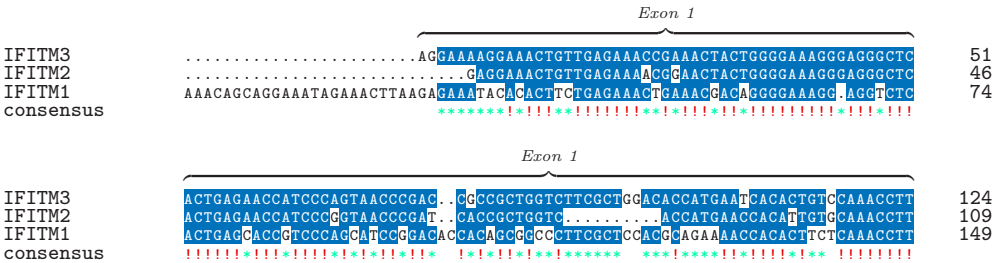


Figure 1.9: Map of the IFITM3 promoter region, showing reported binding sites for associated transcription factors.

contain two CCAAT boxes [216]. The *IFITM3* promoter contains two ISREs and a single GAS. The *IFITM3* promoter also contains a Transcription Enhancer Factor 1 (TEF-1) binding site, which when bound, suppresses *IFITM3* expression [217].

Cell specific methylation of the *IFITM3* promoter has been observed on multiple occasions. In human melanoma cells, methylation was reported to disrupt an interaction with the Specificity Protein 1 (Sp1) TF. This methylation was temporarily abrogated by IFN-α induction [218]. In a more recent report, methylation of the *IFITM3* promoter was observed to disrupt a CCCTC-binding factor (CTCF) binding site, which incurred an overall increase in *IFITM3* transcription [219].

The sequences of antiviral *IFITM* genes have a great deal of similarity, reflecting a recent evolutionary divergence. Complementary DNA (cDNA) sequences of *IFITM1*, *IFITM2*, and *IFITM3* are aligned in section 1.5.3. These sequences are obtained from National Center for Biotechnology Information (NCBI) (*IFITM3*: NM_021034.2, *IFITM2*: NM_006435.2, *IFITM1*: NM_003641.3). Via NCBI Basic Local Alignment Search Tool (BLAST), relative to *IFITM3* cDNA, *IFITM2* cDNA shows 97% sequence alignment, with 92% similarity. *IFITM1* cDNA shows 65% sequence alignment with 85% similarity [220]. The majority of divergence between the three genes is clustered at the sequence termini, with high levels of homology within coding sequences, and near the canonical exon-exon junction.



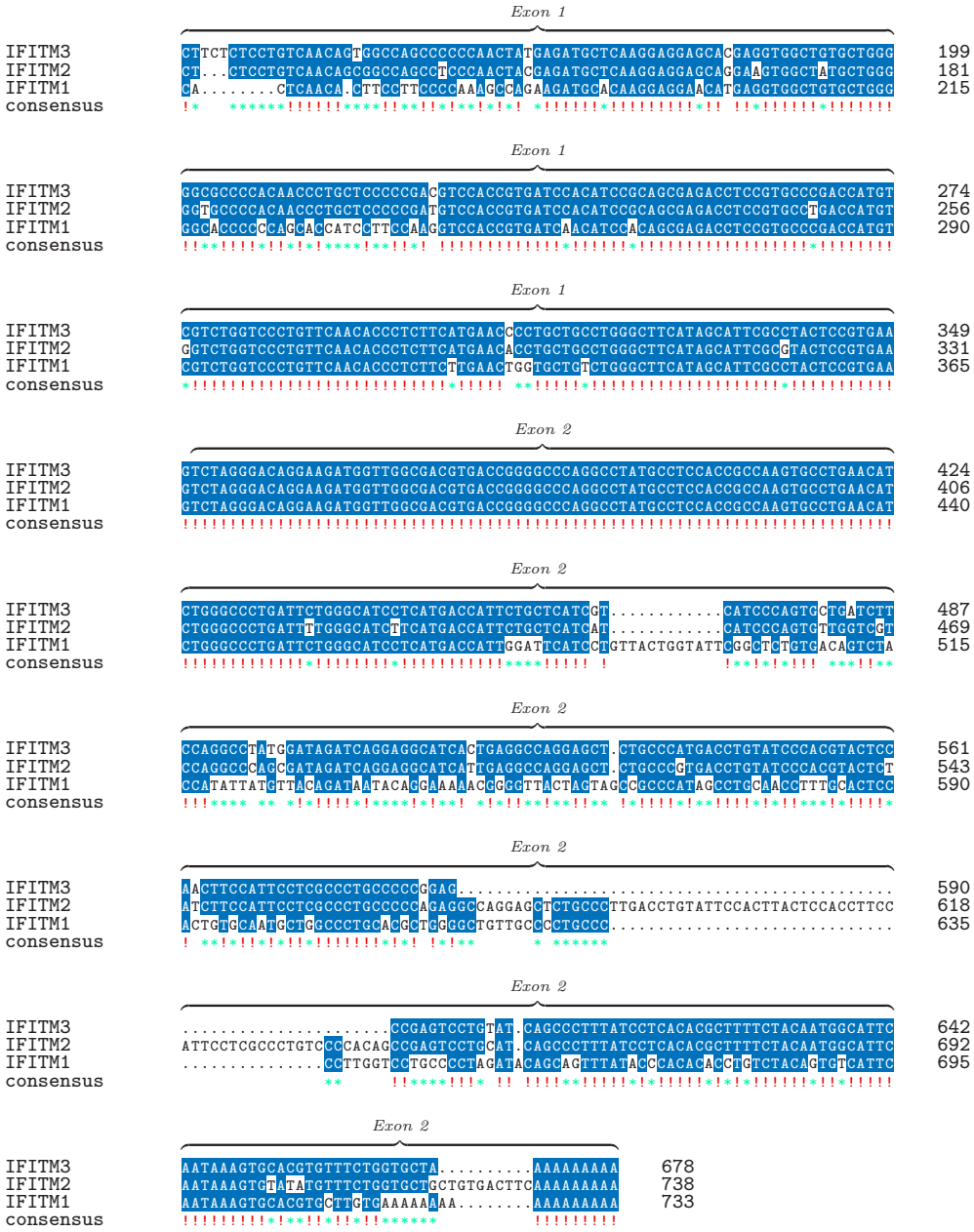


*Figure 1.10: mRNA sequence alignment between IFITM1, 2, and 3. Sequences show high levels of homology. Instances where all sequences agree are denoted in the consensus with a !. Where two of three agree, a * is used. The first and second exons for IFITM3 are shown above the alignment - these exon boundaries are consistent across all mRNA shown. cDNA sequences were obtained from NCBI reference sequences.*

1.5.4 The IFITM3 Protein

The IFITM3 protein is a small, membrane bound, 133aa, 15kilodaltons (kDa) protein, with high numbers of hydrophobic and uncharged amino acid residues. IFITM proteins are distributed throughout the endosome and cell surface, with IFITM1 localising more to the cell surface, and IFITM2 and 3 being more associated with the endosomal pathway [221]. Recently, key insights into the function and post-translational modification of IFITM3 have been made.

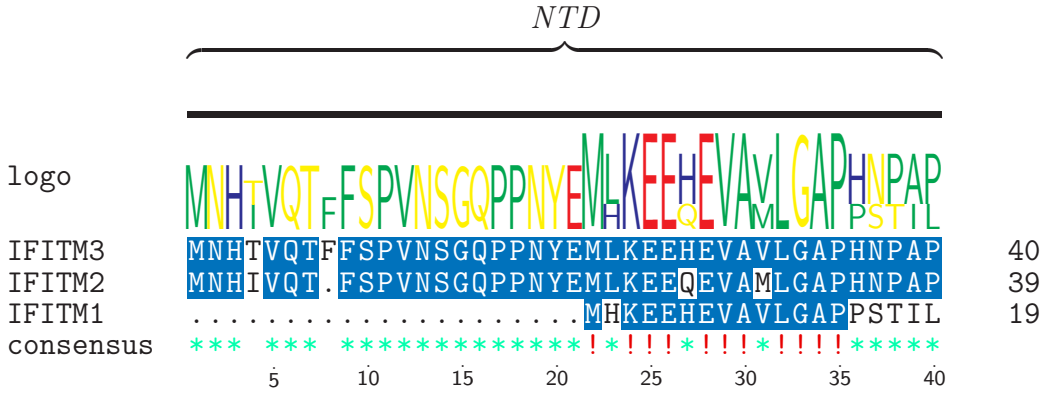
The IFITM proteins are, as predicted by their cDNA sequences in section 1.5.3, highly similar to one another in amino-acid sequence (fig. 11.1). Via BLAST, the human IFITM3 (AFF60304.1) protein shows a 98% alignment with IFITM2 (CAG46672.1), with a 92% sequence similarity. Relative to IFITM3 (NP_003632.3), IFITM1 shows a 79% sequence alignment, and an 84% sequence similarity. Again, the greatest degree of sequence divergence between IFITM proteins is at the C and N termini. The general hydrophobicity of IFITM proteins, alongside cDNA and protein sequence similarities, makes the IFITM genes, especially IFITM3 and IFITM2 challenging targets for protein synthesis, generation of targetting antibodies (addressed in section 3.2.1), and primer design.

In the 2009 Brass *et al* study, which garnered attention for IFITM proteins, five protein domains were suggested for IFITM3. These were:

1. **N-Terminal Domain (NTD)** - residues 1-57. This domain is conserved between IFITM2 and 3, but is divergent in IFITM1.
2. **Intramembrane Domain (IMD)1** - residues 58-80. This domain is highly similar between IFITM genes.
3. **Conserved Intracellular Loop (CIL)** - residues 81-104. This domain shares 100% sequence similarity between human IFITM genes.
4. **IMD2** - residues 105-126. This domain is highly similar between IFITM genes.
5. **C-Terminal Domain (CTD)** - residues 127-133. This domain is the shortest of the five, and is highly divergent between IFITM proteins in terms of length and aa composition.

With respect to the above domains, it should be noted that the IMDs did not initially pertain to a specific membrane orientation. Bearing in mind that the CIL was known to be intracellular, the orientation of IFITM3 was a subject of considerable debate for several years [222]. Initially, three models were proposed for IFITM3; one where the NTD and CTD were in the extracellular space, one where the NTD and CTD were intracellular, and one where the NTD was intracellular and the CTD was extracellular [223–225].

Currently, there is greater consensus on the orientation of IFITM3, based on a study using electron paramagnetic resonance and nuclear magnetic resonance imaging by Ling *et al* in 2016 [226]. In this study, it was reported that the NTD of IFITM3 is intracellular, while the CTD is extracellular. Three α-helices were found within IFITM3; the first two corresponded to IMD1, while the final, larger helix corresponded to IMD2. The first two helices were predicted to lie at the surface of the cytosolic side of the membrane, while the larger helix was a transmembrane domain, passing from the intracellular to the extracellular side of the membrane (fig. 1.12a). In a 2017 study, the aforementioned model was supported by bioinformatic analysis (fig. 1.12b) [227]. Here, the authors also suggested that the first helix was amphipathic, rather than hydrophobic, anchoring it to the membrane surface. It was found that amphipathic nature of this helix was essential for the antiviral activity of IFITM3.



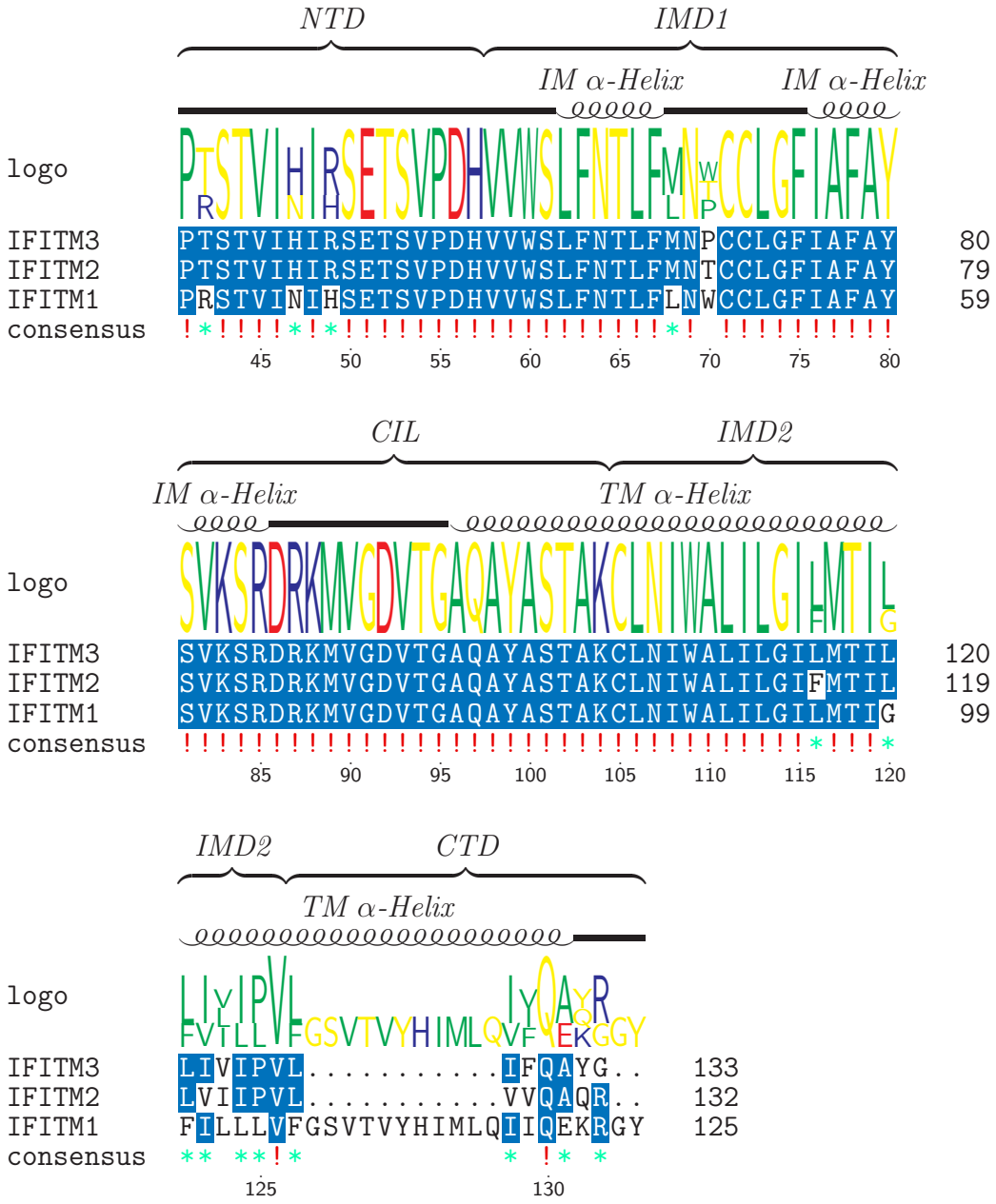
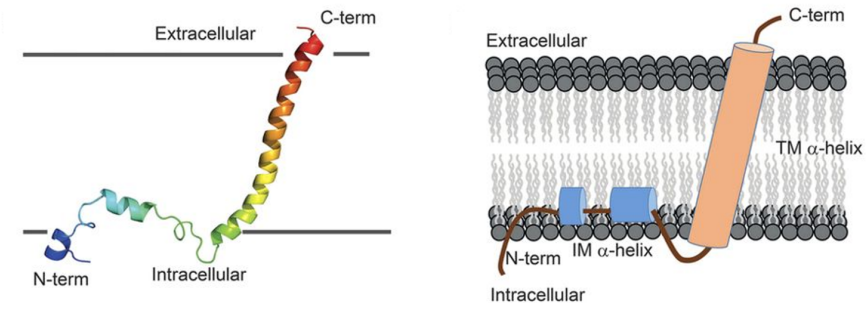
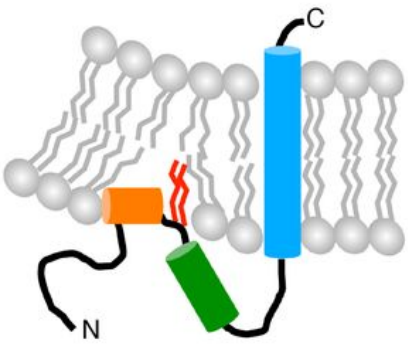


*Figure 1.11: Alignment between IFITM1, 2, and 3 protein structures, from N to C terminus. Conserved residues are shown in blue. In the consensus sequence, residues shared between all proteins are denoted with a !. Instances where two of three sequences agree are denoted with a *. The sequence logo is coloured according to hydrophobicity; hydrophobic residues are shown in green, polar uncharged residues in yellow, acidic residues in red, and basic residues in blue. The alpha helices found by Ling et al are denoted above the structure. The originally declared protein domains are shown above this.*



(a) Ling *et al* structures



(b) Chesarino *et al* structure

Figure 1.12: ***a.*** *Structures of IFITM3 predicted by the Ling et al study, showing the consensus IFITM3 topology. On the right, the surface helices are shown in blue, and the transmembrane helix is shown in orange.* ***b.*** *Predicted structures of IFITM via the Chesarino et al study. The amphipathic helix, found to be essential for antiviral activity, is shown in orange. S-Palmitoylation is shown in red.*

Structural determinants of IFITM3 activity are not fully understood, but have been partially characterised. A 2013 John *et al* study provided critical insights into IFITM3 function via the use of alanine scanning [228]. Here, it was shown that, tyrosine Y20, cysteine C72, arginine R85, tyrosine Y99, and lysine K104 residues were all particularly important for IFITM3 related IAV restriction. With the exception of K104, this was attributed to the fact that changing these residues dramatically redistributed IFITM3 within the cell. The two tyrosine residues appeared to be key phosphorylation sites on the protein. This study also revealed a role for IMD1 in the formation of IFITM-IFITM interaction, although the significance of this is unclear.

The John *et al* study was supported by a 2010 Yount *et al* study [229], where antiviral activity was found to be dependent on the S-palmitoylation

of cysteines C71, C72 and C105. In addition, the same authors showed that lysines K24, K83, K88 and K104 are ubiquitination targets - this was found to negatively regulate IFITM3 activity via redistributing its subcellular localisation [223].

On top of these findings, a 2012 Jia *et al* study showed a role for the IFITM3 NTD in control of its distribution [224]. Without the NTD, this IFITM3 localised to the cell surface instead of the endosomal pathway. This was supported by a 2014 study, where the dephosphorylated form of the NTD Y20 residue was found to form part of a YXXΦ motif, an endocytosis signal [230]. The tyrosine-kinase FYN was identified as the key regulator of Y20 phosphorylation. Further studies of of Y20 revealed an additional role in a PPxY motif, which is critical for the ubiquitination of the protein [231]. Phosphorylation of this residue was found to suppress ubiquitination, and hence degradation of the protein via the ubiquitin ligase NEDD4. This highlights the importance of the Y20 residue, and its dual role in negative and positive regulation of the protein.

The CTD of IFITM3 is less well characterised compared to the NTD. It is worth noting that this is the most divergent region of the protein, suggesting a role for the CTD in determining key differences between IFITM family members. While no studies regarding the IFITM3 CTD have been carried out specifically, it has been shown that the CTD of IFITM1 inhibits HIV-1 entry, potentially via direct interactions [232, 233].

1.5.5 IFITM3 Mechanism of Action

The mechanism by which IFITM proteins are able to restrict viral entry is currently unknown, with conflicting data. Currently, it is understood that IFITM proteins restrict virus replication at entry, and their subcellular localisation is critical to this action [191, 228]. This is consistent with the fact that IFITM3 localisation coincides with the endosomal pathway, where IAV and other restricted viruses enter the cell. It is also clear that IFITM3 alters the fluidity of membranes it inhabits, which is consistent with the properties of amphipathic helices like the one found in IFITM3 [227, 234]. Beyond this, however, the mechanism by which IFITM3, and other IFITM proteins restrict viral entry remains unclear.

One of the first studies investigating the IFTIM3 mechanism of action was published by Li *et al* in 2013 [235]. The authors argued that IFITM3 restricts

viral fusion before hemifusion, or when the outer leaflets of the endosome and the viral envelope fuse. By treatment of cells with chemicals blocking either hemifusion or full fusion, the authors showed that IFITM3-mediated viral restriction could be overcome by induction of hemifusion, but not full fusion from the hemifusion stage. Later in 2013, a report by Amini-Bavil-Olyaee *et al* described an interaction between IFITM3 and Vesicle-Membrane-Protein Associated Protein A (VAPA) via yeast-2-hybrid screening [221]. VAPA is a mediator of cholesterol homeostasis; its interaction with IFITM3 was described to incur cholesterol accumulation in the cell, which culminates in a reduction in membrane fluidity. While abrogation of this interaction was found to reduce viral restriction, this was performed via either deletion or replacement of the IMD2 domain of IFITM3, calling into question potential alternative effects of this change.

Both the Li *et al* and Amini-Bavil-Olyaee *et al* studies were challenged by a 2014 Desai *et al* study, where live cell imaging was used to assess IFITM3 mediated restriction [236]. In IFITM3 expressing cells, lipid dyes associated with endosomes and invading virus were found to mix in acidic endosomes, suggesting hemifusion was occurring. On top of this, it was found that IFITM3 inhibited the formation of fusion pores. The authors also altered cholesterol metabolism, showing through VAPA independent mechanisms, that virus restriction was not inhibited by cholesterol accumulation in the endosome. While this study likely recreates viral infection in the most realistic manner, no studies have managed to study this without the use of artificial systems that potentially alter membrane properties.

In a 2012 study by Wee *et al*, two crucial findings were made in relation to IFITM3 function [237]. Using a mouse *ifitm3*-KO model, the authors showed that a loss of Ifitm3 incurred a loss of clathrin at the cell surface, and decreased rates of clathrin-mediated endocytosis. In addition, a direct interaction was observed between Ifitm3 and v-ATPase; this interaction was crucial for nominal acidification of endosomes. Overall, this study suggests a role for IFITM3 in maintenance of the endosomal pathway.

More recently, the role for IFITM3 in the innate immune response has been expanded beyond its role in the alteration of membrane properties. In 2017, a Fu*et al* study reported an interaction between IFITM3 and Zinc Metallopeptidase STE24 (ZMPSTE24), a transmembrane metalloprotease [238]. This interaction was found to be necessary for viral restriction; IFITM3 did not effectively restrict IAV infection in *ZMPSTE4*-KO cells. This study indicates that the action of IFITM3 may be related to a larger protein

complex on the surface of endosomes, or at least cooperative action with other, non-IFITM proteins.

Interestingly, a 2017 study by Jiang *et al* described an interaction between IFITM3 and IRF3 [239]. It was shown that this interaction facilitated the autophagosomal degradation of IRF3. This effectively suppressed IFN-β induction in cells where IFITM3 was expressed at high levels, presenting, a role for IFITM3 in the wider innate immune signalling network.

Overall, studies into the mechanism by which IFITM3 functions have yielded mixed results. While it is clear that IFITM3 can alter membrane properties, actual mechanisms by which the protein restricts viral entry remain elusive. On top of this, recent evidence has suggested that IFITM3 may complex with multiple membrane proteins, affecting viral restriction, endosome dynamics and innate immune control.

1.5.6 Evolution of *IFITM* Genes

The evolution of the *IFITM* genes is highly dynamic and reflective of their role in the evolutionary arms race between viruses and their hosts. The *IFITM* genes have undergone many gene duplications, as highlighted by the different copy numbers between mammalian species. In primates and rodents, the *IFITM* genes are under positive selection, a common property of antiviral ISGs [185].

A more recent analysis of the *IFITM* locus by Compton *et al* revealed, using comparisons to primate genomes, that *IFITM3* is the most ancient of the mammalian antiviral *IFITM* genes [240]. Interestingly, *IFITM2* is a relatively recent evolutionary development; it only appears in humans, chimpanzees and gorillas. In this study, the authors argue that the NTD of IFITM proteins is crucial for viral specificity, likely due to its regulatory Y20 residue. This may explain the fact that this residue is lost in IFITM1. By diversifying the use of this residue, it is likely that IFITM proteins collectively form a more restrictive barrier against viruses that fuse at both the cell surface and endosome.

1.6 IFITM3 Genetics and Infection

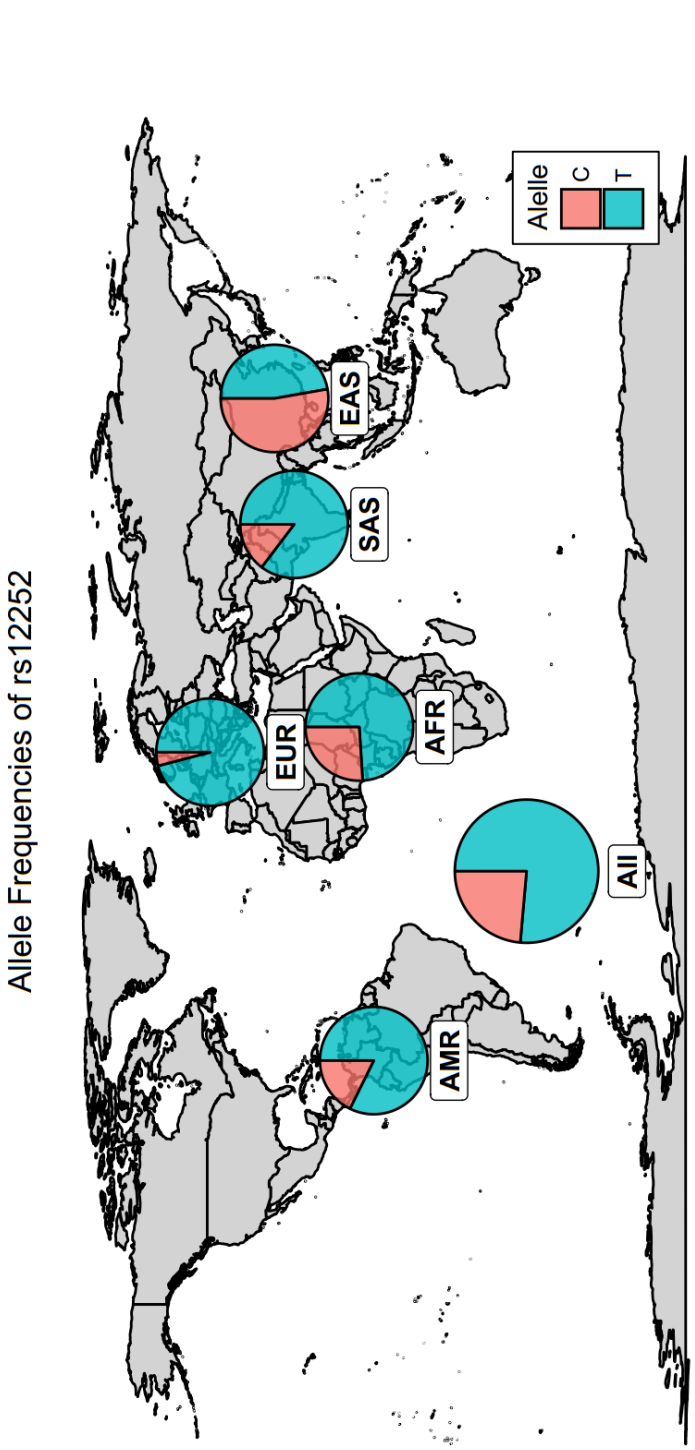
Since its role in viral restriction was characterised in 2009, one of the most intriguing aspects of *IFITM3* has been the role of its genetics in determining susceptibility to infection. The first instance where an *IFITM*-related genetic variant was associated with any phenotype was in 2012, in a seminal study that was published by Everitt *et al* [200]. This paper reported an association between the minor allele of a synonymous SNP in the first exon of *IFITM3* and hospitalisation during the 2009 H1N1 pandemic. This SNP is known as rs12252, and is associated with T (WT/major) and C (minor) alleles. This association was discovered based on sequencing of a 1.8kbp region of *IFITM3* in the aforementioned patient cohort, and investigating the overrepresentation of SNP genotypes compared to the background population rate (via 1000 genomes).

Since the publication of this study, a number of advances have been made in understanding the role of this SNP and others in *IFITM3*-related virus susceptibility. These will be reviewed in the following section. It should be noted that a number of these studies have run in parallel to our research into *IFITM3* genetics, which began in 2014.

1.6.1 rs12225 Population Genetics

Similar to the majority of SNPs in the human population, most information related to the allele frequencies of rs12252 derives from the extensive 1000 genomes project (reviewed in Appendix C.1) [241]. According to this resource the overall allele frequency of the T and C alleles are 0.76 and 0.24 respectively.

Like with a large number of SNPs, the population frequencies of rs12252 vary widely between world populations, outlined in table 1.4 and fig. 1.13. The allele frequencies for rs12252 are highest in East Asian (EAS) populations, with the lowest frequencies in European (EUR) populations. Over the 1,006 individuals genotyped for rs12252 from EUR populations, none were homozygous for rs12252-C. In EAS and EUR subpopulations, the allele frequencies of rs12252 are relatively invariant (table 1.5, fig. 1.14). The highest frequencies of the minor allele (C) were found in Japanese populations from Tokyo.



(a) Global Distribution of rs12252 Allele

Figure 1.13: World map with allele frequencies of world populations superimposed upon it. Data is taken from the 1000 genomes project.

Table 1.4 Global Allele Frequencies of rs12252. More details on population groups are shown in Appendix C.1

(a) Genotype Frequencies

Population	TT Frequency	CT Frequency	CC Frequency
AFR	0.549	0.381	0.070
AMR	0.677	0.291	0.032
EAS	0.244	0.456	0.300
EUR	0.918	0.082	0
SAS	0.736	0.233	0.031
All	0.549	0.381	0.070

(b) Allele Frequencies

Population	T Frequency	C Frequency
AFR	0.740	0.260
AMR	0.823	0.177
EAS	0.472	0.528
EUR	0.959	0.041
SAS	0.853	0.147
All	0.764	0.236

Table 1.5 Allele Frequencies of rs12252 in East Asian Populations

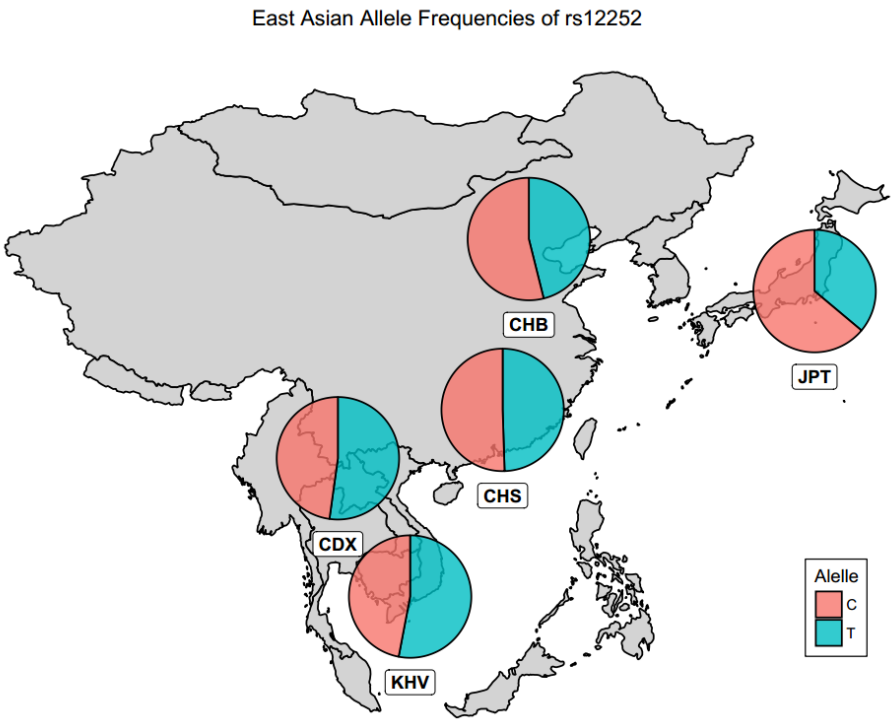
(a) Genotype Frequencies

Population	TT Frequency	CT Frequency	CC Frequency
CDX	0.258	0.527	0.215
CHB	0.184	0.553	0.262
CHS	0.267	0.457	0.267
JPT	0.183	0.356	0.462
KHV	0.333	0.394	0.273

(b) Allele Frequencies

Population	T Frequency	C Frequency
CDX	0.522	0.478
CHB	0.461	0.539
CHS	0.495	0.505
JPT	0.361	0.639
KHV	0.530	0.470

This data is corroborated by other sources of allele frequency data. Chief among these is the gnomAD dataset - this dataset is composed of a combination of genome and exome sequencing data, obtained in parallel



(a) Global Distribution of rs12252 Allele

Figure 1.14: Map of Eastern Asia, with population frequencies for subpopulations superimposed upon the geographical location of sample collection. Data is taken from the 1000 genomes project.

with the ExAC dataset (reviewed in Appendix C.2) [242]. While these samples are divided into world populations, they are more generalised groups. In this cohort however, 9,421 East Asian individuals were genotyped for rs12252, giving a C allele frequency of 0.5611. In European (Non-Finnish) and European (Finnish) populations, allele frequencies for rs12252-C were 0.043 and 0.078 respectively. These cohorts used 33,324 and 3,307 individuals respectively. In the UK 10,000 genome project, allele frequencies were found to be 0.034 for the rs12252-C, again in line with previous findings [243].

As part of the 1000 genomes project, ancestral alleles for all alleles, including rs12252, were inferred from Ensembl Compara - this resource contains a six-way alignment of primate genome sequences [244]. With regards to rs12252, the ancestral allele is considered the C allele. It is interesting to note that the ancestral allele for rs12252 is considered to be the minor allele and risk allele.

1.6.2 Association Between rs12252 and Viral Infection

At this point in time, a wide range of clinical studies have taken place investigating the role of rs12252 in the context of viral infection. In the original Everitt *et al* study, only 53 patients were used for analysis. This is a relatively small number of individuals in a population where the allele frequency of rs12252 is rare. No population stratification analysis was performed on these patients; the study instead relied on self-reported ethnicity, potentially causing undue over-representation of rs12252-C, outside of its potential effects on susceptibility to IAV.

In the following studies, a number of different genetic models were used to assess the effects of rs12252 on risk for viral infection. These models have been standardised within this section and are outlined in table 1.6.

Table 1.6 Genetic Models for rs12252

Model Name	**Group 1**	**Group 2**	**Group 3**
Allelic	C	T	N/A
Additive	TT	CT	CC
Dominant	TT	CT + CC	N/A
Recessive	CT + TT	CC	N/A

Despite the limitations of this study, eleven additional studies have investigated rs12252 and influenza infection (table 1.7). These studies have taken place mainly in European or East Asian populations; studies carry significantly more weight in the latter population given the allele frequencies of rs12252. The first of these was a 2013 Zhang *et al* study, where a significant association was observed between the C allele and risk for severe infection during the 2009 H1N1 pandemic [245]. Using 83 patients, this association was significant under recessive, additive, and allelic genetic models. This cohort is expanded upon in Appendix A.1, and samples from it are used in chapter 6.

At present, there have been four studies in East Asian populations - this association has been observed in all cases. In European populations, associations between rs12252 and this SNP have been investigated seven times in total. Despite this, an association between rs12252 and influenza infection has only been observed in two of the studies. This could reflect the lower statistical power associated with performing these studies in European populations, or differing effects of the SNP within different population

groups.

To address this association more clearly, two meta-analyses have been performed with regards to this SNP and influenza infection. The first of these, authored by Xuan *et al* in 2015, used four studies - the *E*veritt *et al*, Mills *et al*, Zhang *et al* (2013), and Wang *et al* studies (table 1.7) [258]. This study used different genetic models, and split data according to ethnicity and severity of influenza infection. However, it did not split the ethnic groups into different severities. In Europeans, this study found a significant association between rs12252 genotype under the recessive model (Odds Ratio (OR) = 10.6, 95% Confidence Interval (CI) = 2.99-37.74), but not under the allelic and dominant genetic models. In East Asians, the allelic (OR = 1.88, 95% CI = 1.34-2.62), dominant (OR = 2.68, 95% CI = 1.39-5.16), and recessive (OR = 2.04, 95% CI = 1.26-3.03) models all significantly associated with IAV susceptibility. In addition, there was a significant association observed between severe infection and rs12252 genotype under all 3 genetic models.

While the Xuan *et al* meta-analysis was important for establishing the association between rs12252 and influenza infection, it was limited in number of studies. In 2018, a novel meta-analysis was performed by Prabhu *et al*, using all studies in table 1.7 associated with IAV [259]. This was significantly more thorough, using combinations of genetic models, all ethnicities, and all severities. With respect to East Asians, the most significant associations were observed when using allelic (OR = 2.21, 95% CI = 1.22-4.03) and recessive (OR = 3.16. 95% CI = 1.42 - 7.01) models in severe patients. Associations between rs12252 genotype and mild disease were not observed. In Europeans, associations between rs12252 genotype and both disease severities were observed, although this was not consistent across all genetic models. The most significant association between severe disease and rs12252 genotype was observed under the recessive model (OR = 9.80, CI = 2.42-39.62), and the most significant association between mild disease and rs12252 genotype was observed under the allelic model (OR = 1.48, CI = 1.11-1.97).

Overall, it is highly likely that rs12252 genotype is associated with severe influenza infection, especially with respect to East Asian populations. It is notable that observed ORs are widely variable, and relatively high; an OR near 2 associated with rs12252 genotype and severe infection would suggest a profound influence of a relatively common genotype on disease susceptibility. However, whether or not rs12252-C incurs risk of mild-influenza disease risk still appears to be somewhat uncertain, especially in regards to East Asian populations. With this in mind, further studies are required to fully ascertain

Table 1.7 Clinical Studies with rs12252

Study	**Virus**	**N Cases**	**N Control**	**Population**[1]	**Severity**[2]	**Association**[3]
Everitt et al[200]	IAV	53	404[4]	EUR	S	√
Mills et al[246]	IAV	293	2623	EUR	M/S	√/×
Lopez-Rodriguez et al[247]	IAV	118	246	EUR	M/S	×/×
David et al[248]	IAV	41	200	EUR	M/S	×/×
Randolph et al[249]	IAV	185	99[4]	EUR	S	×
Gaio et al[250]	IAV	268	351	EUR	S	×
Carter et al[251]	IAV	180	10013	EUR	S	×
Mehrbod et al[252]	IAV	79	125	Iranian	M	√
Zhang et al[245]	IAV	83	103[4]	EAS	M/S	×/√
Wang et al[253]	IAV (H7N9)	16	93[4]	EAS	S	√
Lee et al[254]	IAV	275	105[4]	EAS	M/S	×/√
Pan et al[255]	IAV (H1N1/H7N9)	245	65	EAS	S	√
Zhang et al[256]	HIV-1	178	196	EAS	N/A	√
Xu-Yang et al[257]	Haantan Virus	69	208[4]	EAS	M/S	×/√

[1] EUR, European; EAS, East Asian [2] M, Mild; S, Severe [3] Between rs12252 and infection; √, Significant; ×, Not Significant
[4] Via 1000 genomes

the relationship between this SNP and mild influenza infection, especially in world populations not previously investigated

rs12252 genotype has also been associated with other viral infections. For example, a 2015 Zhang *et al* study found associations between rate of rapid progression of HIV-1 infection under the dominant model (OR = 3.78, 95%CI = 1.47-9.72). A 2016 Xu-Yang *et al* study also described associations between rs12252 genotype and Haantan virus infection. All of the aforementioned studies found the C allele to be associated with more severe outcomes in the context of infection.

Finally, rs12252-C has been associated with pre-existing antibody titres. In a 2018 Qin *et al* study, young adults with the CC genotype had higher Ab titres than CT or TT carriers to influenza A (H1N1)pdm09 before vaccination [260]. This was not true for influenza A (H3N2) and influenza B strains. The TT genotype was associated with higher Ab titres after vaccination than CT and CC carriers after one month, although titres were equal after 1 year. This suggests that rs12252 may affect vaccine efficacy, although this would need to be validated with more experimental data.

It is interesting to note that, despite this evidence, that associations between rs12252 and any phenotype have never been uncovered in any GWAS. With this being said, as mentioned in section 1.2.2, only one GWAS has been carried out with respect to influenza virus, which used a European population [68]. However, multiple GWASs have been carried out in relation to HIV infection, which have also not flagged rs12252. Strangely, the *IFITM3* gene has only been implicated in a single GWAS, where SNPs were related to the leukocyte composition of blood [261].

1.6.3 The rs12252 Locus

The rs12252 polymorphism is located at position 320772 on chromosome 11. As mentioned previously, this SNP is located within the first exon of *IFITM3*, and confers a synonymous change to the coding sequence (fig. 1.15). It forms part of a codon encoding a serine residue. Notably, rs12252 is considered the first base in the second exon of both alternative protein coding transcripts (ENST00000602735.2 and ENST00000526811.4), placing it within the predicting splicing acceptor site for these two transcripts (fig. 1.16). It is annotated as a splice region variant in GENCODE 27 genome annotations, based on this location.

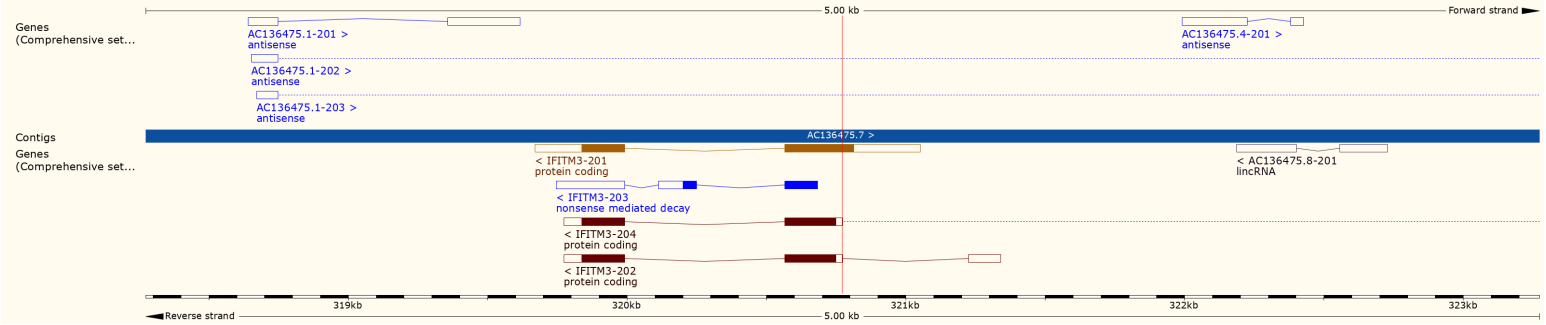
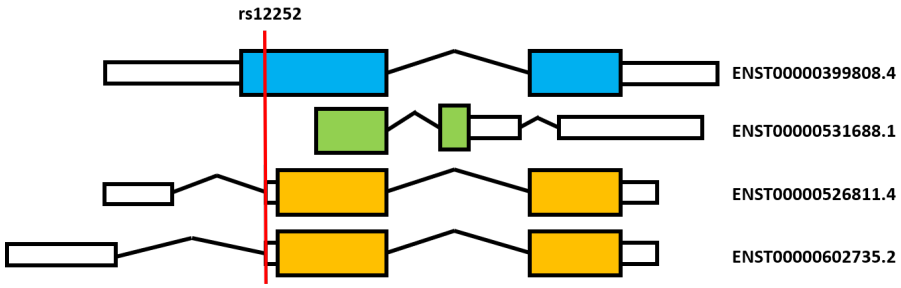


Figure 1.15: Map from ensembl genome browser depicting the location of rs12252 within the genome, shown as a red vertical line. IFITM3 is transcribed from the negative strand, or right to left. The canonical IFITM3 transcript is shown in gold. The NMD transcript is shown in blue, and alternative protein coding transcripts are shown in yellow.



(a) Predicted *IFITM3* Isoforms

Figure 1.16: Cartoon showing predicted IFITM3 isoforms with respect to rs12252 location. Introns are not drawn to scale. The blue isoform is the canonical isoform for IFITM3. The isoform predicted to undergo NMD is shown in green. Both yellow isoforms are predicted to be associated with production of a truncated form of the IFITM3 protein. Each yellow isoform uses a distinct 5' UTR; this is not shown in the case of ENST00000602735.2, as it appears far upstream. The location of rs12252 is shown in red. Cartoon is adapted from alignments visualised in the ensembl genome browser.



Figure 1.17: Map of the canonical IFITM3 transcript, with rs12252 marked in red. The canonical start site, and predicted alternative downstream start site associated with alternative protein coding transcripts are shown in maroon. The rs34481144 and rs3888188 SNPs are shown in blue

1.6.4 Proposed rs12252 Mechanism of Action

In the initial Everitt *et al* report, it was proposed that rs12252 could alter the activity of the splicing acceptor site associated with the SNP. The authors suggest that rs12252-C may be related to differential splicing of *IFITM3* mRNA, specifically the alternative protein coding isoforms . These isoforms are predicted to produce a protein product from a downstream, alternative start codon, given the location of the splice site after the canonical start site on the *IFITM3* transcript (fig. 1.17). Hence, these isoforms are predicted to encode an IFITM3 protein with an N-terminal truncation of 21aas.

This study also included an experiment where the cDNA of N-terminally

truncated IFITM3 protein was transfected into A549 cells in culture, and compared to the same cells with transfected WT IFITM3. In this experiment, the truncated form of the protein was redistributed to the cell surface, and was less effective at restriction of IAV than the WT protein. Experiments similar to this one have been reproduced in other studies [224, 240, 262]. Interestingly, in contrast to the findings of Everitt *et al* and Jia *et al*, a 2014 Williams *et al* study reported that the N terminally trunctated form of the protein had little effect on IAV restriction under a similar transfection system to that used by Everitt *et al*, despite the difference in subcellular distribution.

This hypothesis is addressed directly in chapter 3 and chapter 4. As of August 2018, it has also been addressed in a report by Randolph *et al*, which is discussed in section 3.3 [249].

1.6.5 Other Notable *IFITM3* Polymorphisms

While the rs12252 SNP is by far the most discussed *IFITM3* SNP within scientific literature, several other SNPs have been reported to modulate IFITM3 related control of pathogens.

The first, and most prolific of these was first reported on in 2017 by Allen *et al* [219]. This remains as the only other *IFITM3* SNP associated with IAV infection. In this study the minor allele of the rs34481144 (fig. 1.17) (major G, minor A) SNP was reported to associate with severe influenza infection using three different, largely European cohorts. This SNP is found in the promoter region of *IFITM3*, and was shown to associate strongly with expression levels of *IFITM3* mRNA, making it an Expression Quantitative Trait Locus (eQTL). This was shown both using patient cohorts and via *in-vitro* systems. Further work implicated the A allele of this SNP in enhanced binding of CTCF to the *IFITM3* promoter, which has a repressive effect on *IFITM3* expression. Finally, this study associated the SNP with enhanced *IFITM3* promoter methylation of CD8+ T cells in nasal washes, and general transcriptional repression of the *IFITM3* genomic neighbourhood. The latter phenomenon was attributed to the fact that CTCF has been found to associate with broad changes in regional expression via alteration of chromosome topology.

It should be noted that the population genetics of rs34481144 are widely different from those of rs12252. The rs34481144 shows relatively even allele frequencies in European populations (G: 54%, A: 46%, via 1000 genomes).

In contrast, in East Asian populations, the A allele has an allele frequency of merely 1%. All other world popluations have Minor Allele Frequencys (MAFs) between those in European and East Asian populations. This makes combinatorial analyses between these two SNPs difficult, although it is potentially feasible in populations with intermediate allele frequencies for both alleles. Interestingly, the Randolph *et al* paper reported that every instance of the CC rs12252 genotype was associated with the GG genotype of rs34481144, indicating that, in European populations, the risk variants for each allele are on opposite haplotypes. If this is applicable generally across world populations, it would make disentanglement of their effects extremely difficult outside of the context of *in-vitro* DNA editing.

Outside of IAV infection, two studies have reported associations between an *IFITM3* promoter SNP and risk of various forms of tuberculosis. The first of these was published by Shen *et al* in 2013 [211]. Here, the minor allele of an *IFITM3* promoter SNP, rs3888188 (fig. 1.17) (major T, minor G), was associated with susceptibility to paediatric tuberculosis. The authors, through a series of *ex vivo* and *in vitro* experiments attributed this to a reduction in *IFITM3* expression in carriers of the G allele. This observation was supported by a 2016 Naderi *et al* report, which observed an association between rs3888188 and risk of pulmonary tuberculosis in adults [212]. The allele frequencies of this SNP are more consistently distributed than rs12252 or rs34481144. Via the gnomAD study, the highest frequency of the minor allele is found in East Asian populations (T: 32%, G: 68%) and the lowest is found in Finnish populations (T: 80%, G: 20%). rs3888188 was not included in the 1000 genomes panel, and therefore Linkage Disequilibrium (LD) values between it and other SNPs common in East Asian populations is unknown.

Overall, these studies point to a role for *IFITM3* expression in the modulation of disease severity. In all, this presents a different model by which genetic variation of *IFITM3* can affect disease severity than the one proposed for rs12252, and suggests that, looking forward, SNPs that can modulate *IFITM3* expression may be influential in determining outcome to infection. When rs12252 was initially associated with severe IAV, it was thought its effects should potentially be visible in different world populations. The finding that SNPs common in European populations are associated with severe infection alongside rs12252 may help to explain the fact that rates of severe influenza infection do not appear to vary widely across world populations when discounting socioeconomic factors [43]. Despite this, the fact that these SNPs are fixed at high allele frequencies in some world populations is not well understood, and may suggest positive aspects to these

genotypes that are not current appreciated.

1.7 Overall Aims

In our research, we aimed to elaborate on the work of Everitt *et al* and Zhang *et al* in identifying rs12252 as a risk variant. Specifically, we aimed to investigate the mechanistic hypotheses postulated by Everitt *et al*. In addition, we were interested in gaining a greater understanding of the molecular effects of the risk variant in the context of health and infection. This was facilitated largely via collaborative works with Dr. Yan Zhao and Dr. Yonghong Zhang of Beijing You'An Hospital (of the 2013 and 2015 Zhang *et al* studies). Through this collaboration, we had unique access to patient samples with high frequencies of all rs12252 genotypes. With this in mind, our initial research aims were:

- Investigate the role of alternative splicing in the effects of rs12252 (chapter 3)
- Investigate the hypothesis that rs12252 is in LD with a true risk variant for the observed phenotype (chapter 3)
- Investigate other potential mechanisms associated with rs12252 (chapter 4)
- Investigate the molecular effects of rs12252 on IAV infection (chapter 6)

Chapter 2

Methods and Materials

2.1 General Methods and Materials

2.1.1 Extraction of gDNA from Cells

Adherent cells were washed with Dulbecco's Phosphate Buffered Saline (DPBS). Trypsin-Ethylenediaminetetraacetic Acid (EDTA) 0.05% (Thermo Fisher Scientific) was added to cells for 1-5 minutes. D10 was added to neutralise the trypsin-EDTA. Cells were centrifuged (423g, 5 minutes). Cell pellet was frozen at -20 C or immediately used for Genomic DNA (gDNA) extraction.

Cells were resuspended in 600µL of Cell Lysis Solution (QIAGEN). Cells were vortexed, and incubated at Room Temperature (RT) for 30 minutes. 200µL of Protein Precipitation Solution (QIAGEN) was added to lysis solution. Cell lysate was vortexed and transferred to 1.5ml microcentrifuge tubes. Cell lysate was centrifuged (17,000g, 25 minutes). Supernatant was transferred to a new microcentrifuge tube. 600µL isopropanol was added, and mixed well before centrifugation (17,000g, 5 minutes). Supernatant was discarded, and 70% ethanol was added before centrifugation (17,000g, 5 minutes). Supernatant was discarded before centrifugation (17,000g, 3 minutes). Remaining liquid was removed via pipette. Tubes were air dried for 10 minutes, before addition of 30µL Distilled Water (dH_2O), or Tris EDTA Buffer (TE) buffer. gDNA concentration was measured via nanodrop, and frozen at -20°C.

2.1.2 Amplification of *IFITM3* Exon 1

Amplification of *IFITM3* Exon 1 was performed with the Advantage 2 Polymerase Chain Reaction (PCR) kit (Clontech). Reaction was performed as in table 2.1

Table 2.1 PCR Reagents and Cycling Conditions for Amplification of IFITM3 Exon 1

(a) PCR reagents

Reagent	**Quantity**
10X Advantage 2 PCR Buffer	2µl
10mM dNTP mix	0.4µl
Ex1_Seq_F primer (10µM)	1.6µl
Ex1_Seq_R primer (10µM)	1.6µl
50X Advantage 2 Polymerase Mix	0.4µl
gDNA template	25ng
Nuclease-free water	to 20µl total volume

(b) Cycling parameters

Step	**Temperature** (°C)	**Time**[1]
Denaturation	95	1:00
35 cycles of:		
Denaturation	95	1:00
Annealing	64	0:30
Extension	68	1:00
Final Extension	68	2:00

[1] minutes:seconds

Resultant PCR products were run on a 1% Tris-Acetate-EDTA (TAE) agarose gel for 40 minutes at 60V. PCR product is 337bp in length.

2.1.3 Purification of PCR Products

PCR products were purified with the QIAquick PCR Purification kit (QIAGEN) according to manufacturer's instructions. PCR product was eluted into 30-50µL of dH_2O or TE. Concentration of DNA was determined by nanodrop, and product was frozen at -20°C.

2.1.4 DNA Gel Electrophoresis

Molecular Grade Agarose (Bioline) was used for gel electrophoresis. 1% agarose solution was made with TE buffer, then microwaved until agarose was completely dissolved. Ethidium Bromide was added to agarose ($1\mu g\,ml^{-1}$). Agarose was poured into gel mould. After setting, gel was immersed in 1X TAE buffer in a gel tank (Mini-Sub Cell GT Cell, Biorad).

DNA products were mixed in a 5:1 ratio with 6X DNA gel loading dye (Thermo Fisher Scientific). Ladder and DNA products were added to wells. After running, bands were visualised with a Gel Doc XR+ imaging system (Bio-Rad).

2.1.5 Gel Extraction of PCR Products

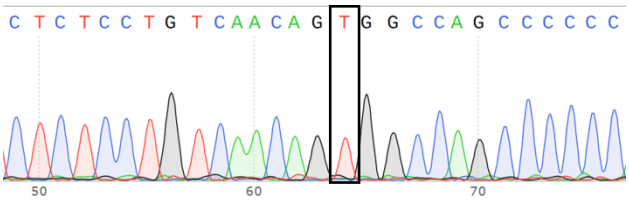
Electrophoresis gel was run as in section 2.1.4. Gel extraction was performed with the QIAquick Gel Extraction Kit (QIAGEN) according to manufacturer's instructions. DNA was eluted in 20µL of TE Buffer.

2.1.6 Genotyping of rs12252

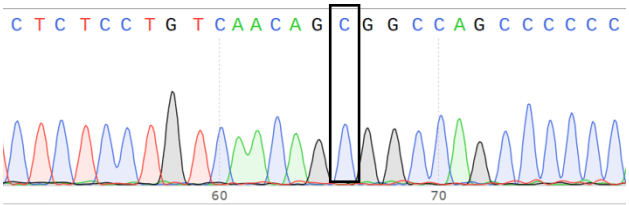
IFITM3 exon 1 PCR was carried out (section 2.1.2). PCR products were purified (section 2.1.3). Samples were sequenced with Ex1_Seq_F and Ex1_Seq_R primers. Both returned sequences were aligned with reference sequences, and cross checked to determine rs12252 genotype. Heterozygous samples appear with overlapping T and C peaks at the rs12252 position (fig. 2.1).

2.1.7 TOPO-TA Cloning

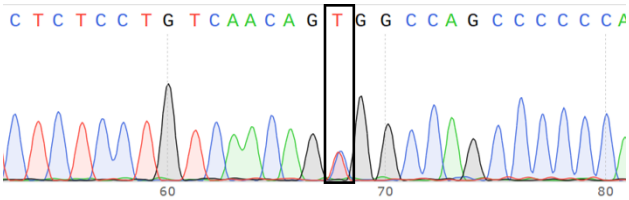
TA cloning was performed with the TOPO TA Cloning Kit (Thermo Fisher Scientific) with PCR product of choice according to manufacturers instructions. PCR products were purified beforehand via gel extraction (section 2.1.5)



(a) TT Genotype



(b) CC Genotype



(c) CT Genotype

Figure 2.1: Genotyping of rs12252 using sequence traces for exon 1 of IFITM3. The rs12252 SNP is shown with a black box.

2.1.8 Small Scale Plasmid Preparation (Miniprep)

Miniprepping was performed with the QIAprep Spin Miniprep Kit (QIAGEN). Single *Escherichia coli* (*E. coli*) colonies were picked and added to 1-2mL of Lysogeny Broth (LB) media with appropriate selective antibiotic. Cells were incubated at 37°C for 16-24 hours. Plasmid purification was performed as per manufacturer's instructions. Plasmid DNA was eluted in 30µL TE. DNA concentration was determined by nanodrop.

2.1.9 Large Scale Plasmid Preparation (Maxiprep)

Maxiprepping was performed with the HiSpeed Plasmid Maxi Kit (QIAGEN). Singel *E. coli* colonies were picked and added to 1mL of LB medium with appropriate selective antibiotic before subsequent addition to 200mL of LB with appropriate antibiotic. Bacteria were allowed to grow for 16-24 hours. Plasmid purification was performed as per manufacturer's instructions. Plasmid DNA was eluted in 1mL of TE. DNA concentration was determined by nanodrop.

2.1.10 RNA extraction via RNeasy PLUS Mini Kit

β-Mercaptoethanol (βME) was added to RLT plus buffer with a ratio of 10µL per mL. Cells were washed before addition of 350µl RLT plus buffer. Cell lysate was added directly to QIAshredder columns (QIAGEN).

QIAshredder columns were centrifuged (17,000g, 2 minutes). Flow-through was frozen at -80°C, or used immediately for RNA extraction. RNeasy PLUS mini kit (QIAGEN) was used following manufacturer's protocol. RNA was eluted into 30µL dH_2O. Concentration of RNA was determined by nanodrop. RNA samples were stored at -80°C.

2.1.11 Preparation of cDNA from RNA using High-Capacity RNA-to-cDNA Kit

Conversion of RNA to cDNA was performed wit the High-Capacity RNA-to-cDNA kit (Thermo Fisher Scientific) according to manufacturer's

instructions. A consistent quantity of RNA was used across samples to ensure reverse-trascriptase efficiency was consistent across samples. Retured cDNA was diluted with dH_2O in a ratio of 1:4.

2.1.12 Quantification of mRNA Transcripts via qPCR

qPCR mix was prepared with TaqMan Gene Expression Master Mix (Thermo Fisher Scientific) and TaqMan probes (Thermo Fisher Scientific). In each reaction, 5µL of mastermix was mixed with 0.5µL of probe, and 4.5µL of cDNA mix. Volumes were halved in cases where a 384 well qPCR plate was used, as opposed to 96 well plates. *Glyceraldehyde 3-Phosphate Dehydrogenase (GAPDH)* was used as an internal control gene. Plates were run on a QuantStudio 7 Flex Real-Time PCR System (Thermo Fisher Scientific) according to standard protocols.

2.1.13 Western Blotting For IFITM3

Adherent cells were collected and counted (section 2.1.15). $5x10^5$ cells were seeded in each well of a 6 well plate. 24 hours later, cells were harvested into 1.5mL microcentrifuge tubes and counted. Cells were centrifuged (865g, 2 minutes). Cells were resuspended in RIPA lysis buffer (Thermo Fisher Scientific) with 10µL mL^{-1} protease inhibitor cocktail (Sigma-Aldrich). Cells were vortexed for 1 minute, and left to incubate on ice for 30 minutes. Lysate was centrifuged (17,000g, 10 minutes). Supernatant was combined with 4x Laemmli Sample Buffer (Bio-Rad) at a ratio of 3:1 before incubation at 90°C for 5 minutes.

Samples were loaded in a 16% Novex Tris-Glycine Protein Gel (Thermo Fisher Scientific) in a XCell SureLock Mini-Cell Gel Tank (Thermo Fisher Scientific). Samples were run in 1X Novex Tris-Glycine SDS Running Buffer (Thermo Fisher Scientific) alongside a Precision Plus Protein duo ladder (Bio-Rad). Gel was stacked at 60V for 15 minutes, before running for 90 minutes at 120V.

Transfer sandwich was prepared with sponges, gel, Western Blotting Filter Paper (Thermo Fisher Scientific) and activated (methanol, 5 seconds) Immobilon-FL Membranes (Merck). Transfer module was placed in gel tank with 1X Novex Tris-Glycine Transfer Buffer (Thermo Fisher Scientific).

Samples were run at 26V for 75 minutes.

5% skim milk in Tris-Buffered Saline with Tween 20 (TBST) (Thermo Fisher Scientific) was prepared. After transfer, membrane was removed, and blocked in skim milk for 1 hour on shaker. Membrane was transferred to primary antibody incubation mixture with $3\mu g\,ml^{-1}$ anti-IFITM3 Ab (Non-Commercial, Mouse Anti-Human) and 1/500 anti-GAPDH Ab (Clone 6C5, Mouse Anti-Human, Merck). Membrane was incubated overnight on shaker at 4°C. Membrane was washed three times in TBST. Membrane was incubated with IRDye 680LT Goat Anti-Mouse fluorescent antibody (Li-Cor) for 1 hour at RT. Membrane was washed in TBST three times. Membranes were visualised using the Li-Cor Odyssey Imaging System (Li-Cor).

2.1.14 Culturing of HEK293 and A549 Cells

Human Embryonic Kidney Cell 293 (HEK293) (Gifted from Dr. Jan Rehwinkel) and A549 (Gifted from Dr. Lili Wang) cell lines were grown in cell culture flasks in D10 medium. Cells were grown until 70-90% confluency before collection and splitting. Cells were incubated at 37°C and 5% CO_2.

2.1.15 Collection and Counting of Cells in Culture

Media was aspirated from cells. Cells were washed once with DPBS. Cells were incubated for 2-5 minutes in a solution of Trypsin-EDTA. An equal volume of D10 medium was added to the cell culture dish, and cells were collected in a centrifuge tube. Cells were centrifuged (423g, 5 minutes). Cells were resuspended in fresh D10 medium. 10µL cell solution was combined with 10µL 0.4% trypan blue solution (Thermo Fisher Scientific). Resultant solution was used for counting via a Countess II FL Automated Cell Counter (Thermo Fisher Scientific)

In the case of suspended cells, cells were agitated to ensure even distribution of cells. Cells were counted directly with Countess system.

2.2 Methods and Materials for Chapter 4

2.2.1 Culturing Primary Human Fibroblasts

Primary human fibroblasts were obtained via surgical skin waste from anonymous donors. Study and collection of these cells was approved by the local ethics committee of Beijing You-An Hospital, and the University of Oxford Tropical Ethics Committee, under OXTREC 1001-13. Skin was cut into 0.2 - 0.5cm pieces with a scalpel. Cell culture flask was partially filled with AminoMAX C-100 Basal Medium (Thermo Fisher Scientific)with added AminoMAX C-100 Supplement (Thermo Fisher Scientific). Skin waste pieces were left partially covered by medium. Flasks were left for 2-3 days to allow fibroblasts to grow. Skin pieces were removed, and cells were incubated for 3-5 minutes with Trypsin-EDTA 0.05% (Thermo Fisher Scientific) solution. Trypsin-EDTA was neutralised via added D10 medium, and cell solution was centrifuged (423g, 5 minutes). Cell pellet was resuspended with fresh AmnioMAX medium.

2.2.2 Preparation of cDNA via Superscript III

cDNA was prepared with the Superscript III First Strand Synthesis System (Thermo Fisher Scientific) according to manufacturer's instructions. RNA quantity and concentration was made consistent across samples. Resultant cDNA was frozen at -20°C.

2.2.3 Targeted RNA Sequencing of rs12252

Targeted RNA sequencing of rs12252 was performed via a modified version of the Illumina 16S Metagenomics Sequencing Library. cDNA was prepared (section 2.2.2) and purified (section 2.1.3). Concentration was determined via nanodrop. Library preparation was performed in two PCR steps:

Amplicon PCR

The first step, amplicon PCR, used custom amplification primers designed to amplify the rs12252 locus. Each primer contained overhang adapters for later use. KAPA HiFi HotStart ReadyMix (KAPA Biosystems) was used for amplicon PCR. Reactions were prepared in a 96 well plate format as in table 2.2

Table 2.2 PCR reagents and Cycling Conditions for Amplicon PCR (Reaction 1 of Targetted rs12252 Sequencing)

(a) PCR reagents

Reagent	**Quantity**
2X HiFi HotStart ReadyMix	12.5µl
Ampl_Seq_F Primer (1µM)	5µl
Ampl_Seq_R primer (1µM)	5µl
cDNA template	12.5µg
10mM Tris pH 8.5	to 25µl total volume

(b) Cycling parameters

Step	**Temperature (°C)**	**Time**[1]
Denaturation	95	3:00
25-28 cycles of:		
Denaturation	95	0:30
Annealing	55	0:30
Extension	72	1:00
Final Extension	72	5:00

[1] minutes:seconds

Resultant products were used immediately in PCR Clean up 1.

PCR Clean Up 1

PCR Cleanup was performed with Agencourt AMPure XP Beads (Beckman Coulter). Beads were brought to RT and vortexed for 30 seconds. 20µL of beads were added to each sample and mixed. Beads were incubated at RT for 5 minutes. Plate was placed on a Dynamag 96-Side Skirted Magnet (Thermo Fisher Scientific) for two minutes. Supernatant was discarded. Beads were

washed with 200µL of 80% ethanol for 30 seconds. Supernatant was removed, and ethanol wash was repeated. Supernatant was removed, and beads were allowed to air dry on for 10 minutes. Plate was removed from magnet. 52.5µL of 10mM Tris pH 8.5 was added to each well before incubation at RT for 2 minutes. Plate was placed back on the magnet for 2 minutes. 50µL of the supernatant was transferred to a new plate.

Samples were stored at -20°C or used immediately in Indexing PCR

Index PCR

A second PCR step was performed to add indices from the Nextera XT Index Kit (Illumina) to amplicons. The indexing PCR was performed with KAPA HiFi HotStart ReadyMix as in table 2.3

Table 2.3 PCR reagents and Cycling Conditions for Index PCR (Reaction 2 of Targetted rs12252 Sequencing)

(a) PCR reagents

Reagent	**Quantity**
2X HiFi HotStart ReadyMix	25µl
Nextera XT Index 1 Primer	5µl
Nextera XT Index 2 Primer	5µl
Amplicon PCR Sample Template	5µg
Nuclease Free Water	10µL

(b) Cycling parameters

Step	**Temperature (°C)**	**Time**[1]
Denaturation	95	3:00
8 cycles of:		
Denaturation	95	0:30
Annealing	55	0:30
Extension	72	1:00
Final Extension	72	5:00

[1] minutes:seconds

Resultant PCR samples were used immediately in PCR Clean Up 2.

PCR Clean Up 2

The second round of PCR Cleanup was again performed with the use of Agencourt AMPure XP Beads. 56µL of beads were added to each well of the 96 well plate. Clean up was performed as in clean up 1. Samples were eluted in 25µL of TE buffer in this case.

Samples were stored at -20°C.

Quantification and Quality Control of Library

Concentration of samples was determined via Qbit dsDNA HS Assay Kit (Thermo Fisher Scientific) as per manufacturer's instructions. DNA quality and average band size of samples was determined with a High Sensitivity DNA Kit (Agilent) on a Bioanalyzer 2100 Instrument (Agilent). Average band size and concentration was used to pool all samples equally into a 4nM library.

Sample Denaturation and Sequencing

Sample denaturation was performed with a MiSeq Reagent Kit v2 for 500 cycles (Illumina). A PhiX Control library (Illumnia) was combined with the prepared library. Denaturation and PhiX library combination was performed as per manufacturer's instructions. Samples were sequenced on a MiSeq platform at the WIMM Sequencing Facility in a paired-end fashion, with 250bp reads. Analysis was performed as in section 2.2.4.

2.2.4 Analysis of rs12252 Allelic Expression

In cases where we performed targetted RNA Sequencing, quality control and alignment of samples was performed on the Illumina BaseSpace Platform. FASTQ Toolkit app (Illumina) to filter out reads with low quality scores, with a Phred threshold of 20. We then used the BWA aligner tool [263] for alignment and indexing.

When samples were derived from larger RNA-Sequencing experiments, we used aligned samples as explained in section 2.5.

Alignment files were exported to Integrated Genomics Viewer (IGV) [264], with the reference genome loaded (GRCh38). The number of reads of each nucleotide that aligned to the rs12252 locus were obtained. We did not consider rare instances of inappropriate alleles. From raw alignment counts we calculated percentage of T and C alignments.

2.2.5 DNA Analysis via Bioanalyser 2100 System

DNA analysis was performed with a Aglient High Sensitivity DNA Kit (Aglient) on a Bioconductor 2100 System. Procedure was carried out as per manufacturer's instructions.

2.2.6 Extraction of RNA via TRIzol

TRIzol RNA extraction was performed with adherent cells. Cell culture media was removed from cells. TRIzol reagent (Thermo Fisher Scientific) was added to cells and left to incubate at RT for 5 minutes. Cell lysate was collected.

Chloroform was added to TRIzol lysate in a ratio of 1:5 and mixed vigorously. Mixture was incubated for 2-3 minutes. Solution was centrifuged (13000g, 15 minutes, 4°C). Aqueous phase was transferred to new tube. A volume of isopropanol equivalent to half of the initial volume of TRIzol was added to aqueous phase and mixed well. Solution was centrifuged (13000g, 10 minutes, 4°C). Supernatant was discarded and RNA pellet was washed with 75% ethanol. Solution was centrifuged (13000g, 5 minutes, 4°C). Supernatant was discarded and tubes were centrifuged again briefly. A pipette was used to remove remaining ethanol. RNA pellet was resuspended in dH_2O in 20-100µL dH_2O. RNA concentration was determined by nanodrop.

2.2.7 RNA Purification via RNeasy MinElute Cleanup Kit

RNA was purified via the RNeasy MinElute Cleanup Kit (Qiagen) according to manufacturer's instructions. 20µL of RNase free water was used to elute

RNA. Concentration of eluted RNA was determined by nanodrop. RNA was stored at -80°C.

2.2.8 Nascent RNA Pulldown via 4SU Treatment

Cells were grown in 15cm dishes until 70-80% confluency. 4SU (Sigma Aldrich) was added to media at 300μM, and incubated for 1 hour. TRIzol RNA extraction was performed (section 2.2.6). Biotin-HPDP (Thermo Fisher Scientific) was dissolved in Dimethylformamide (DMF) (Sigma Aldrich) at 1mg mL^{-1} Biotin labelling reaction was made, with 2µL Biotin-HPDP, and 1µL 10x Biotinylation Buffer per 1µg of RNA. Reaction volume was made up with dH_2O to 10µLs per 1µg of RNA. Biotinylation reaction was incubated at RT for 1.5 hours.

An equal volume of Chloroform/Isoamylalcohol (24:1) was added to reaction mixture and mixed vigorously. Solution was incubated at RT for 2-3 minutes until phases were separated. Mixture was transferred to 2mL MaXtract High Density Tubes (QIAGEN) and centrifuged (17000g, 10 minutes). Upper phase was transferred into new microcentrifuge tubes. 5M NaCl was added to tubes in ratio of 1:10. An equal volume of isopropanol was added and tubes were centrifuged (17000g, 25 minutes). RNA was washed with of 75% ethanol, and centrifuged (17000g, 15 minutes). RNA was resuspended in TE Buffer.

RNA was heated to 65°C for 10 minutes, and subsequently transferred to ice. Separation of labelled/biotinylated RNA from unlabelled RNA was performed with the μMACS Streptavidin Kit (Miltenyi Biotech) using a μMACS Separator (Miltenyi Biotech) according to manufacturers instructions. Columns were pre-equilibrated with 0.9mL biotin washing buffer. RNA was eluted with 100µL Dithiothreitol (DTT) into 700µL of RLT Buffer, from the RNeasy MinElute Cleanup Kit (QIAGEN). Elution was performed again into the same tube.

RNA was cleaned with the RNeasy MinElute Cleanup Kit, as in section 2.2.7.

2.2.9 Nascent RNA pulldown via Click-iT Kit

Adherent cells were grown in 6 well plates until 70-80% confluency. 0.5mM of 5-Ethynyl Uridine (EU) was added to each well and incubated for 1 hour. TRIzol extraction of RNA was performed as in section 2.2.6. 3-4µg of RNA was used in each reaction. Nascent RNA pulldown and cDNA preparation was performed with the Click-iT Nascent RNA capture Kit (Thermo Fisher Scientific) as per manufacturer's instructions. cDNA was stored at -20°C.

2.2.10 Bisulfite Conversion of gDNA

gDNA was bisulfite treated with the EZ DNA Methylation-Gold Kit (Zymo Research) as per manufacturer's instructions. 500ng of gDNA was used per reaction. Bisulfite converted DNA was eluted in 10µL of M-Elution Buffer. DNA was frozen at -20°C.

2.2.11 Amplification of Bisulfite Treated rs12252 Locus

Amplification of the rs12252 locus using bisulfite-treated gDNA was performed with the Advantage 2 PCR Kit (Clontech). Template DNA was prepared as in section 2.2.10. Reaction was performed as in table 2.4
PCR products were run on a 1% TAE gel and extracted (section 2.1.5).

2.2.12 Detection of Methylation State of rs12252

gDNA was extracted from cells of interest. Bisulfite conversion was performed on gDNA (section 2.2.10). The rs12252 locus was amplified (section 2.2.11). The PCR product was cloned into the pCR4-TOPO TA Vector (Thermo Fisher Scientific) (section 2.1.7) and miniprepped (section 2.1.8). Resultant plasmids were sent for Sanger sequencing using the M13_F and M13_R primers.

Returned electropherograms were checked for clear signal and lack of background. Clear electropherograms were aligned to the predicted bisulfite converted sequence of exon 1 of IFITM3. Forward and reverse sequences

Table 2.4 PCR Reagents and Cycling Conditions for Amplification of rs12252 locus from Bisulfite Treated gDNA

(a) PCR reagents

Reagent	**Quantity**
10X Advantage 2 PCR Buffer	2µl
10mM dNTP mix	0.4µl
Meth_Seq_F Primer (10µM)	1.6µl
Meth_Seq_R primer (10µM)	1.6µl
50X Advantage 2 Polymerase Mix	0.4µl
Bisulfite-Treated gDNA template	3µL
Nuclease-free water	to 20µl total volume

(b) Cycling parameters

Step	**Temperature (°C)**	**Time**[1]
Denaturation	95	3:00
30 cycles of:		
Denaturation	95	1:00
Annealing	64	0:30
Extension	68	1:00
Final Extension(1)	68	1:00
Final Extension(2)	72	10:00

[1] minutes:seconds

from a single sample were cross-checked. The base shown at the rs12252 locus was used to determine the methylation state.

2.2.13 Cloning sgRNA into px458 Plasmid Backbone

The px458 vector was used for all Single Guide RNA (sgRNA) cloning. sgRNAs were ordered as two short complementary Single Stranded DNA (ssDNA) Oligonucleotides (oligos). The forward oligo was designed with an additional CCAC sequence , and the reverse was designed with an additional AAAC sequence, both at the 5' ends. These overhangs were used as sticky ends for cloning.

Oligos were made to 100µM stock solutions in dH_2O. Double stranded oligos were prepared in NEB Buffer 2.1 (New England Biolabs) as in table 2.5. Solution was incubated at 99°C for 5 minutes, and allowed to cool slowly.

Oligos were precipitated by adding 30µL 3M NaOAC pH 5.6 and 300µL 100% ethanol before centrifugation (17,000g, 30 minutes, 2°C). Oligos were washed with 70% ethanol. Solution was centrifuged (17,000g, 10 minutes). Supernatant was discarded, and tubes were air-dried for 10 minutes. Oligo concentration was determined by nanodrop.

Table 2.5 Reagents Used for Cloning of sgRNA Construct into px458 Vector

(a) Preparation of Double Stranded Oligos

Reagent	**Quantity**
Forward Oligo (100µM)	10µl
Reverse Oligo (100µM)	10µl
10X NEB Buffer 2.1	10µl
dH_2O	70µL

(b) Phosphorylating Double Stranded Oligos

Reagent	**Quantity**
10X NEB Ligation Buffer	2µl
T4 PNK	1µl
Double Stranded Oligonucleotide	1µg
dH_2O	to 20µL

(c) Ligation of Double Stranded Oligo and px458 Vector

Reagent	**Quantity**
10X NEB Ligation Buffer	1µl
T4 Ligase	1µl
Double Stranded Oligonucleotide	25ng
Pre-Cut px458 Vector	20ng
dH_2O	to 10µL

1µg of oligo was phosphorylated with NEB T4 Polynucleotide 5'-Hydroxyl-Kinase (PNK) (New England Biolabs) in T4 Ligase Buffer (New England Biolabs) as in table 2.5. Reaction was incubated at 37°C for 1 hour.

25ng of oligo was ligated with px458 plasmid vector cut with BbsI enzyme, as shown in table 2.5. Reaction was incubated at RT for 1 hour.

2.2.14 Transfection of HEK293 Cells

HEK293 cells were transfected with Lipofectamine 2000 (ThermoFisher Scientific). Cells were counted and seeded at of 3.5×10^5 cells per well in a 6 well plate. Cells were supplemented with 2mL of antibiotic-free D10 medium. The following day, Lipofectamine transfection cocktail was prepared as in table 2.6. When Homology Directed Repair (HDR) was desired, a Single Stranded Donor Oligonucleotide (ssODN) was co-transfected with plasmids.

Table 2.6 Reagent Cocktail for Transfection of HEK293 Cells

(a) Reagent Cocktail

Reagent	**Quantity**
Total Plasmid DNA	2.5µg
ssODN Template (100µM) (Optional)	1µL
Lipofectamine 2000 Reagent	5µL
Opti-MEM Low Serum Medium	to 500µL

Lipofectamine 2000 was allowed to incubate for 5 minutes in Opti-MEM Low Serum Medium (Thermo Fisher Scientific) at RT. DNA was combined with Lipofectamine 2000 and incubated for 20 minutes at RT. Cocktail was added to cells drop by drop. Cells were left to rest for 48-72 hours.

When HDR activity was desired, we used two HDR enhancing drugs: SCR7 and later, RS1. 0.1µM SCR7 or 7.5µM RS1 was added [265, 266] to the cell mixture.

2.2.15 Transfection of A549 Cells

A549 cells were transfected with Lipofectamine LTX (ThermoFisher Scientific). A549 cells were collected and counted. Cells were seeded at $3x10^5$ cells per well in a 6 well plate. Cells were supplemented with 2mL of antibiotic-free D10 medium. The following day, Lipofectamine transfection cocktail was prepared as in table 2.7.

DNA was incubated with PLUS reagent in Opti-MEM Low Serum Medium (Thermo Fisher Scientific) at RT for 10 minutes. Lipofectamine LTX was incubated for 5 minutes in Opti-MEM Low Serum Medium for 5 minutes

Table 2.7 Reagent Cocktail for Transfection of A549 Cells

(a) Reagent Cocktail

Reagent	**Quantity**
Total Plasmid DNA	2.5µg
Lipofectamine LTX Reagent	5µL
PLUS Reagent	1.5µL
Opti-MEM Low Serum Medium	to 500µL

at RT. The aforementioned mixtures were combined and incubated for 20 minutes at RT. Cocktail was added to cells drop by drop. Cells were left to rest for 48-72 hours.

2.2.16 Single Cell Cloning of Transfected Cells

Cell culture media was collected from growing cells as conditioned media. Transfected cells were collected into Fluorescence Activated Cell Sorting (FACS) tubes. Flat bottom 96 well cell culture plates were filled with 100µL of a 1:1 mixture of conditioned media and fresh D10. Collected cells were sorted based on transfected fluorescent marker, with one cell per well. Gating strategy is shown in fig. 2.2. Wells were topped up every 3-5 days with fresh D10. After 3-4 weeks, clones were collected and analysed. Roughly 10-30% of wells had productive growth.

Single cell sorting was done with the help of Dr. Kevin Clark and Dr. Craig Waugh.

2.2.17 Single Cell Dilution of HEK293 Cells

HEK293 cells were collected and counted. A cell solution of 5 cellsmL^{-1} was made up in fresh D10 media. 200µL of cell solution was added to each well of a flat bottom, 96 well plate. Cells were grown for 2-3 weeks. Resulting cell clones were collected.

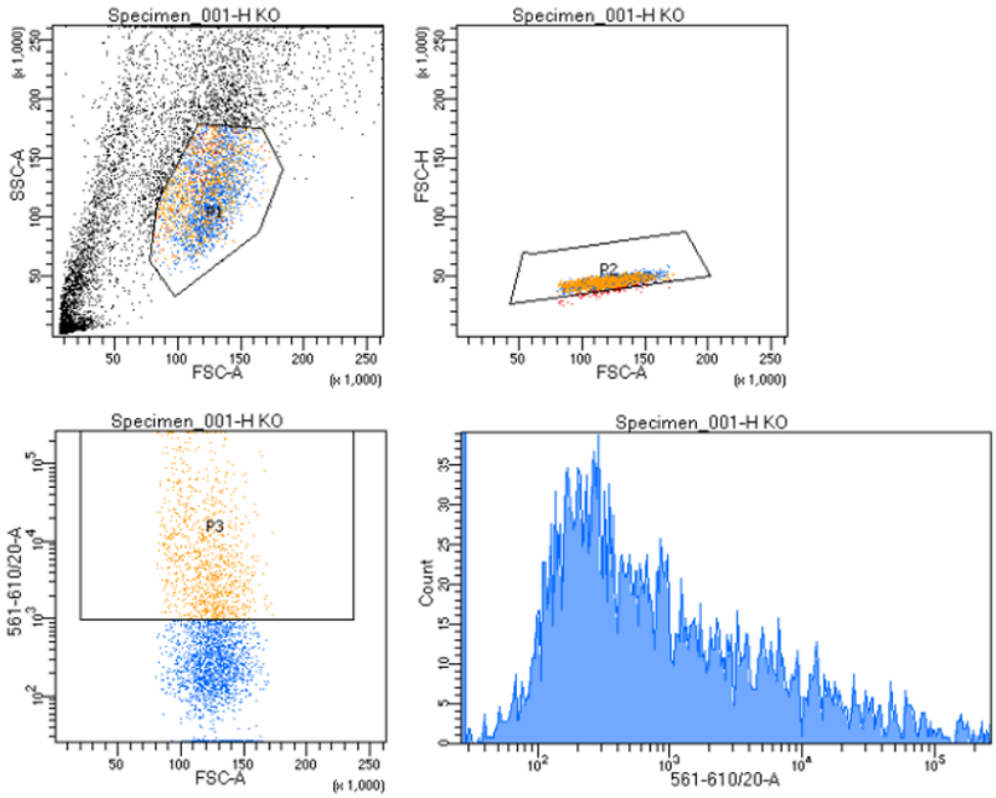


Figure 2.2: Single cell sorting gating strategy for HEK293 cell lines. Cells were gated on FSC-A/SSC-A (top left), before selection of singlets via FSC-A/FSC-H. Singlets were gated on positive signal for fluorescent marker. Transfected cells were gated by presence of fluorescent marker (mRuby2 or GFP, detected via Texas red or FITC channels). Gating of positive cells was done relative to an uninfected control population of cells to account for autofluorescence Gating strategy was the same for A549 cells.

2.2.18 Screening for *IFITM3* Knock-Out Cell Lines

Cell clones were grown until at least 2 million cells were present. For each clone, two wells of a six well tissue culture dish were seeded with 0.5 million cells. Cells were left for 24 hours; one well was left untreated, while the other was treated with 500U/ml Interferon-α2 (IFN-α2). Western blotting was used to determine the KO status of each cell line(section 2.1.13).

2.2.19 Amplification of *IFITM3* Exon 2

Exon 2 of IFITM3 was amplified with the Phusion High-Fidelity DNA Polymerase (New England Biolabs). Reaction was performed as in table 2.8

Table 2.8 PCR Reagents and Cycling Conditions for Amplification of IFITM2 Exon 1

(a) PCR reagents

Reagent	**Quantity**
5X Phusion HF Buffer	10µl
10mM dNTP mix	1µl
IFITM3_E2_F primer (10µM)	2.5µl
IFITM3_E2_R primer (10µM)	2.5µl
DMSO	1.5µL
50X Phusion HF Polymerase	1µl
gDNA template	50ng
Nuclease-free water	to 50µl total volume

(b) Cycling parameters

Step	**Temperature (°C)**	**Time**[1]
Denaturation	98	0:30
35 cycles of:		
Denaturation	98	0:10
Annealing	69	0:30
Extension	72	0:30
Final Extension	72	2:00

[1] minutes:seconds

2.2.20 Verification of *IFITM3* Knock-Out Status

gDNA was extracted from HEK293 KO clones (section 2.1.1). Two PCR reactions were performed, amplifying over both regions of sgRNA specificity (section 2.1.2 and section 2.2.19). Samples were purified as in section 2.1.3, and submitted for sequencing alongside reactions with the WT parental cell line. The TIDE webtool [267] was used to determine the nature of the edited cell lines. Cells with Insertions or Deletions (indels) of base numbers not divisible by 3 were saved as KO cell lines.

2.2.21 Screening of rs12252 for Site Directed Mutagenesis

Screening of *IFITM3* for HDR at the rs12252 locus was performed via amplification of the rs12252 locus, followed by treatment with EcoRV-HF. PCR reaction was performed with Phusion High-Fidelity DNA Polymerase (New England Biolabs) as in table 2.9

PCR product was purified, as in section 2.1.3. Cleaned product was used for restriction digest with EcoRV-HF (New England Biolabs). Restriction digest reaction was established as follows:

Resultant product was run on a DNA gel (section 2.1.4). Gel images were assessed for the presence of EcoRV cutting.

2.2.22 Revised Screening of rs12252 for Site Directed Mutagenesis

Revised screening of rs12252 knock-in cells was performed via direct sequencing of rs12252. Cell clones in 96 well plates were washed once with DPBS. Cells were lysed by incubation at RT for 20 minutes in 30µl Terra Direct Buffer from the Terra Direct Genotyping Kit (Takara). This lysate was used as a template amplification of the rs12252 locus. was performed as in table 2.10

Amplicons were purified as in section 2.1.3. Nanodrop was used to determine the concentration of PCR products before sequencing with both forward and

Table 2.9 PCR Reagents and Cycling Conditions for Amplification of IFITM2 Exon 1

(a) PCR reagents

Reagent	**Quantity**
5X Phusion HF Buffer	10µl
10mM dNTP mix	1µl
IFITM3_HDR_F primer (10µM)	2.5µl
IFITM3_HDR_R primer (10µM)	2.5µl
DMSO	1.5µL
50X Phusion HF Polymerase	1µl
gDNA template	50ng
Nuclease-free water	to 50µl total volume

(b) Cycling parameters

Step	**Temperature** (°C)	**Time**[1]
Denaturation	98	0:30
35 cycles of:		
Denaturation	98	0:10
Annealing	68	0:30
Extension	72	0:30
Final Extension	72	2:00

[1] minutes:seconds

Reagent	**Quantity**
10X CutSmart HF Buffer	1µl
EcoRV-HF	1 unit
gDNA template	100ng
Nuclease-free water	to 10µl total volume

reverse primers at the WIMM sequencing facility. rs12252 genoytype was determined by examination of individual electropherograms.

2.2.23 Amplification of *IFITM2* Exon 1

Exon 1 of IFITM2 was amplified with Phusion High-Fidelity DNA Polymerase (New England BioLabs). Reaction was performed as in table 2.11

Table 2.10 PCR Reagents and Cycling Conditions screening CRISPR Clones

(a) PCR reagents

Reagent	**Quantity**
IFITM2_E1_F primer (10µM)	1.5µl
IFITM2_E1_R primer (10µM)	1.5µl
50X Terra PCR Direct Polymerase Mix	1µl
2X Terra PCR Direct Buffer with Lysate	25µl
Nuclease-free water	to 50µl total volume

(b) Cycling parameters

Step	**Temperature** (°C)	**Time**[1]
Denaturation	98	2:00
35 cycles of:		
Denaturation	98	0:10
Annealing	64	0:15
Extension	68	1:00
Final Extension	68	2:00

[1] minutes:seconds

2.2.24 Verification of *IFITM2* Gene Integrity

Amplification of Exon 1 of *IFITM2* was performed as in section 2.2.23. Amplicon was purified as in section 2.1.3 and sent for Sanger sequencing with both IFITM2_E1_F and IFITM2_E1_R primers at the WIMM sequencing facility. Returned sequences were aligned to the *IFITM2* reference sequence. Sequences with a clear signal and a one-to-one alignment with the reference sequence were judged to have retained *IFITM2* integrity.

2.2.25 Extraction of RNA from Whole Blood

Whole blood was collected in Tempus Blood RNA tubes (Thermo Fisher Scientific). Samples were frozen at -80°C. RNA extraction from whole blood was performed via the Tempus Spin RNA Isolation Kit (Thermo Fisher Scientific) according to manufacturer's instructions. RNA was eluted into 90µL of Elution Buffer. RNA concentration was determined usng a Qbit RNA HS Assay Kit (Thermo Fisher Scientific). RNA samples were stored at -80°C.

Table 2.11 PCR Reagents and Cycling Conditions for Amplification of IFITM2 Exon 1

(a) PCR reagents

Reagent	**Quantity**
5X Phusion HF Buffer	10µl
10mM dNTP mix	1µl
IFITM2_E1_F primer (10µM)	2.5µl
IFITM2_E1_R primer (10µM)	2.5µl
DMSO	1.5µL
50X Phusion HF Polymerase	1µl
gDNA template	25ng
Nuclease-free water	to 50µl total volume

(b) Cycling parameters

Step	**Temperature (°C)**	**Time**[1]
Denaturation	98	0:30
35 cycles of:		
Denaturation	98	0:10
Annealing	66	0:30
Extension	72	0:30
Final Extension	72	2:00

[1] minutes:seconds

2.3 Methods and Materials for Chapter 6

2.3.1 *In-Vitro* Infection of Adherent Cells with Virus

Cells were collected and counted (section 2.1.15). Cells were seeded in a 24 well cell culture plate at 1×10^5 cells per well and left overnight. Media was aspirated and cells were washed once with DPBS. Virus was diluted from stock solution to the appropriate Multiplicity Of Infection (MOI)/dilution factor in DPBS. MOI was determined by manufacturer when available, or titration in Madin-Darby Canine Kidney (MDCK) cells (via Dr. Alain Townsend). 100µL of virus solution was added to cells and left for 1.5 hours. 1mL of fresh D10 medium was added to cells. Cells were incubated for 16-24 hours. This protocol was performed with Sendai Virus (SeV), Influenza A/Puerto Rico 8/1934 (PR8), and S-Flu.

2.3.2 *In-Vitro* Infection of Suspended Cells with Virus

Cells were collected and counted (section 2.1.15). The appropriate number of cells were added to a 15ml centrifuge tube. Cells were centrifuged (423g for 5 minutes). Cells were washed with 1ml DPBS and centrifuged (423g, 5 minutes). Supernatant was removed via pipette, ensuring no liquid remained. 100µl virus solution in DPBS was added to each sample and cells were resuspended, before incubation for 1.5 hours. After, fresh R10 medium was added to cells, and cells were transferred to a 6 well cell culture plate. Cells were allowed to incubate for 8 hours. This protocol was performed with S-Flu.

2.3.3 Separation of PBMCs from Whole Blood

30-50mL of blood was drawn from donors with informed consent. Study and collection of these samples was approved by the local ethics committee of the University of Oxford Tropical Ethics Committee, under OXTREC 1001-13. 16mL of Lymphoprep (STEMCELL Technologies) was added to 50mL LeucoSep tubes (Greiner Bio-One) and centrifuged (751g, 5 minutes). A maximum of 25mL of blood was added to each tube, and centrifuged (751g, 20 minutes) without brake. Lymphocytes were pipetted out, and added to a 50mL tube before topped up to 50mL with Roswell Park Memorial Institute Medium (RPMI). Cells were centrifuged at 4°C three times at speeds of 609, 423 and 270g, each time washing with RPMI. Cells were resuspended in R10, or used for preparation of DCs.

2.3.4 Isolation of Monocytes and Generation of Monocyte-Derived DCs

PBMCs were extracted from whole blood as in section 2.3.3. PBMCs were washed once with 30ml ice-cold FACS buffer (270g, 10 minutes, 4°C). PBMCs were resuspended in 2.75mL FACS buffer, and 0.25mL human Cluster of Differentiation 14 (CD14) microbeads (Miltenyi Biotec) before incubation for 20 minutes at 4°C. Cells were resuspended with 30mL ice-cold FACS buffer and centrifuged (270g, 10 minutes, 4°C). Cells were resuspended in 3mL ice-cold FACS buffer. LS MACS columns (Miltenyi Biotec) were attached to a QuadroMACS Separator (Miltenyi Biotec) on a

MACS Multistand (Miltenyi Biotec). Columns were equilibrated with 3mL FACS buffer. Cells were passed through the column, before washing with 3mL FACS buffer. Column was removed from separator, and 6mL FACS buffer was added to the column. Plunger was placed at the top of the column and depressed, eluting monocytes into a 50mL tube. Cells were resuspended with 30mL ice-cold RPMI, and centrifuged (270g, 10 minutes, 4°C).

Monocytes were resuspended in warm R10, with 500U/ml Interleukin-4 (IL-4) and 800U/ml Granulocyte-Macrophage Colony-Stimulating Factor (GM-CSF). Cells were plated in six well dishes, and left to incubate for 6 days to allow for differentiation into DCs. On day three, fresh IL-4 and GM-CSF was added.

2.3.5 Characterisation of DCs via FACS

We characterised DCs via FACS using Abs for moncytes CD14), DCs (Dendritic -Specific Intercellular Adhesion Molecule-3-Grabbing Non-Integrin (DC-SIGN)), and DC maturity (Cluster of Differentiation 83 (CD83)) (table 2.12).

Table 2.12 DC Characterisation Panel

Target	**Marker**	**Volume** ($\mu l\,ml^{-1}$)	**Ab Clone**
Live/Dead	N/A	1	N/A[1]
CD14	FITC[3]	50	M5E2[2]
DC-SIGN	APC[4]	30	9E9A8[2]
CD83	PE[5]	5	HB15e[2]

[1] LIVE/DEAD Fixable Violet Dead Cell Stain Kit (Thermo Fisher Scientific)
[2] Via Biolegend
[3] Fluorescein Isothiocyanate
[4] Allophycocyanin
[5] Phycoerythrin

In staining, washing was done with 1ml wash solution (DPBS or FACS wash buffer, as indicated) and centrifugation (423g, 5 minutes).

DCs were aspirated from 6 well plates into 15ml centrifuge tubes and centrifuged (423g, 5 minutes). Cells were washed once with DPBS before counting (section 2.1.15). A consistent number of DCs was used across

all samples, ranging from 0.5 to 2 million cells. Cells were added to clean 15ml centrifuge tube, and washed once with FACS wash buffer. Cells were resuspended in 100µl Fixable Live/DEAD solution. Cells were left for 20 minutes at 4°C in the dark before washing twice with FACS wash buffer. 100µl of Ab cocktail, as in table 2.12, was added to each sample before incubation at 4°C for 20 minutes in the dark. Cells were again washed twice with FACS wash buffer. 1X CellFIX solution (BD Biosciences) was prepared by adding dH_2O to 10X solution. Cells were resuspended in 1x CellFIX, and stored at 4°C in the dark.

Cells were analysed via FACS as in section 2.3.7. Gating strategy is indicated in section 6.2.2.

2.3.6 Examination of S-Flu Infection via FACS

When working with adherent cells, cells were collected (section 2.1.15) in 1.5ml microcentrifuge tubes. Cells were washed once with DPBS, and centrifuged (423g, 5 minutes). Cells were resuspended in 500µl DPBS, and used for FACS analysis as in section 2.3.7.

When working with suspended cells, cells were collected in round bottom polystyrene tubes, and centrifuged (423g, 5 minutes). Cells were washed once with DPBS, and centrifuged (423g, 5 minutes). Cells were resuspended in 500µl DPBS, and used for FACS analysis.

Infection was investigated via Enhanced Green Fluorescent Protein (eGFP) signal associated with S-Flu. Uninfected cells were used to discriminate between cell autofluorescence and S-Flu fluorescence.

2.3.7 Flow Cytometry

Flow Cytometry was performed on an Attune NxT Flow Cytometer (Thermo Fisher Scientific). Cells were run in 1.5ml microcentrifuge tubes (Eppendorf), or round bottom polystyrene tubes (Corning). Consistency between experiments was ensured by daily laser calibration via Attune Performance Tracking Beads (Thermo Fisher Scientific). The same machine was used for all replicates. Attune machines were maintained by the flow cytometry facility team. Compensation was not performed; no experiments

required the use of the same laser for multiple fluorescent markers. Data was analysed in FlowJo 10.

2.3.8 Determination of IP10 Concentration using ELISA

If cells were grown in an adherent monolayer, supernatant was collected directly. When cells were grown in suspension, cells were centrifuged (423g, 5 minutes) before collecting supernatant. Supernatant was stored at -80°C.

Concentration of Interferon γ-Induced Protein 10 (IP10) was determined via the Human CXCL10/IP-10 DuoSet ELISA kit (R&D Systems). ELISA was performed as per manufacturer's instructions. After manufacturer's protocol, 50µl TMB Chromogen Solution (Thermo Fisher Scientific) per well was added as a substrate for Horseradish Peroxidase (HRP)-conjugated Abs. Plates were incubated in the dark for 15-30 minutes. Reaction was stopped with 0.5M Sulfuric Acid. ELISA plate was read on an iMark Microplate Reader at 450nm.

Standards were used to generate a standard curve via four parameter logistic regression in Graphpad Prism 7. Standards were used to infer sample IP10 concentration.

2.3.9 Determination of Type-I IFN Concentration via 3C11 Bioassay

Assay was developed and shared by the Jan Rehwinkel group. This assay uses HEK293 cells transduced with a lentiviral construct expressing eGFP and firefly luciferase under the control of an ISRE (3C11 Cells). 3C11 cells respond to type-I IFN induction via expression of eGFP and firefly luciferase.

If cells were grown in an adherent monolayer, supernatant was collected directly. When cells were grown in suspension, cells were centrifuged (423g, 5 minutes) before collecting supernatant. Supernatant was stored at -80°C.

3C11 cells were collected and counted (section 2.1.15) before seeding in the central 60 wells of a 96 well cell culture plate at $2.5x10^5$ cells/well in 100µl D10 per well. The next day, media was aspirated from cells. For each sample,

50µl supernatant was added to each well. Recombinant IFN-α2 in 50µl D10 was added to wells in duplicates as a standard. Cells were incubated for 24 hours.

After 24 hours, 50µl of ONE-Glo Assay Buffer with Substrate (Promega) was added to each well. Plates was incubated for 3 minutes in the dark. 75µl of substrate mixture was transferred to an opaque white 96 well plate. Luminosity of firefly luciferase was measured on a GloMax Multi Detection System (Promega).

Standards were used to generate a standard curve via four parameter logistic regression in Graphpad Prism 7. Standards were used to infer sample type-I IFN concentration.

2.4 Buffers and Reagents

2.4.1 D10 Medium

- Dulbecco's Modified Eagle's Medium (DMEM) (Sigma)
- 10% Fetal Calf Serum (FCS) (Sigma)
- 100U/ml Penicillin (Sigma)
- 100U/ml Streptomycin (Sigma)
- 2mM Glutamine (Sigma)

2.4.2 R10 Medium

- RPMI (Sigma)
- 10% FCS (Sigma)
- 100U/ml Penicillin (Sigma)
- 100U/ml Streptomycin (Sigma)
- 2mM L-Glutamine (Sigma)

2.4.3 Viral Growth Medium

- DMEM (Sigma)
- 0.1% Bovine Serum Albumin (BSA) (Sigma)
- 100U/ml Penicillin (Sigma)
- 100U/ml Streptomycin (Sigma)
- 2mM Glutamine (Sigma)

2.4.4 TAE Buffer

TAE buffer was made as a 50X stock solution, and diluted to a 1X working solution in dH_2O.

- 242g Tris Base
- 18.61g EDTA
- 57.1ml Glacial Ascetic Acid
- dH_2O to 1L

2.4.5 FACS Buffer

- 474ml DPBS
- 2ml 0.5M EDTA
- 34ml FBS

2.4.6 Freezing Medium

Freezing medium was made with FCS supplemented with 10% DMSO.

2.4.7 Biotinylation Buffer and Biotin Wash Buffer

Biotinylation buffer and biotin wash buffer were prepared in DEPC-treated water.

Biotinylation buffer was made with the following:

- 100mM Tris pH 7.5
- 10mM EDTA

Biotin wash buffer was made with the following:

- 100mM Tris pH 7.5
- 10mM EDTA
- 1M NaCl
- 0.1% Tween20

2.4.8 S-Flu Virus

We performed infection experiments using a pseudotyped influenza virus that is not replication competent, known as S-Flu. This virus was developed, produced, and given as gift from the Alain Townsend group [268]. This virus is produced in MDCK cells via a series of cDNA constructs. The majority of progeny virus proteins and genes are derived from the Cambridge strain of the PR8 virus. The exceptions are the psuedotyped HA protein (S-Flu strain dependent), and the HA gene - the coding sequence of this gene is replaced by an eGFP gene. Cells infected by S-Flu virus will express eGFP.

In experiments where S-Flu was used, we used virus pseudotyped with H1/England/2015 HA containing two mutations. The first was G225D - this mutation is naturally occuring in H1, and has been shown to be a key substitution in avian influenza switching from avian to human sialic acid receptors [269, 270]. The second mutation was K154E - this has been shown to facilitate higher yields of progeny virus [271].

2.5 Bioinformatic Analyses

We used bioinformatics programs for RNA-Sequencing analysis. Our RNA-Seq pipeline is outlined in fig. 2.3. When performing analysis, we plotted gene expression in units of Fragments Per Kilobase of transcript per Million mapped reads (FPKM) or counts normalised by the DESeq2 program after transformation with a variance stabilising transformation. This transformation is similar to a log_2 transformation, and units should be treated as if they are on a log_2 scale, rather than as if they are linear. FPKM units are linear.

We ran bioinformatics programmes on various platforms. FastQC and all programs run from the unix terminal were run on CBRG bioinformatics servers. R based programmes were run locally [272].

2.5.1 Quality Control with FastQC

FastQC was used to get an early profile of the quality of raw sequence files [273]. This program offers a series of quality checks for sequencing data.

When using FastQC, we do not require or expect RNA-Sequencing data to pass all quality checks. For example, an even distribution of per base sequence content is not expected for the first 10-12 bases as a result of random hexamer priming. We also expect to see high levels of sequence duplication in RNA-Seq data - gene quantification is based on alignment of overlapping sequences to a specific region.

2.5.2 Quality Control with cutadapt

We used cutadapt version 1.8.3 to trim contaminating sequences [274]. We also restricted the minimum read length to 1bp with the –minimum-length option after trimming. Without this option we found that cutadapt would produce empty reads that would interfere with genome alignment.

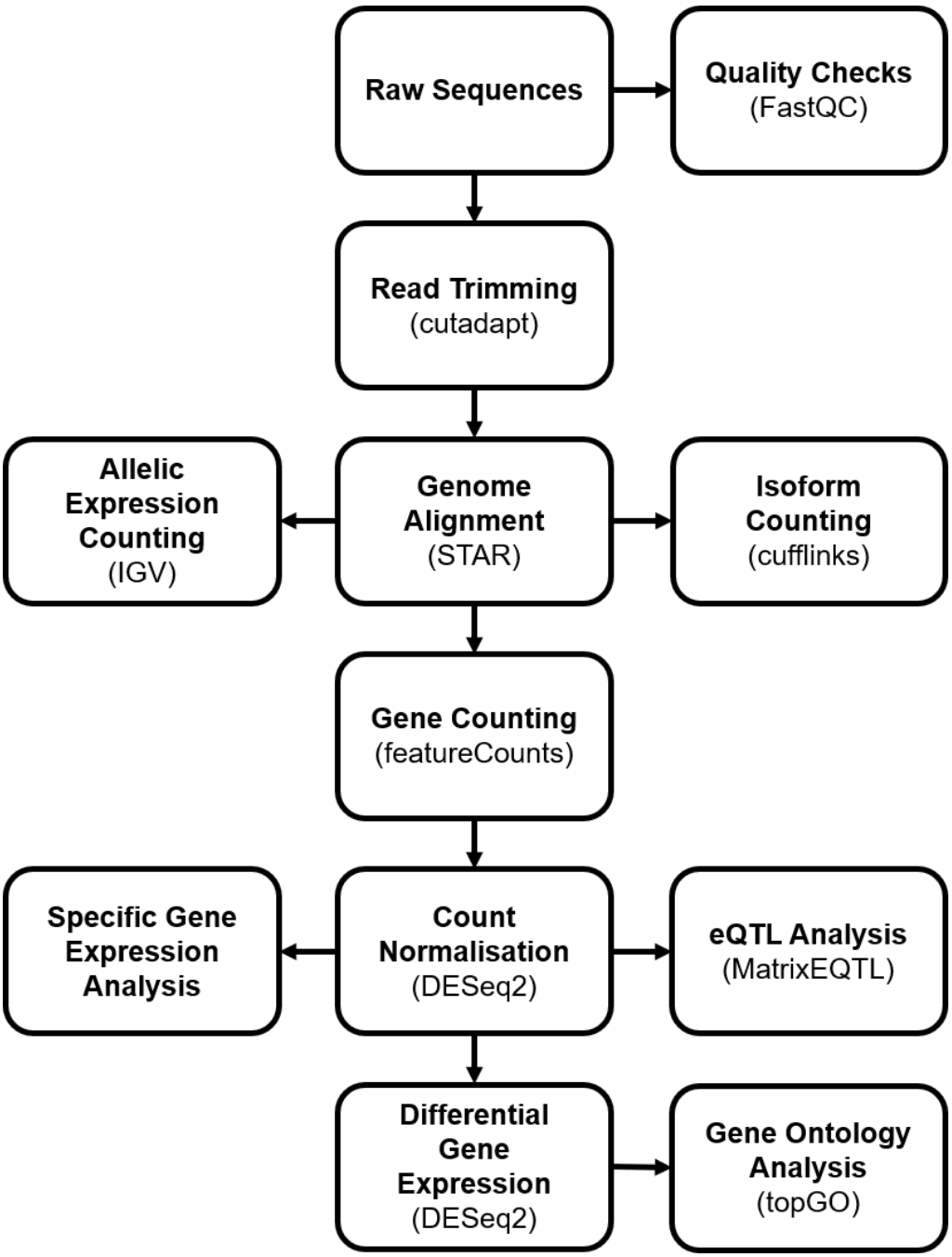


Figure 2.3: Flow Chart showing an overview of programs and techniques used in analysing RNA-Sequencing data.

2.5.3 Genome Alignment with STAR

We performed the majority of genome alignments with STAR version 1.1.8 [275]. When performing alignments, we used the most recent version of the human reference genome from the ensembl database [276].

We mostly used default STAR options. We chose to not allow for multi-mapped reads, restricting each read to a single alignment location in the genome. Files were generated in *.bam* format, with reads sorted according to genomic coordinate.

2.5.4 Feature Counting with featureCounts

We counted aligned sequences using featureCounts version 1.4.5, a part of the Subread package [277]. featureCounts is used to obtain count values for gene expression. This programme was used immediately after alignment with STAR, to aggregate counted features for all samples in a dataset. For gene counting, we used the genome annotation that pertained to the reference genome used for alignment. We ran featureCounts in paired-end mode. We counted features by exon alignment and aggregated counts by gene id.

2.5.5 Production of Index Files for Alignment Visualisation

In order to visualise aligned reads as a track on genome viewers, we used samtools version 1.3 [278]. This produces an accessory index file to accompany an alignment file necessary for read visualisation.

2.5.6 Quantification of Isoform Level Expression Using Cufflinks

We used cufflinks version 2.21 to count isoform level expression [279]. Cufflinks uses a probabilistic model of splicing, where estimates are made for the likelihood of each read originating from specific isoforms of a gene. This probability is largely based on prevalence of exon-exon junctions.

Cufflinks was used with default parameters, with the exception of the `-u` option. The -u option helps to accurately assign multiple mapping reads - this was important for us, given the high homology of *IFITM3* to other *IFITM* genes.

2.5.7 Differential Gene Expression With DESeq2 and Accessory Packages

We performed differential gene expression analysis with DESeq2 [280]. This analysis was done in concert with the biomaRt package [281], and uses the ggplot2 package for plotting [282]. We performed standard analysis with DESeq2, usually performing comparisons directly between two factors. DESeq2 uses a Wald test to directly compare gene expression levels across sample groups. This incurs a large multiple test correction via Benjamini-Hochberg. In testing for additive models of gene expression, we converted genotypes into a continuous scale. Individual genes are filtered out of testing by default, based on the mean counts of the respective gene. Low expression genes are often removed as well.

2.5.8 Gene Ontology Enrichment with topGO

Gene Ontology (GO) analysis is used to identify generalised trends in gene groups. GOs are annotated groups of genes categorised by the Gene Ontology Consortium [283, 284]. Ontologies are updated with respect to novel biological information, and are generated based on a combination of experimental information, phylogenetic inference, and computational inference. There are over 40,000 ontologies annotated as of 2018. Ontologies are generated based on three concepts:

- Cellular Compartment - localisation of gene products
- Molecular Function - molecular activity of gene products
- Biological Process - pathways and interactions between gene products

We used the topGO [285] package to identify enriched genes. This package functions based on statistical tests for enrichment of specific ontologies within

a gene set, with comparison to gene incidence expected to occur at random. We used a fisher test in our analyses. We filtered gene subsets from obtained differential expression based on criteria appropriate to the experiment.

2.5.9 eQTL Analysis with matrixEQTL

We used MatrixEQTL to perform eQTL analysis [286]. We used variance stabilised normalised counts data from DESeq2 as expression values, and inputted patient sex and age as covariates. We used standard linear modelling as our method of testing for eQTLs. Like in DESeq2, MatrixEQTL applies a standard Benjamini-Hochberg multiple test correction in all analyses to control false discovery rate.

In chapter 4, we used a single SNP for analysis (rs12252), and in chapter 5, we used SNP genotypes obtained from the Axion Precision Medicine Research Array (Thermo Fisher Scientific). We converted the obtained SNP matrix into a numerical matrix, where reference SNP homozygotes were denoted by a 0, heterozygotes were denoted by a 1, and alternative SNP homozygotes were denoted by a 2. If genotyping was not successful, or gave an unexpected genotype it was converted to an "NA". This was done with a python script created by Philip Staupe.

2.5.10 Deconvolving IFN Signature from *IFITM3* Expression

We used a custom methodology to deconvolve the IFN Signature from *IFITM3* expression. This is based on expression values from the ontology "response to type I interferon" (GO:0034340). We removed *IFITM1* and *IFITM3* from this ontology and genes that had not undergone testing with DESeq2.

We obtained a matrix of gene expression as in table 2.13:

Table 2.13

	Patient 1	**Patient 2**	**Patient 3**
Gene 1			
Gene 2			
Gene 3			

We filtered out all genes that were not distributed normally. For each gene/row of this matrix, we performed a Shapiro-Wilk test for normality. We used test results for gene filtering, with significance threshold of $\alpha = 0.05$ after a Benjamini-Hochberg multiple-test correction.

For each patient, we standardised gene expression values using the following equation. This equation is performed for a single gene in a single patient:

$$Ep_{std} = \frac{Ep_{s1} - \bar{x}_{g1}}{\sigma_{g1}} \tag{2.1}$$

Where Ep_{std} is standardised gene expression, Ep_{s1} is a gene expression value for a single patient for gene 1, $\bar{x}_{g1}$ is mean gene expression across all patients for gene 1 and σ_{g1} is standard deviation for gene 1 across all patients. After expression standardisation, each gene had a mean expression of 0, and a standard deviation of 1.

This operation was performed throughout the aforementioned gene/patient matrix. For each patient, an ISG expression index was made:

$$ISGIndex = \frac{\sum Ep_{std}}{n_{genes}} \tag{2.2}$$

Where n_{genes} is the total number of genes used.

For each patient, the ISG Index was subtracted from the *IFITM3* Standardised Expression, or from an alternative gene of interest.

Chapter 3

Alternative Splicing of *IFITM3* in Relation to rs12252

3.1 Introduction

As discussed in section 1.6.4, the seminal Everitt et al study proposed a mechanistic link between rs12252 genotype and splicing of *IFTIM3* [200]. In this paper, it was proposed that rs12252, which lies in an *IFITM3* splicing acceptor site, affected the splicing profile of *IFITM3*. In particular, it was proposed that this SNP shifted the splicing profile of *IFITM3* in favour of an *IFITM3* transcript isoform known as ENST00000526811.4 within the ensembl genome browser.

As explained previously, the current *IFITM3* annotation in reference genomes predicts four different isoforms for *IFITM3*. These include the canonical transcript, a transcript predicted to undergo NMD, and two transcripts associated with truncation of the IFITM3 protein (ENST00000526811.4 and ENST00000602735.2). It should be noted that the two truncated isoforms are each predicted to use a non-canonical 5'UTR distinct from the one used by the accepted *IFITM3* promoter. This splicing model includes more transcripts than presented by the Everitt *et al* study, wherein only two *IFITM3* transcripts are discussed. It is likely that the other two transcripts were removed for simplicity. As outlined in the aforementioned study, rs12252 is oriented at the predicted splicing acceptor site needed for production of the two truncated isoforms. Again,

the truncated *IFITM3* isoforms are predicted to incur the production of a truncated IFITM3 protein product, due to the loss of the canonical start site, and usage of a downstream start codon.

The idea that rs12252-C affected the splicing of *IFITM3* was presented without evidence of this association. It is likely this connection was assumed given the association between rs12252-C and severe influenza, and the fact that this SNP lies in a predicted splicing acceptor site. While the authors merely proposed this as a link between rs12252 severe outcomes to infection, this hypothesis has resulted in the publication of a number of papers where this association has been accepted as a certainty, leading to inherent flaws within these studies. Notably, multiple studies have been performed utilising the truncated *IFITM3* cDNA in artificial expression systems as a surrogate for the rs12252-C polymorphism without establishing a real link between these protein isoforms and rs12252 [262, 287]. In fact, as mentioned previously, the genome annotations for the truncated isoforms of *IFITM3* did not have any experimental support at the time of the Everitt et al publication and at present [276].

In studies using the truncated IFITM3 protein isoform, is has been shown that the cellular distribution of the truncated isoform of IFITM3 is vastly different from the full length protein. This result is largely expected in the context of the literature - the NTD of IFITM3 contains an endocytosis motif discussed previously. With the fact that a large body of research was formed on relatively shaky assumptions, we decided it was incumbent on us to investigate the link between rs12252 and *IFITM3* splicing.

In the literature, there has also been relatively little investigation into the wider genomic context of rs12252. Often, disease associated SNPs do not directly cause their associated phenotype, but rather are coinherited with a disease SNP [288]. This phenomenon is largely attributable to LD, a phenomenon where two genetic variants are coinherited more often than expected. LD between two SNPs can be derived from a variety of factors, but is commonly the result of close genomic distances, reducing the chance of dissociation due to crossing over in meiosis. The degree of LD between two variants is influenced by the distance between variants and the time that has passed since both genetic variants arose. LD is generally expressed in terms of D' or r^2.

D' measurements are attained by dividing the observed disequilibrium between alleles by the maximum possible disequilibrium. Using this, variants

in full disequilibrium will have a D' of 1, and variants in complete equilibirum will have a D' of 0. Notably, this measurement becomes unreliable when there are extreme differences of the two variants. r^2 values are derived from correlation coefficients, similar to those used in linear regression. Like with D', variants in full disequilibrium will have an r^2 value of 1, and variants in complete equilibirum will have an r^2 value of 0. This metric is generally considered more robust than D'.

With all this considered, we aimed to investigate existing hypotheses regarding the effects of rs12252, and explore the *IFITM3* genetic locus to get a full picture of potential alternative causative variants.

3.1.1 Project Aims

The aims of this project were as follows:

1. With respect to the hypothesis that rs12252-C results in the production of truncated IFITM3:
   (a) Investigate the expression of full length IFITM3 protein in CC and TT cells
   (b) Investigate the splicing profile of *IFITM3* with respect to rs12252
2. Examine rs12252 in the wider context of its genomic locus, looking for potential proxy SNPs in LD

3.2 Results

3.2.1 Examining the Splicing Profile of IFITM3 via western Blot

Our first initiative with regards to *IFITM3* splicing was to probe for the existence of full length and truncated forms of the *IFITM3* protein. With the help of Dr. Boquan Jin, from the Fourth Military Medical University in Xian, China, we produced a panel of Abs specific to the N and C termini of *IFITM3*. Briefly, these were produced by immunizing mice with 21 amino

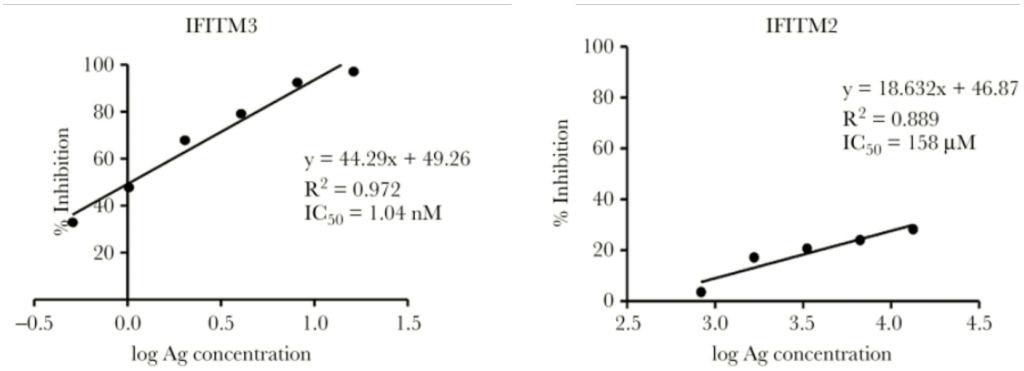


Figure 3.1: Inhibition ELISAs performed with our in-house IFITM3-NTD Ab. Increasing concentrations of free IFITM3 Ag (left) and IFITM2 Ag (right) were used to inhibit the binding of our in-house Ab to coated IFITM3. IFITM3 was more effective at inhibiting this binding by 10^5*-fold. Data provided by Dr. Shokouh Makvandi-Nejad*

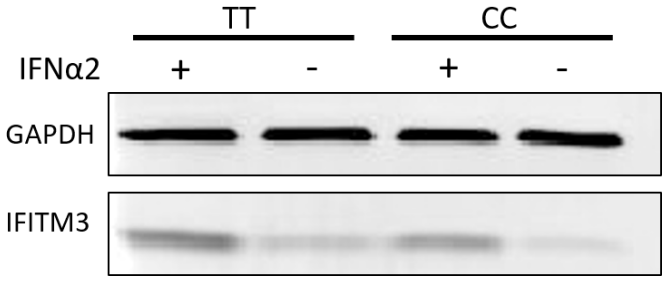


Figure 3.2: Western blot of TT and CC fibroblast cell lines showing expression of full length IFITM3 protein in both CC and TT fibroblasts. In both cell lines, the full length IFITM3 protein was visible before and after IFN induction.

acids from the N and C termini of IFITM3, conjugated to either BSA or Keyhole Limpet Hemocyanin (KLH).

Leading up to this project, characterisation of the aforementioned antibodies was performed by Dr. Shokouh Makvandi-Nejad. She showed specific IFITM3 protein binding by a single Ab specific to the IFITM3 NTD. This Ab had low cross-reactivity to IFITM2 as determined by inhibition Enzyme-Linked Immunosorbent Assay (ELISA). The IC_{50} of this antibody to IFITM3 was found to be 1.04 nM; in contrast, the Half Maximum Inhibitory Concentration (IC_{50}) of this antibody with respect to IFITM2 was 158µM (fig. 3.1). We did not identify any CTD specific antibodies.

With an antibody specific to the N-terminus of *IFITM3*, we looked for evidence of the full length protein in both TT and CC cell lines. We used primary human skin fibroblasts as a model cell line in this instance - we used one CC and one TT fibroblast cell line, due to the unavailability of a larger panel at the time. We used IFN stimulated (500U/ml IFN-α2, 24hrs), and untreated fibroblasts for this. We observed expression of the full length protein in both TT and CC cell lines (fig. 3.2). However, we could not confirm whether or not expression of truncated *IFITM3* was occurring using this system.

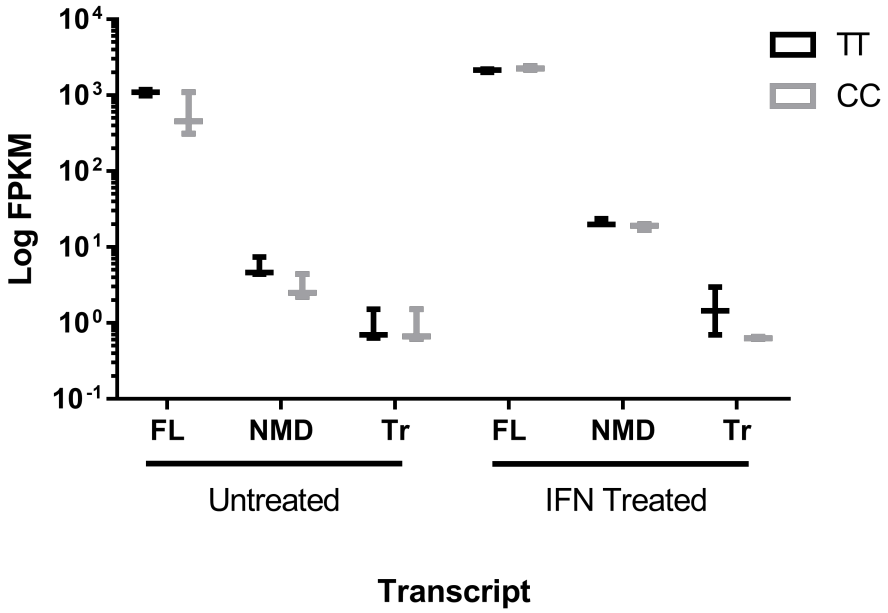
3.2.2 Examining the Splicing Profile of *IFITM3* via RNA-Sequencing

To obtain an in-depth profile of the splicing of *IFITM3*, we performed RNA-Sequencing to obtain quantitative measurements of isoform expression. We used primary human fibroblasts, with the intention to follow up this experiment in patient samples at a later date. At this point in time, this was our only cell culture system where both CC and TT genotypes were available. We sequenced fibroblasts that were either untreated, or stimulated for 24 hours with 500U/ml of IFN-α2. RNA-Sequencing was performed using standard Illumina protocols by the Wellcome Trust Centre for Human Genetics. Quality control is shown in Appendix D.1

Isoform Specific Expression Analysis

To obtain quantitative values for isoform expression of *IFITM3*, we performed isoform specific analysis using cufflinks 2.2.1 [289].This analysis returned quantitative FPKM values for all isoforms, including those for *IFITM3*. We performed splicing analysis with respect to the annotated *IFITM3* transcripts available through the ensembl genome browser, annotated via GENCODE 27.

The values obtained showed a clear dominance of the canonical full length transcript (ENST00000399808.4) predicted to make the full length *IFITM3* protein (fig. 3.3 and table 3.1). This isoform made up, on average, 98.9% and 99.2% of all *IFITM3* transcripts in TT and CC cells respectively under IFN stimulated condition. Similarly, when cells were not stimulated, 99.4% (TT) and 99.4% (CC) of transcripts expressed were the canonical full length



(a) Overview of *IFITM3 Splicing in Different Tissues*

Figure 3.3: Expression values of IFITM3 transcripts from primary human fibroblasts, shown on a log_{10} scale. Transcripts shown are full-length (FL), nonsense-mediated decay (NMD) and the sum of both truncated transcripts (Tr) Expression of the canonical transcript is overwhelmingly the most common. In addition, we observed expression of ENST00000531688.1, a transcript thought to be associated with NMD. The splicing pattern observed here does not differ between CC and TT fibroblasts. Data shown represents three technical replicates, with error bars showing SD

transcript.

Further examination of the splicing profile of these cell lines showed the expression of the isoform predicted to undergo NMD (ENST00000531688.1) (fig. 3.3). The expression of this transcript made up the bulk of the transcripts not accounted for by the canonical transcript, but nonetheless was expressed at relatively low levels; its expression made up 0.5-1% of total *IFITM3* expression. Via t-test, we saw that these these expression levels, as percentages of total *IFITM3* expression, were not significantly different between genotype groups,

Expression of the two transcripts associated with the truncated *IFITM3* protein product (ENST00000602735.2 and ENST00000526811.4) were found

Table 3.1 Expression of IFITM3 Transcripts in TT and CC Primary Cells

(a) Splicing in Untreated TT Cells

Transcript	**Ensembl ID**	**Expression (FPKM)**	**Percent Total**
Full Length	ENST00000399808.4	1087.1	99.4
NMD	ENST00000531688.1	5.48	0.50
Truncated	ENST00000602735.2 & ENST00000526811.4	0.952	0.09

(b) Splicing in Untreated CC Cells

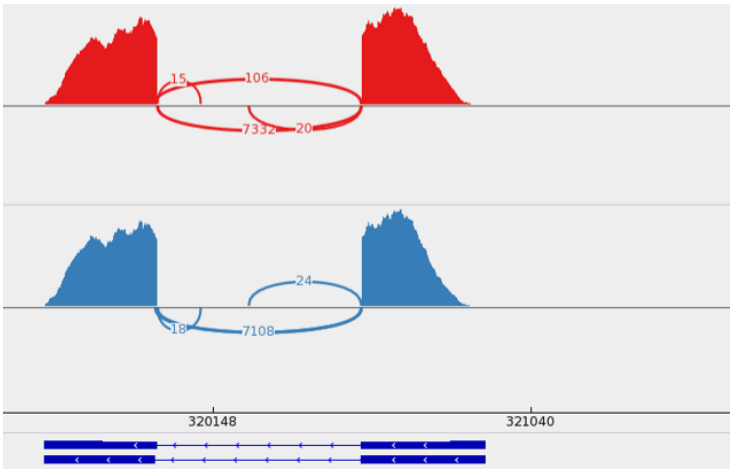
Transcript	**Ensembl ID**	**Expression (FPKM)**	**Percent Total**
Full Length	ENST00000399808.4	619.7	99.4
NMD	ENST00000531688.1	3.04	0.49
Truncated	ENST00000602735.2 & ENST00000526811.4	0.933	0.15

(c) Splicing in IFN Stimulated TT Cells

Transcript	**Ensembl ID**	**Expression (FPKM)**	**Percent Total**
Full Length	ENST00000399808.4	2117.9	98.9
NMD	ENST00000531688.1	21.17	0.99
Truncated	ENST00000602735.2 & ENST00000526811.4	1.71	0.08

(d) Splicing in IFN Stimulated CC Cells

Transcript	**Ensembl ID**	**Expression (FPKM)**	**Percent Total**
Full Length	ENST00000399808.4	2284.9	99.2
NMD	ENST00000531688.1	18.68	0.81
Truncated	ENST00000602735.2 & ENST00000526811.4	0.63	0.02



(a) Sashimi Plots of TT and CC Fibroblasts

Figure 3.4: Sashimi plot showing splicing splicing profile graphic for CC (red, top) and TT (blue, bottom) fibroblasts. Splicing profiles look roughly identical here, with no splicing events observed between the canonical first exon of IFITM3 and any 5' UTRs.

to be extremely low compared to the other *IFITM3* transcripts.

Visual Examination of *IFITM3* Splicing Profile

We also used the sashimi plot function on the alignments files that corresponded to the aforementioned analysis (fig. 3.4) [264, 290]. The use of sashimi plots allows for visualisation of events wherein aligned reads span two different exons, while additionally showing plotting density with respect to each exon. In visualising splicing using sashimi plots, the vast majority of exon spanning reads spanned the canonical first and second exons of *IFITM3* (ENSE00001540254 and ENSE00001540252 respectively); this was consistent with the finding that the bulk of *IFITM3* transcripts were the full length transcript. Again, consistent with the findings seen in cufflinks, there were a small number of reads that spanned the exon specific to the NMD transcript (ENSE00002175954), and zero exon spanning events specific to the two exons associated with the truncated transcripts.

3.2.3 Analysis of *IFITM3* Splicing in Other Cohorts

We used the methodology explained above on multiple sets of RNA-Sequencing samples, each time performing the same analysis to identify the splicing profile of the *IFITM3* gene. This analysis was performed on:

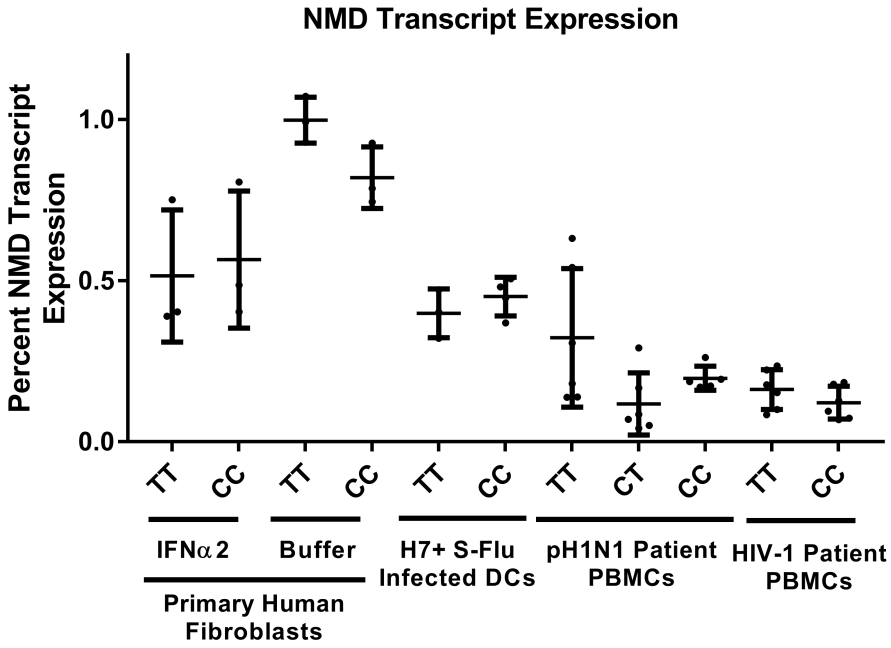
1. Monocyte-derived DCs positive for infection with IAV (section 6.1).
2. PMBCs from the 2009 H1N1 pandemic. These patients all had mild IAV infection (Appendix A.1).
3. PBMCs from acute HIV patients (Appendix A.2).

These cohorts are expanded upon in indicated sections. Quality control is shown in section D.1.1.

Isoform Specific Expression Quantification

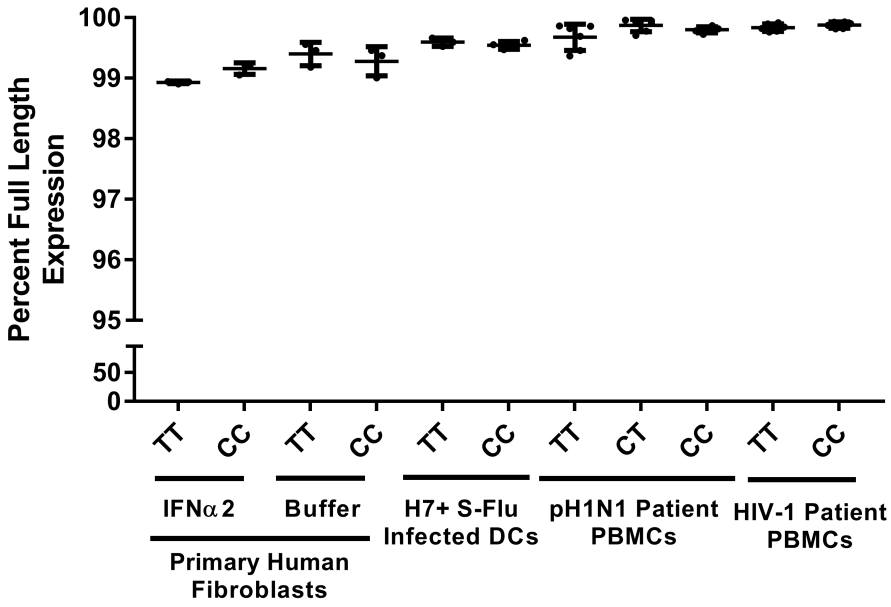
In all of these cohorts, the splicing pattern of *IFITM3* was overwhelmingly similar to that observed in the fibroblasts previously (table 3.2, table 3.3, and table 3.4). In general, we could always observe expression of the NMD transcript (ENST00000531688.1). The NMD transcript was consistently expressed at low levels in these cohorts; this was overwhelmingly below 1% of total expression. Interestingly, the expression level of this transcript did appear to vary marginally by tissue type, although not by genotype (fig. 3.5).

Overall, we obtained evidence that *IFITM3* splicing was almost exclusively tied to the canonical transcript in these systems (fig. 3.6).



(a) Overview of *IFITM3 Splicing in Different Tissues*

Figure 3.5: Percent of total expression used by the IFITM3 isoform associated with NMD (ENST00000531688.1). Samples were taken from a wide range of in vitro and in vivo sources, under infection and normal conditions. Overall expression of this transcript makes up a small proportion of total transcripts, although this does appear to vary slightly between tissue types. Error bars show mean and SD.



(a) Overview of *IFITM3 Splicing in Different Tissues*

Figure 3.6: Percent of total expression used by the canonical, full length IFITM3 splicing isoform. Samples were taken from a wide range of in vitro and in vivo sources, under infection and normal conditions. Regardless of rs12252 genotype, IFITM3 splicing strongly favours the canonical isoform. Error bars show mean and SD.

Table 3.2 Percent Expression of IFITM3 Transcripts in pdmH1N1 Patient PBMCs

Transcript	**Ensembl ID**	**Expression (FPKM)**	**Percent Total**
TT Patients			
Full Length	ENST00000399808.4	397.2	99.7
NMD	ENST00000531688.1	1.17	0.3
Truncated	ENST00000602735.2 & ENST00000526811.4	0	0
CT Patients			
Full Length	ENST00000399808.4	1009.9	99.8
NMD	ENST00000531688.1	1.61	0.2
Truncated	ENST00000602735.2 & ENST00000526811.4	0.13	0
CC Patients			
Full Length	ENST00000399808.4	1943.6	99.8
NMD	ENST00000531688.1	4.21	0.2
Truncated	ENST00000602735.2 & ENST00000526811.4	0.3	0

Table 3.3 Percent Expression of IFITM3 Transcripts in TT and CC IAV-infected Monocyte Derived DCs

Transcript	**Ensembl ID**	**Expression (FPKM)**	**Percent Total**
TT Cells			
Full Length	ENST00000399808.4	1959.3	99.6
NMD	ENST00000531688.1	8.2	0.3
Truncated	ENST00000602735.2 & ENST00000526811.4	0.07	0
CC Cells			
Full Length	ENST00000399808.4	2929.9	99.5
NMD	ENST00000531688.1	12.70	0.5
Truncated	ENST00000602735.2 & ENST00000526811.4	0.12	0

Table 3.4 Percent Expression of IFITM3 Transcripts in TT and CC HIV Patient PBMCs

Transcript	Ensembl ID	Expression (FPKM)	Percent Total
TT Patients			
Full Length	ENST00000399808.4	977.2	99.9
NMD	ENST00000531688.1	1.51	0.1
Truncated	ENST00000602735.2 & ENST00000526811.4	0	0
CC Patients			
Full Length	ENST00000399808.4	978.3	99.5
NMD	ENST00000531688.1	1.43	0.2
Truncated	ENST00000602735.2 & ENST00000526811.4	0	0

3.2.4 Examining the Splicing of *IFITM3* via the ExAC Database

We additionally used the exome data available on the ExAC database to observe the incidence of reads aligning to exons pertaining to truncated forms of the *IFITM3* transcript(ENSE00002162154 and ENSE00003321292) [242]. As outlined in Appendix C.2, the ExAC database does not disclose information about the tissue state or disease state of the individuals it obtains exome data from, so this data could not be, and was likely not explicitly linked to infection related contexts. 2144 samples were homozygous for rs12252-C. Consistent with our earlier findings, there was no evidence of any alignment to either of the two predicted *IFITM3* exons associated with the truncated IFITM3 protein product (fig. 3.7). We were unable to observe any evidence of transcript alignment to the 5' UTRs associated with the truncated transcripts whatsoever. Additionally, a relatively small degree of alignment to the NMD transcript specific exon was observable.

3.2.5 Analysis of SNPs in LD with rs12252

Given our conclusions that rs12252 did not appear to affect the splicing of *IFITM3*, we had no mechanistic hypothesis to explain the observed association between rs12252 and its disease phenotype. With this in mind, we decided to examine SNPs in LD with rs12252. We hypothesised that the



(a) ENST00000526811.4



(b) ENST00000602735.2



(c) ENST00000531688.1

Figure 3.7: ***a*** *Alignment to ENST00000526811.4 (truncated) for transcripts within the ExAC browser, showing the mean alignment depth to each exon.* ***b.*** *Alignment to ENST00000602735.2 (truncated) for transcripts within the ExAC browser, showing mean alignment depth to each exon.* ***c.*** *Mean alignment to the ENST00000531688.1 (NMD) transcript for transcripts within the ExAC browser, showing a small degree of alignment to the exon specific to the aforementioned transcript. For both truncated transcripts, it can be observed that there is no recorded alignment to the 5' UTRs incorporated into both truncated transcripts. Direction of transcription for all is right to left.*

observed effects of rs12252 might be explained by a causative SNP in LD with rs12252.

To examine the presence of SNPs in LD with rs12252, we used the phase 3 of the 1000 genomes project database (reviewed in Appendix C.1), with the LDlink tool [241, 291]. When investigating potential proxy SNPs, we used a relatively liberal r^2 cutoff threshold of 0.7.

We first examined proxy SNPs for rs12252 using all pooled populations (fig. 3.8). As expected, in this analysis, SNPs with higher r^2 values clustered near to rs12252. Here, no SNPs had r^2 values greater than 0.5, indicating that worldwide, there is no clear surrogate SNP for rs12252 sufficient to explain its observed association with viral disease.

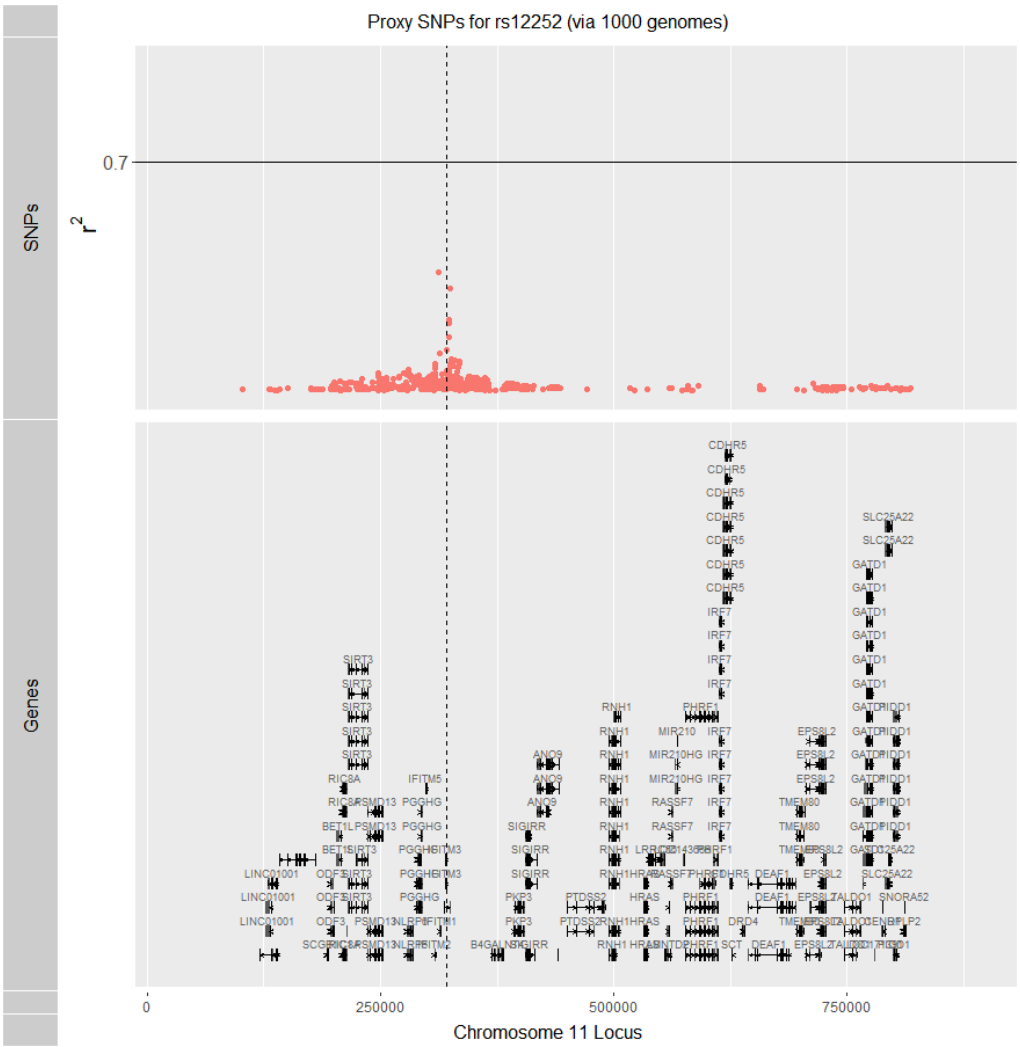
We chose to narrow our analysis to find all proxy SNPs for rs12252 using three population groups as datasets: EUR, EAS and Han Chinese in Beijing China (CHB) (fig. 3.9). These populations covered all world populations with reported associations between rs12252 genotype and infection outcome. While CHB was more specific to the Chinese populations we observed this association in, this dataset confers less statistical power than the EAS dataset. Using this threshold, we found that no SNPs were effective proxies for rs12252 in EUR - this is not unexpected given the low allele frequency in this population. When examining proxy SNPs in EAS and CHB populations, we observed three SNPs with high r^2 values relative to rs12252 - these SNPs are shared between these two population groups (table 3.5) .

Table 3.5 High LD SNPs with rs12252

SNP Number	**EAS**			**CHB**		
	MAF[1]	r^2	**D'**	**MAF**[1]	r^2	**D'**
rs7948108	0.4593	0.9304	0.8220	0.4320	0.7811	0.9374
rs77537847	0.4504	0.9249	0.7835	0.4369	0.8333	0.9588
rs10794307	0.4157	0.9864	0.7736	0.4029	0.7536	0.9776

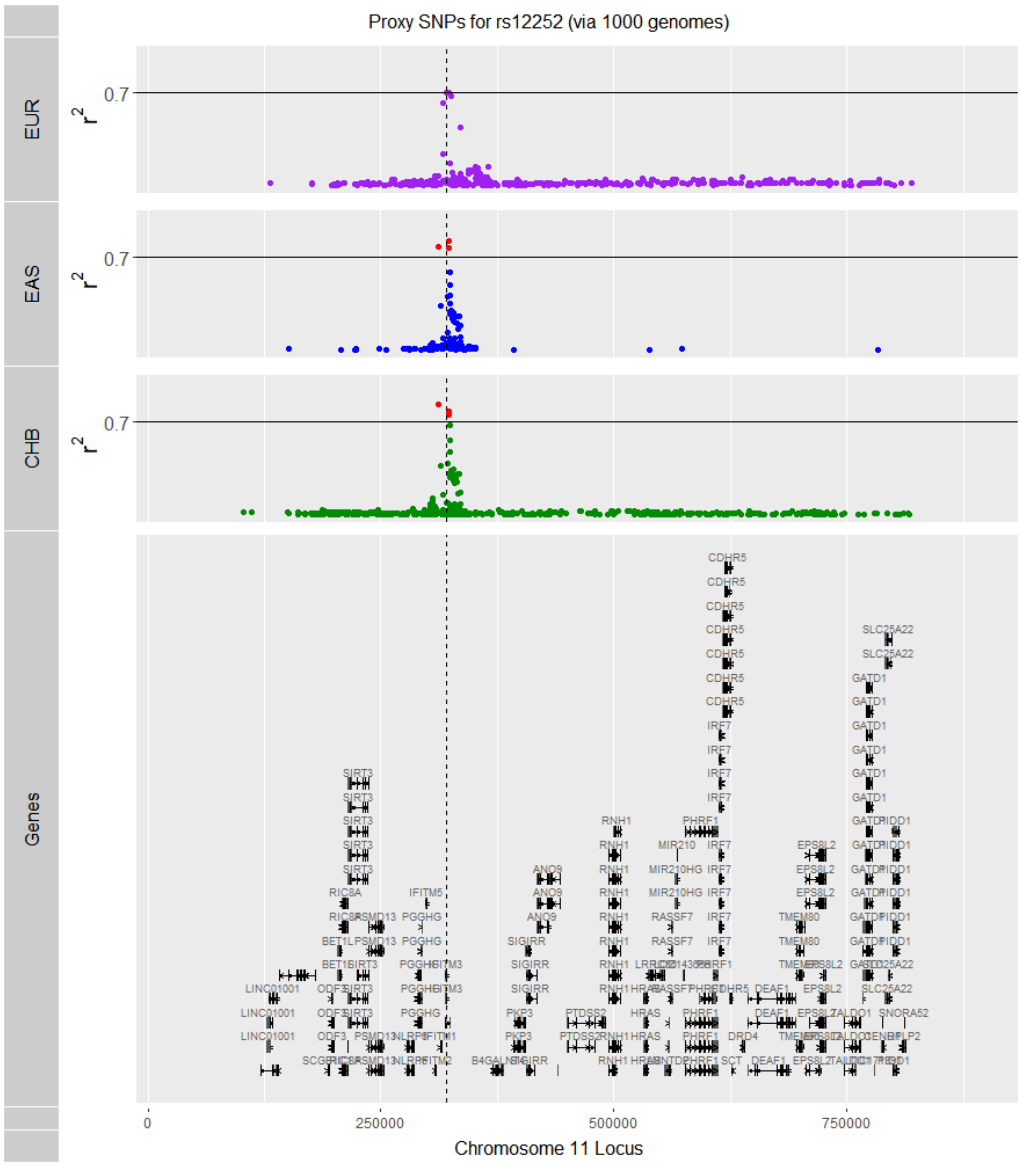
[1] Minor Allele Frequency

Upon the discovery of these SNPs as proxies for rs12252, we looked further into their loci, and what effects may be incurred by them, if any (fig. 3.10 and **??**). The rs7948108 and rs10794307 SNPs were located downstream of the *IFITM3* locus, 2735 and 1950bp away from rs12252 respectively. The rs77537847 SNP conversely, was found upstream of *IFITM1*, 8521bp from rs12252.



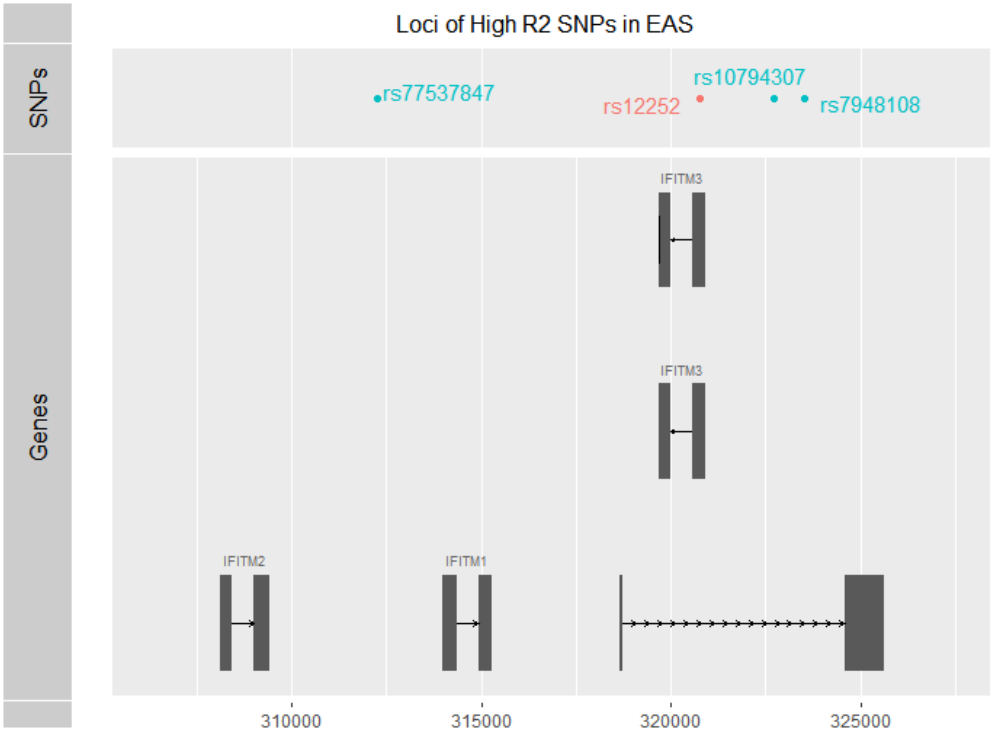
(a) Proxy SNPs for rs12252 in All World Populations

Figure 3.8: Proxy SNPS for rs12252 using all world populations in the 1000 genomes database. Top track shows SNPs in the 1000 genomes database, arranged by genomic locus and r^2 value with respect to rs12252. Bottom track shows annotated genes. The rs12252 locus is shown with the vertical dotted line. No SNPs were seen to have high r^2 values, indicating that there are no high LD variants for rs12252 worldwide.



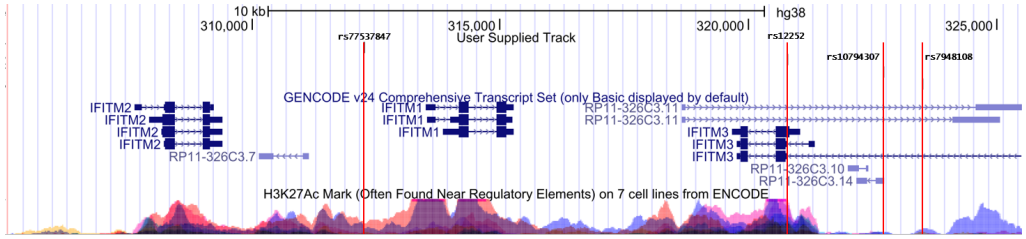
(a) Proxy SNPs for rs12252 (via 1000 Genomes)

Figure 3.9: Proxy SNPS for rs12252 using the EAS, EUR and CHB datasets in the 1000 genomes database (Top 3 tracks, arranged by locus and r^2 value.). Bottom track shows annotated genes. SNPs with r^2 values higher than 0.7 are shown in red. The rs12252 locus is shown with the vertical dotted line. Both the EAS and CHB datasets show SNPs with r^2 values higher than 0.7; none of these are seen using the EUR dataset.



(a) High r^2 Value SNPs with rs12252 in EAS and CHB

Figure 3.10: Proxy SNPs (shown in blue) for rs12252 (shown in orange) using the EAS and CHB datasets in the 1000 genomes database. Bottom track shows annotated genes. Unannotated gene is a non-coding RNA. SNP loci fell close to the IFITM3 locus, but were not found within the core promoters or coding sequences of any genes.



(a) High LD SNPs in Relation to H3K27Ac Markers

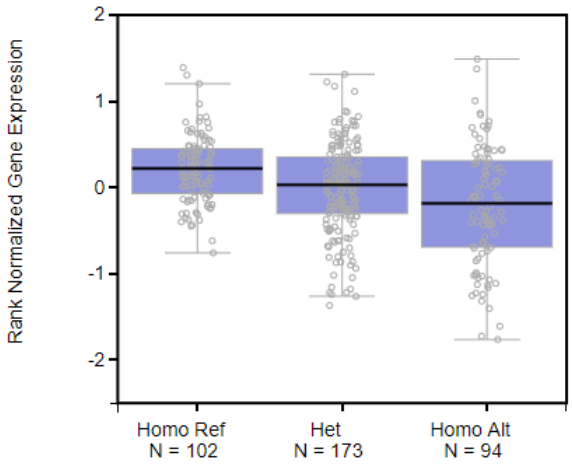
Figure 3.11: Proxy SNPs for rs12252 Overlayed with H3K27Ac marker. These markers can serve as an indication for the location of genomic regulatory elements. Figure shows the IFITM genomic locus, with SNP loci denoted by red lines. H3K27Ac marker track is shown as an overlay between seven different cell lines, represented by different colours. Both rs7948108 and rs10794307 are not located near markers, but rs77537847 overlaps with markers in some cell lines.

None of these SNPs were located within coding regions or the core promoters of any genes. We used the overlapping ENCODE 27 Histone H3 Lysine 27 Acetylation (H3K27Ac) marker track on the UCSC genome browser relative to these SNPs (fig. 3.11) [292, 293]. H3K27Ac serves as a marker of active transcription, and can be used as a marker for potential enhancer sites. Only rs77537847 was seen to be associated with relatively high levels of H3K27Ac markers. These markers were specific for two of the seven possible tracks generated from seven different cell lines; the GM12878 (immortalised B cell line) and K562 (mylogenous lukaemia cell line) cell lines. Other cell lines tested included lung fibroblasts (NHLF) and embryonic stem cells (H1-hESC). In addition, this SNP overlaps with a Deoxyribonuclease-I (DNase-I) hypersensitivity site discovered as part of the ENCODE project. DNAse-I hypersensitivity sites also generally indicate regulatory features. However, this SNP does not overlap any described regulatory features in the ensembl build.

3.3 Discussion

3.3.1 Full Length IFITM3 Protein is Expressed Regardless of rs12252 Genotype

In studying the IFITM3 gene, one of the largest challenges we encountered was the lack of viable reagents for the IFITM3 protein. In particular, there



(a) High r^2 Value SNPs with rs12252 in EAS and CHB

Figure 3.12: eQTL boxplot examining the relationship between rs7948108 and IFITM3 expression in whole blood. Genotype significantly differentially affects the expression of IFITM3 ($p = 6.4 \times 10^8$).

was no commercially available IFITM3 specific Abs, as evidenced by the fact that published IFITM3 studies largely use either cross-reactive IFITM3 Abs, or recombinant protein. Both of these systems suffer from drawbacks. When using cross-reactive Abs, IFITM3 and IFITM2 cannot be differentiated between, even in a size discriminating assay such as western blotting.

It is interesting that there were no Abs raised specific to the CTD of IFITM3. The IFITM3 CTD is the area of the protein with the most sequence divergence from IFITM2 - only two of the six aas are shared. Despite this, the CTD of IFITM3 represents a small epitope for Ab binding. Abs generally bind to epitopes with greater than 20 contact residues, a criteria not fulfilled here [294].The NTD of IFITM3 is a larger target, which may be a prerequisite for Ab production.

The generation of an IFITM3 specific antibody is a valuable reagent for the study of this protein. Generally speaking, *in vitro* studies have used IFITM3 constructs stably transduced into cell lines, often with an added protein tag for detection. This practice is not ideal; it is unclear whether or not these tags affect IFITM3 function in the aforementioned studies. Protein tags have, in some cases, been shown to incur a abnormal effects on the protein they are attached to [295]. These studies often overlook the fact that a wide range of

cell lines express IFITM3 (including HEK293 and A549) - its expression can generally be induced by interferon. The acquisition of an antibody specific to IFITM3 will be an important reagent for examining the function and localisation of the protein in the future.

With our development of the an antibody specific to the N-terminus, we were able to, for the first time, show true expression of full length IFITM3 protein in a physiological context. When using this antibody on primary human fibroblasts raised from CC and TT donors, we could observe clear expression of the full length protein in both cases. While this assay was not quantitative, this clearly indicated that CC and TT individuals were capable of producing full length IFITM3. While showing this at the protein level was an important step in assessing the claims made by Everitt *et al*, it is important to note that this did not rule out the possibility of expression of the truncated protein form. Without a C-terminal antibody, this task was still not possible at this stage without the use of full-proteome techniques such as SWATH mass-spectrometry [296].

3.3.2 The Splicing Profile of *IFITM3* is Unaffected by rs12252

We used a combination of four different RNA-Sequencing cohorts to examine the splicing profile of *IFITM3*. These cohorts all, in some way, pertained to viral infection. In all cases, we found little variability in splicing profile with respect to rs12252 genotype. Overwhelmingly, we observed expression of the canonical *IFITM3* transcript, with effectively no expression of truncated transcripts. Overall, this finding strongly rebukes the idea that rs12252 is associated with splicing of *IFITM3*. This was consistent across different tissues, and in the context of conditions related to viral infection.

It is important to note that tissue specific isoform expression is not an uncommon occurrence in the human transcriptome [297]. It is possible that we missed the production of truncated *IFITM3* isoforms due to the use of incorrect tissue types. It could be argued that, given its relationship to IAV infection, this splicing activity may take place in lung epithelial cells. It was not possible for us to obtain lung samples from IAV patients with active infection, as this is not routinely collected. This idea is somewhat contradicted by the fact that rs12252-C appears to also affect active HIV-1 infection. HIV primarily infects CD4+ T cells - a cell subset captured

within the PBMCs we performed RNA-Sequencing on. If alternative splicing explained observed differences in HIV infection with respect to rs12252 genotype, it would likely be captured in this set of patient samples.

In using the ExAC database, we again found no indication of expression of transcripts including 5' UTRs exclusive to truncated *IFITM3* isoforms. This provides strong support that this finding is broadly applicable for a wide range of tissues and genetic backgrounds. In addition, the use of the ExAC browser provides evidence for our finding using a higher powered cohort than those curated by us personally.

When analysing the splicing profile of *IFITM3*, we observed expression of a transcript predicted to undergo NMD. This prediction was based on the fact that this mRNA has a Premature Termination Codon (PTC) in its open reading frame occurring before the final exon of the full mRNA. This is a well established signal for NMD [298]. RNA-Seq underestimates the production rate of PTC+ transcripts due to their fast rate of degradation [299]. This could suggest that production of this alternative isoform occurs at a substantially higher rate than its observed frequency implies. This is likely not relevant to rs12252, given that the rates of the NMD isoform are consistent across all genotypes. The fact that this transcript appears to make up different proportions of total *IFITM3* expression in different tissues could suggest that production or degradation of this transcript is favoured on a tissue-by-tissue basis. This would need further investigation to verify - it is again expressed at very low levels, where RNA-Seq is somewhat unreliable, and this observation could be the result of different sequencing protocols.

What is often not discussed with respect to the alternative splicing hypothesis is that the production of the *IFITM3* isoforms associated with truncation is entirely dependent on transcription initiation from alternative promoters to the canonical *IFITM3* promoter. Alternative gene promoters and 5' UTRs do occur naturally in mammalian genomes, and are often used for complex regulatory purposes [300]. Gene products with alternative 5' UTRs can also vary drastically in terms of their translation initiation potential [301]. With this in mind, it would not be unusual for *IFITM3* to use multiple promoters. The idea that splicing would be controlled by acceptor site affinity however, overlooks the fact that predicted alternative promoters need to be viable in the first place. It is also worth noting that hypothetical transcripts initiating from alternative *IFITM3* promoters are predicted to have a single viable initial splice site, and would be forced to make a mature mRNA associated with the truncated variant. However, it would be

possible, in this hypothetical model, that affinity for splicing machinery, as conferred by rs12252 could affect processing efficiency of *IFITM3* pre-mRNA. However, there is no indication from experimental evidence that predicted alternative *IFITM3* promoters actually show transcriptional activity in any circumstance. While it is possible that these are viable promoters, the fact that there is no supporting evidence for their activity makes the existence of their associated transcripts even less likely. In the future, this could be addressed by ribosomal profiling.

In the 2017 Randolph *et al.* study, RNA sequencing was performed on PBMC samples from paediatric influenza patients [249]. These patients were partially from an African-American background, and thus had a relatively higher rate of the C allele than in other, studies using individuals with European ancestry. In this study, the authors did not observe differential splicing of *IFITM3* with respect to rs12252 genotype, in concordance with our findings. The authors additionally identified the presence of the NMD transcript, and another, unannotated transcript. This transcript is almost identical to the canonical *IFITM3* transcript - the only difference is the loss of 5bp from the end of the first exon of *IFITM3*. While we did not observe this transcript, it is likely that this transcript would cause a frameshift due to this loss of 5bp, and produce a truncated protein product at the C terminus. It is unlikely this would have any functional relevance to rs12252 or viral restriction - it was found to be expressed at low levels, and was not differentially expressed according to genotype. We could not detect this transcript in our datasets.

In addition to this Randolph *et al* study, a 2017 Wu *et al* study has also provided novel evidence for a lack of truncated *IFITM3* isoform expression [302]. In this study qPCR was performed with primers specific to the ENST00000526811.4 truncated transcript in an attempt to amplify the associated cDNA. This experiment was performed in Jurkat E6-1 immortalized T lymphocytes, and Lymphoblastoid Cell Lines (LCLs) immortalised with Epstein-Barr Virus (EBV). LCLs were of TT, CT and CC rs12252 genotype - in all cases, no amplification of the ENST00000526811.4 transcript was observed. It is worth noting that the authors did not attempt to amplify the other transcript associated with truncated *IFITM3*, ENST00000602735.2.

What is interesting to note is that in this study, the authors observed expression of an *IFITM2* isoform associated with truncation at the NTD domain of the protein (ENST00000533141.1). The authors additionally

find that this protein form behaves similarly to artificial truncated IFITM3 proteins - its distribution is moved to the cell surface, and the different proteins differentially restrict infection on a virus-by-virus basis. It is interesting to note that this *IFITM2* isoform is predicted to be produced in a manner synonymous with that of the truncated *IFITM3* isoforms. Both of these alternative isoform sets use analogous splicing acceptor sites within the canonical first exon of *IFITM2* or *IFITM3*. A key difference between the *IFITM2* and *IFITM3* transcripts is the fact that alternative 5' UTRs/promoters have no sequence relation. This finding implies that the *IFITM3* splicing acceptor site associated with rs12252 is a viable acceptor site, given its identical sequence alignment to its analogue in *IFITM2*. With this evidence in mind, it is likely that it is the lack of activity from alternative *IFITM3* 5'UTRs/promoters is the key determinant underlying the fact that truncated *IFITM3* isoforms are not observable. This splicing site is clearly conserved within the *IFITM* family. This splicing site could help to facilitate the already plastic nature of *IFITM* gene N-termini in the context of gene evolution. As discussed previously in the Compton *et al* and Foster *et al* studies [240, 303], this plasticity could allow for stronger anti-HIV activity.

Despite the fact that we have accumulated strong evidence for a lack of rs12252-associated splicing of IFITM3, it is still possible that this SNP could be involved in the production of truncated IFITM3 proteins. A number of human mRNAs have been shown to be bicistronic, where two different proteins can be translated from a single mRNA. A 2014 study performed by Brubaker *et al* [304] showed that MAVS mRNA could produce a truncated version of the protein dubbed miniMAVS from the canonical mRNA. The miniMAVS protein retained some properties of the full length protein. In addition, the authors performed a wide scale ribosomal screening investigation, in U937 cells - a human monocyte cell line. Here, *IFITM2* was identified as a gene with bicistronic mRNA capable of producing a truncated protein product. This truncated product was translated from the *IFITM2* alternative start site analogous to the one previously discussed within *IFITM3* While no further investigation or validation has taken place with regards to this finding, it is possible that, given the sequence similarities between *IFITM3* and *IFITM2*, this is an unidentified property of *IFITM3* mRNA. In addition, this could be affected by rs12252 genotype, given the proximity of rs12252 to the predicted downstream start codon within *IFITM3* itself. However, without an anti-IFITM3 antibody specific to the CTD of the protein, showing the presence, or lack thereof, of this protein is not currently possible.

Overall, the evidence we have obtained from RNA-Sequencing cohorts is a strong refutation of the idea that the rs12252-C allele is associated with alternative splicing of *IFITM3*, and that this alternative splicing results in severe viral disease. The fact that this data is is backed up from a wide range of sources gives us a great deal of confidence in the conclusions we have reached. The finding that the truncated IFITM3 is likely non-existent is highly important for the IFITM3 field at large. It is unfortunate that a number of studies have been published under the assumption that this truncation exists in a physiological context. From this information, we have worked to re-examine this SNP from the ground up, looking to investigate its effects on on *IFITM3* and viral infection.

3.3.3 rs12252 is in LD with Three Proxy SNPs

With the finding that the presence of rs12252 did not affect the splicing profile of *IFITM3*, and that the canonical isoform was the overwhelmingly dominant *IFITM3* transcript, we decided to take a wider look at potential SNPs in LD with rs12252.

In GWAS studies in general, SNPs are often pulled down that have no discovered mechanistic link to the trait they are associated with. Identifying a true causal variant associated with a mechanism is a major challenge in these studies that often is never completed [305]. While methods have been developed to better address this, identifying true causal SNPs through LD is difficult for a number of reasons. For one, if a SNP does not have an obvious effect on a gene, such as incurring a missense mutation, its effects can be incurred through intergenic regions with no functional annotation. In addition, a SNP associated with a trait can be extremely far from its associated gene - it has been seen that these interactions can occur up to 2Mbps away [306]. In our investigation, we were confident that, if rs12252 was not a causative SNP, it would be in LD with a SNP that would affect *IFITM1*, *IFITM2* or *IFITM3*. This was owing to the wealth of experimental data showing strong associations between *IFITM* genes and restriction of viral infection.

In numerous patient studies related to rs12252, we observed strong risk associated with the C allele in EAS populations. It is unlikely this would be caused by a rare variant in LD with rs12252; if this was the case, we would likely not be able to observe any effects when discriminating patients

based on rs12252 genotype. For this reason, we felt it was most likely that a hypothetical, disease causing proxy SNP for rs12252 would have similar allele frequencies to rs12252 itself. For this reason, we used r^2 instead of D' as a measurement for LD. D' cannot distinguish between low and high allele frequency variants. In performing this analysis, we used a liberal r^2 threshold of 0.7. In the majority of published studies, an r^2 threshold of 0.8 is used [307]. We felt that by lowering this threshold by 0.1 we could get a more holistic view of potential proxy variants for rs12252.

In EAS and CHB patients, we observed a largely similar pattern of proxy SNPs for rs12252. This is unsurprising given that EAS is a superset of CHB, and that all sub-populations within EAS come from a similar genetic background. CHB yields a significantly noisier picture of proxy SNPs - this is likely a result of the fact that this cohort is underpowered compared to EAS. EAS contains 504 samples, in contrast to the 103 from CHB. We observed three SNPs with r^2 values above 0.7 in both cases - these were rs77537847, rs7948108, and rs10794307. With respect to the latter two SNPs, we did not find any strong indications of functional activity based on genomic context.

When investigating rs77537847, we found some indication of association between this SNP and regulatory regions of the genome, with respect to both H3K27Ac and DNase-I hypersensitivity markers. In both of these cases, this association appeared to be specific to only a subset of cells tested. This indicates that this potential regulatory region may be cell line or cell condition specific. These markers, in combination with one another give a stronger degree of support for this SNP acting as a causative SNP for the phenotypes associated with rs12252.

Taken as a whole, we argue that none of the SNPs in strong LD with rs12252 are likely to be causative SNPs. When compared to rs12252, none of these SNPs have obvious functional effects. It is worth noting that, in comparison to rs77537847, rs12252 concurs with much higher levels of H3K27Ac markers, and is located within *IFITM3* itself. While it is worth considering these SNPs in the future, we feel as if investigating rs12252 directly is the most effective course of action with respect to SNPs in LD with rs12252.

When performing LD analysis on rs12252, we were basing our findings entirely on SNPs called by the 1000 genomes project. It is important to recognise that it is possible that a potential causative SNP was missed, due to its lack of incorporation into the 1000 genomes database. LD analysis performed with the 1000 genomes database is performed with the phase 3

dataset exclusively; this is the first dataset to include haplotype data. It is documented that 2.3 million variant sites were lost between phase 1 and phase 3 of the project for a wide range of reasons. Ultimately, this indicates that it is possible that this analysis has entirely missed a SNP site that is in strong LD with rs12252. This database is, however, by far the strongest resource currently available for identifying proxy variants with relation to a SNP of interest.

With this in mind, we did not have sufficient access to a cohort with enough statistical power to get a convincing picture of SNPs in LD, while also parsing their effects with respect to viral infection. While we could observe this with rs12252 in the past, these studies were reliant on the fact that rs12252 has a high MAF in EAS populations, and thus a smaller patient number was needed to observe a genotype effect. The causative SNP would need to have a similarly high MAF to observe this association. While we could obtain data from healthy individuals to identify new SNPs in LD, this would not come with any functional associations. For reasons of practicality, we decided to investigate functional impacts of rs12252 via the CRISPR-Cas9 system. In using this system to generate different alleles of this SNP on an isogenic background, we could remove any effects of LD. This would ultimately allow us to investigate whether or not this SNP incurred effects on *IFITM3*, and increased cell susceptibility to infection.

3.3.4 Concluding Remarks

In this investigation, we have shown, with a great deal of evidence, that the observed association between the rs12252 SNP and severe outcomes to viral infections is not explained by the alternative splicing of *IFITM3*. We have observed a consistent *IFITM3* splicing profile across rs12252 genotypes, in contrast to the hypothesised alternative splicing associated with rs12252-C. On top of this, we find that the vast majority of *IFITM3* isoforms are consistent with the canonical transcript - overall, we hypothesize that *IFITM3* splicing is simple, and likely does not contribute meaningfully to the regulation of the protein as a whole. These findings were published in the Journal of Infectious Diseases under the title *Lack of Truncated IFITM3 Transcripts in Cells Homozygous for the rs12252-C Variant That is Associated With Severe Influenza Infection* [308].

With the idea that rs12252 does not result in the alternative splicing of

IFITM3, we reapproached our hypotheses with respect to this SNP. We did not find any proxy SNPs for rs12252 that are likely to underlie the observed association between rs12252 and an increased likelihood to develop viral infection. Based on this evidence, we hypothesised that the rs12252 was still mechanistically linked to increased rates of severe infection, but via an unknown mechanism. We showed, via some preliminary evidence from CC and TT fibroblast cell lines, that the C allele was potentially linked to reduced expression of *IFITM3*. This is discussed further in chapter 4.

Chapter 4

Expression of IFITM3 in Relation to rs12252

4.1 Introduction

Results from chapter 3 showed no evidence that splicing was modulated by rs12252 genotype. In making this finding, we were left with no mechanistic link between the SNP its associated clinical phenotypes. In addition, our analysis of SNPs in LD with rs12252 did not reveal any convincing candidate SNPs to explain the aforementioned clinical phenotypes. With this in mind, we chose to re-evaluate the likely mechanistic consequences of rs12252 with regards to *IFITM3*.

There are a number of well accepted mechanisms by which SNPs can perturb a gene, all of which have can relevance to disease phenotypes [309]. Broadly speaking, SNPs can alter a gene by:

- Generation of a premature stop codon (nonsense mutation)
- Changing the amino acid sequence (missense mutation)
- Changing the splicing profile of a gene
- Changing the expression of a gene

Since rs12252 is a synonymous SNP, and we had found no evidence that it

changed the splicing profile of *IFITM3*, the most likely mechanism by which this SNP could associated with perturbations of *IFITM3* and disease would be through expression of the gene itself. There are a large number of different mechanisms by which SNPs have been shown to change the expression of a gene. However, these mechanisms generally affect the abundance of their constitutive gene's mRNA, or the abundance of the resultant protein. Currently, associations between SNPs and mRNA expression are better understood than associations between SNPs and protein expression. This is due to the fact whole transcriptome analyses are much simpler than whole proteome analyses at present [310].

Associations between SNP genotype and mRNA expression of a gene are known as Expression Quantitative Trait Loci. Genome-wide studies have revealed thousands of eQTLs that act in cis and trans [311]. While a great number of eQTLs exist, they are largely dissociated from disease. In fact, a large number of eQTL SNPs are not mechanistically linked to their gene of interest, but are merely carried with another SNP via LD. Briefly, SNPs can affect the expression of mRNA through alteration of TF binding sites, promoter elements, miRNA binding sites, methylation sites, mRNA stability and other mechanisms. These are discussed in more detail, with relevance to rs12252 in section 4.3. SNPs can also alter protein abundance via mRNA folding, codon usage, and other mechanisms affecting translation efficiency.

When beginning this investigation, there was sparse evidence that *IFITM3* expression could be a determinant in infection. In the aforementioned Shen *et al.* study, it was reported that an *IFITM3* promoter SNP, rs3888188, affected the expression levels of *IFITM3* [211]. Lower levels of *IFITM3* expression correlated with increased susceptibility to paediatric Tuberculosis (TB).

Given this, decided to investigate whether or not *IFITM3* expression, and not splicing was modulated by rs12252. It was our hope that undertaking this study would help to discover the true mechanism by which rs12252 was associated with disease outcomes. In the following chapter, we provide evidence that rs12252 is associated with the abundance of *IFITM3* mRNA and protein, and potentially the transcriptional rate of *IFITM3*. We also find that this relationship is not consistent between sample types, and that rs12252 may have more broad effects of gene expression beyond *IFITM3*.

4.1.1 Project Aims

The aims of this project were as follows:

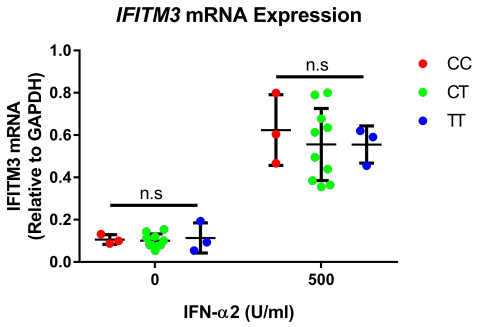
1. Investigate the idea that rs12252 is an modulates *IFITM3* expression using:
   (a) *In vitro* cell culture systems
   (b) Healthy donor samples
2. If the former is true, investigate specific cellular conditions in which rs12252 influences *IFITM3* expression
3. If the former is true, investigate potential mechanisms of rs12252
4. Investigate general transcriptome-wide expression trends associated with rs12252 genotype

4.2 Results

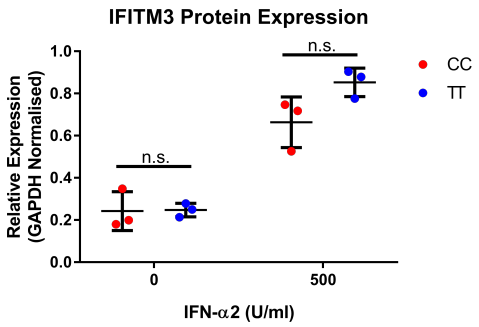
4.2.1 Examination of *IFITM3* Expression Relative to rs12252 Using Primary Human Fibroblasts

Initially, we decided to use primary human fibroblasts as a model to study the effects of rs12252 on *IFITM3* expression. These cells were partially collected in Oxford, but were overwhelmingly obtained with help of Dr. Yan Zhao at Beijing You'An Hospital.

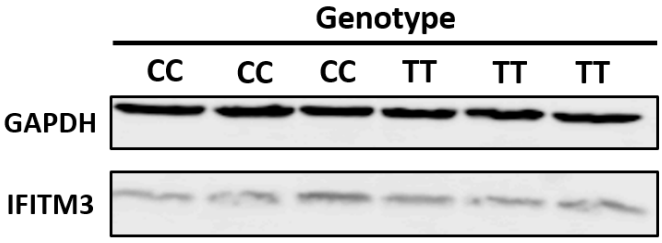
Upon obtaining sixteen cell lines (with three CC and TT homozyogotes), we profiled the expression of *IFITM3* via western blot and qPCR. We used 16 fibroblasts total for qPCR, and 6 for western blot. In each case, we had 3 homozygous fibroblasts of each genotype. We stimulated each line with IFN-α2 (500U/ml, 24hrs). We were unable to observe any significant expression level differences at the mRNA or protein level (fig. 4.1) at baseline or upon stimulation.



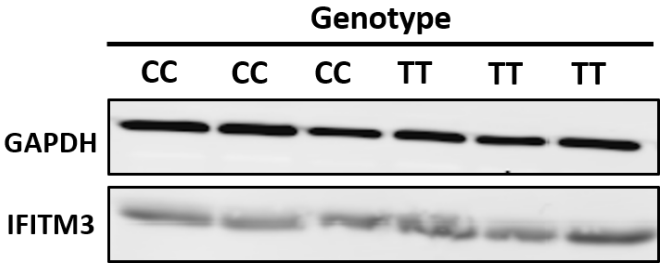
(a) *IFITM3* Expression



(b) IFITM3 Protein Expression



(c) IFITM3 Western Blot of Untreated Fibroblasts



(d) IFITM3 Western Blot of IFN Stimulated Fibroblasts

Figure 4.1: ***a.*** *IFITM3 expression was profiled in primary human fibroblasts using qPCR, and was compared with respect to rs12252. With 16 cell lines in total, and 6 homozygous fibroblasts, we could not observe any difference in IFITM3 expression at the mRNA level. Error bars show mean and SD.* ***b.*** *Quantified IFITM3 protein expression was profiled in primary human fibroblasts using western blot. We could not observe any expression differences with regards to rs12252 genotype using 3 homozygous cell lines of each genotype.* ***c and d.*** *Western blots showing IFITM3 protein expression in CC and TT fibroblasts under untreated and IFN stimulated conditions. We could not observe any difference in expression between genotypes.*

4.2.2 rs12252 eQTL Analysis Using Public Databases

Analysis of *IFITM3* Expression Relative to rs12252 Using eQTL Databases

As a method of accessing a wider range of donor samples, we decided to investigate whether or not rs12252 was considered to be an eQTL by available databases. In testing this, we used the ExSNP database, a database of 16 eQTLs, and the GTEx database (reviewed in Appendix C.3). In the former, we could not observe any associations between rs12252 and *IFITM3* expression.

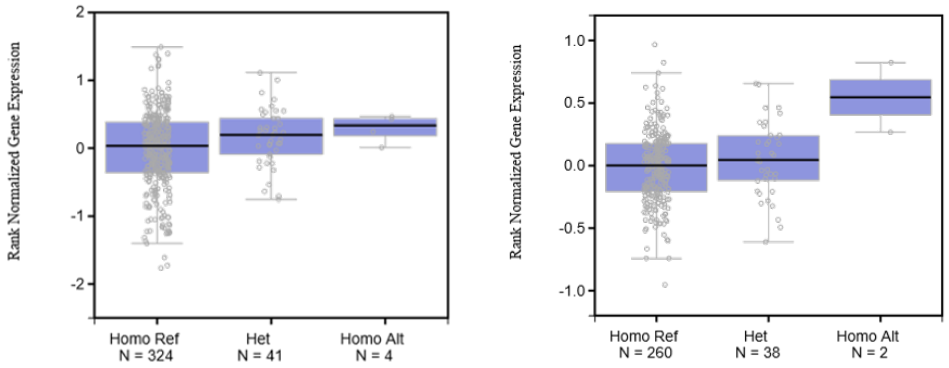
Table 4.1 eQTLs for rs12252 and IFITM3

Tissue	**Effect Size**[1]	**p-value**
Whole Blood	0.26	0.013
Lymphocytes	0.11	0.64
Liver	-0.077	0.36
Lung	-0.047	0.42
Fibroblasts	0.14	0.017
Spleen	-0.0032	0.98

[1] Positive effect size indicates higher expression of rs12252-C

We searched the GTEx database for associations between rs12252 and any gene within the genome - this returned two significant eQTLs: *protein-glucosylgalactosylhydroxylysine glucosidase (PGGHG)* in the Frontal Cortex, and *Plakophilin-3 (PKP3)* in the Thyroid (fig. 4.2). These genes are within 100kbp of *IFITM3* and rs12252. In these cases, the C allele was associated with reduced expression of *PGGHG* (effect size = -0.55, p = 0.000012) and increased expression of *PKP3* (effect-size = 0.30, p = 0.000031). The distribution of genotypes was biased in favour of the T allele - each eQTL had only two samples from rs12252-C homozygotes.

We also profiled the effects of rs12252 on the expression of *IFITM3* in a multitude of tissues: whole blood, EBV-transformed lymphocytes, liver, lung, transformed fibroblasts, and spleen (table 4.1). In two of these tissues - whole blood and transformed fibroblasts - the presence of rs12252-C was associated with higher expression of *IFITM3*. Again, there was a lack of rs12252-C homozygotes, casting some doubt onto the validity of the obtained results. Donor demographics within this cohort showed that merely 1.1%, or



(a) Whole Blood (b) Transformed Fibroblasts

Figure 4.2: Boxplots showing eQTL data from the GTEx database, examining IFITM3 expression in relation to rs12252. Expression is plotted in terms of rank normalised gene expression. In the above figures, the reference allele (Ref) is T and and the alternate allele (Alt) is C.

8 donors were of East Asian descent in this database, explaining the lack of the rs12252-C allele.

4.2.3 Further Investigation of rs12252 as an eQTL using Primary Human Fibroblasts

Analysis of Allelic Expression of rs12252 Heterozygote Samples

In the absence of compelling evidence for differential expression of *IFITM3*, likely with regards to the fact that we had limited sample numbers, we decided to approach this problem using allelic expression as a surrogate for overall gene expression. Allelic expression refers to the individual expression levels, or ratios between of each allele in an individual sample, generally in heterozygote donors. In these heterozygotes, expression preferences for a single allele can indicate a reduced level of expression associated with the allele in lower abundance.

We used RNA-Sequencing data to examine whether or not allelic expression effects were present in rs12252 heterozygotes. We used H1N1pdm2009 patients with mild influenza infection described in Appendix A.1. This cohort contained 6 CT donors. These patients showed a clear bias towards

production of the T allele in five of the six patients. On average, 55.7% of *IFITM3* transcripts produced by these donors contained the T allele (fig. 4.4).

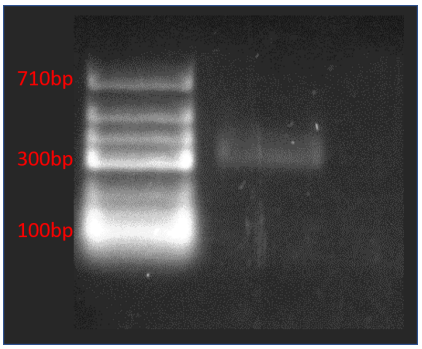
Subsequently, we obtained data from the ExAC database [242]. We obtained alignment data from ten different heterozygous samples. The nature of ExAC database samples is explained in Appendix C.2. In this data set, 9 out of 10 samples examined expressed a higher level of the T allele than the C allele. On average, donors in this dataset produced 60.8% of their transcripts with T alleles, as opposed to 39.2% with C alleles (fig. 4.4).

Lastly, we developed a targeted RNA-Sequencing protocol to examine the splicing profile of cultured cells. We designed primers specific to the rs12252 locus (rs12252_MiSeq; Appendix B), with Illumina specific overhangs. We optimised these primers with sample cDNA to ensure specificity and adequate amplification. These primers specifically amplified the rs12252 locus (fig. 4.3a). Subsequently, we sequenced the resultant amplicon; this sequence aligned to the *IFITM3* locus with no mismatches. In addition, to account for the potential of PCR bias in the amplification of *IFITM3* cDNA, we performed control reactions with gDNA carrying the CT genotype, based on the fact that gDNA will carry C and T alleles in equal quantities. These primers specifically amplified the rs12252 locus using gDNA as a template.

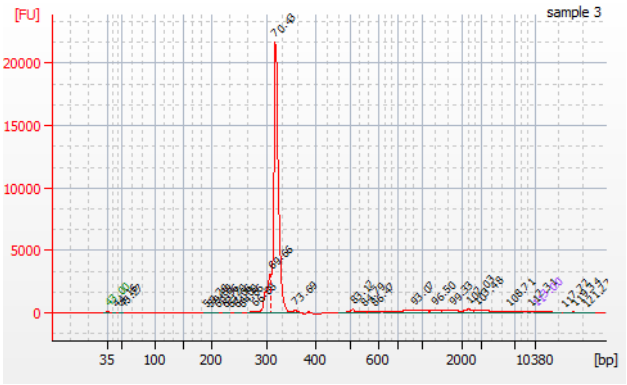
We used 10 different human skin fibroblast cell lines, all of which had the CT genotype. Cells were treated with 500U/ml IFN-α2 for 24 hours, or left untreated. After purification of individual sample libraries, we analysed samples via the Bioanalyser 2100 system. The produced libraries were highly uniform with single peaks as expected. Amplicons were roughly 310bp, the expected sample size. (fig. 4.3b).

With these samples, we observed a bias towards amplification of the C allele in gDNA samples - 48.1% of counted reads contained the T allele. Despite this, we observed overexpression of the T allele in this dataset. In untreated fibroblasts, 9 of the 10 samples preferentially expressed the T allele, with an mean T allele expression of 60.8%. Upon IFN stimulation, all samples preferentially expressed the T allele, with a mean T allele expression of 55.8% (fig. 4.4).

Overall, the results support the idea that the T allele is preferentially expressed in rs12252 heterozygotes.

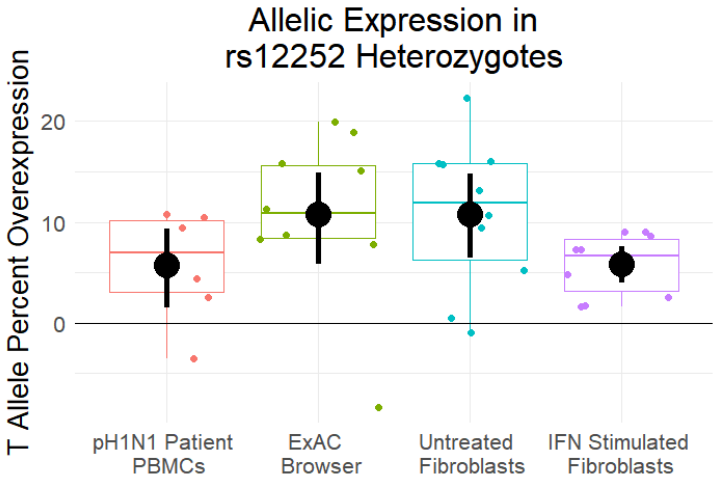


(a) rs12252 Locus PCR

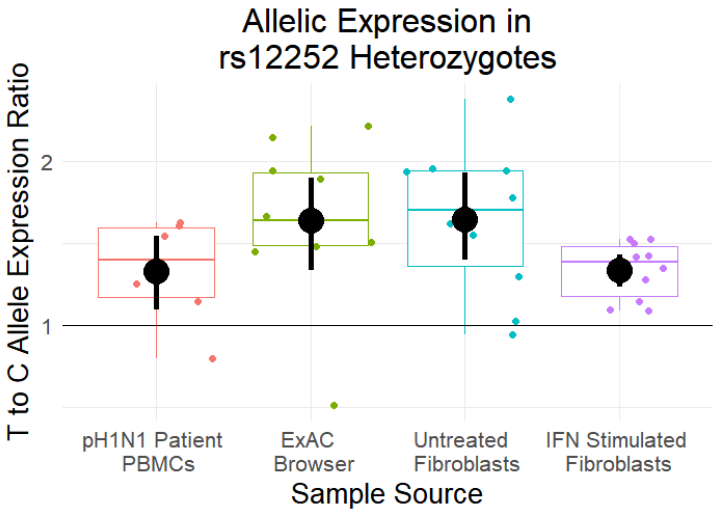


(b) Example Bioanalyser Run

Figure 4.3: ***a.*** *DNA gel showing specific amplification of targetted RNA-Sequencing amplicon from cDNA. Band is roughly 310bp, the expected length of the amplicon.* ***b.*** *Example Bioanalyser 2100 elecropherogram of targetted RNA-Sequencing individual sample library. X axis represents DNA length, and y axis indicates abundance of transcript. Library is highly uniform, with a single peak.*



(a) Allelic Expression (Percentage)



(b) Allelic Expression (Ratio)

Figure 4.4: ***a.*** *Allelic expression of rs12252, as shown by plotting T allele expression percentage over 50%. Graph shows data from pandemic H1N1 donors, the ExAC database, and primary human skin fibroblasts. Black points represent mean and SEM.The T allele is consistently overexpressed, with mean overexpression falling roughly between 55-60% expression.* ***b.*** *Allelic expression data shown as the T to C expression ratio.*

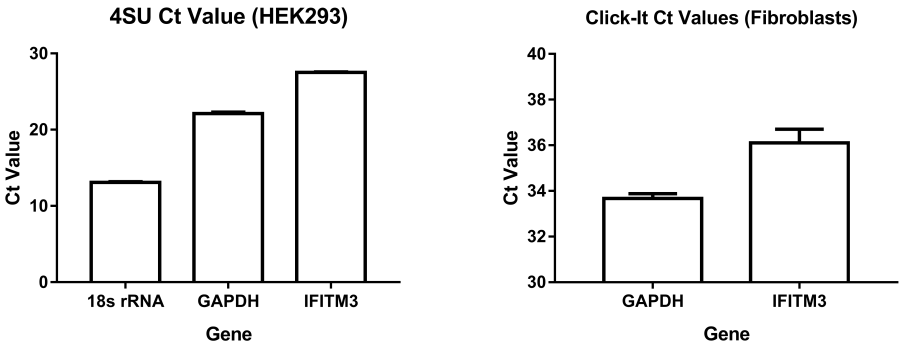
Analysis of Allelic Expression in Recently Transcribed *IFITM3* Transcripts

To further investigate our previous findings regarding the allelic imbalance in rs12252 heterozygotes, we decided to investigate the transcriptional rate of *IFITM3* with regards to rs12252 allele. We pulled down nascently transcribed *IFITM3* mRNA, before repeating our previously developed targeted RNA-Sequencing protocol. We hypothesised that in recently transcribed mRNA we would be able to observe the same differential allelic expression observed when using total RNA.

We initially developed a protocol based on incorporating 4-Thiouridine (4SU), a uridine analogue, into mRNA. We used HEK293 cells alongside fibroblast cells, given the fact that they were generally easier to obtain large volumes of RNA from. We treated cells with 4SU for various time windows, before pulling down RNA using methods described in section 2.2.8. We used numerous concentrations and incubation times for 4SU. Based on pre-existing literature, we decided to use 300µM 4SU [312]. We optimised this concentration to 50µM, based on reports that high concentrations of 4SU can interfere with transcription [313]. We also changed our RNA extraction protocol from the RNEasy kit to TRIzol extraction - this allowed us to obtain greater quantities of RNA, given the upper limit of column based approaches. We also changed our method of streptavidin pulldown from MyOne Streptavidin T1 beads (Thermo Fisher Scientific) to the µMACS Streptavidin Kit (Miltenyi Biotec).

Eventually we pulled down nascent mRNA in HEK293 cells. From an initial quantity of 100µg of RNA, we obtained roughly 850ng of RNA - less than 1% of the initial quantity. This yield is far less than expected, suggesting that this methodology was still inefficient. With this mRNA, we prepared cDNA, and subsequently performed qPCR to ensure that the nascent mRNA could be used for amplification. As expected, we were able to amplify *18S rRNA (RNA18S)*, *GAPDH* and *IFITM3* genes, with Cycle Threshold (Ct) values close to what would be expected in HEK293 cells (fig. 4.5a). However, we were unable to obtain RNA using this methodology with fibroblast cell lines.

We decided to change our strategy to using the Click-It Nascent RNA capture kit. We found that this kit was effective at pulling down RNA using lower initial RNA quantities - this yielded a mean pulldown percentage of roughly 30%. We tested this RNA with qPCR, ensuring that genes were expressed at expected levels. We observed normal amplification of *GAPDH* and *IFITM3*



(a) qPCR Test: 4SU Pulldown

(b) qPCR Test: Click-It

Figure 4.5: ***a.*** *Test qPCR reaction showing normal amplification of nascently transcribed cDNA from HEK293 cells. X axis represents raw Ct values.* ***b.*** *Test qPCR reaction showing normal amplification of nascently transcribed cDNA from primary human fibroblasts*

(fig. 4.5b).

In amplifying the rs12252 locus from cDNA, we found that the previously used 25 cycles was not sufficient. For this reason, we tested the protocol at 28 and 35 cycles (fig. 4.6). In both cases, we could amplify the rs12252 locus. We used 28 cycles in our library preparation protocol in the interest of maintaining as much parity as possible with the original protocol. After library preparation, as with total RNA, we again obtained highly uniform libraries.

After analysis, we observed a slight bias in nascent expression favouring the T allele. In untreated fibroblasts, there was a mean T allele expression of 54.5%, representing a reduction in T allele expression compared to total RNA. In IFN stimulated fibroblasts, there was a mean T allele expression of 55.1% - this was roughly equivalent to our observations made with total RNA (fig. 4.7). In our gDNA controls, we observed a mean T allele expression of 45.1% - this represented a reduction in T allele expression compared to our experiment with total RNA, and may explain the reduced preferential allelic expression observed here.

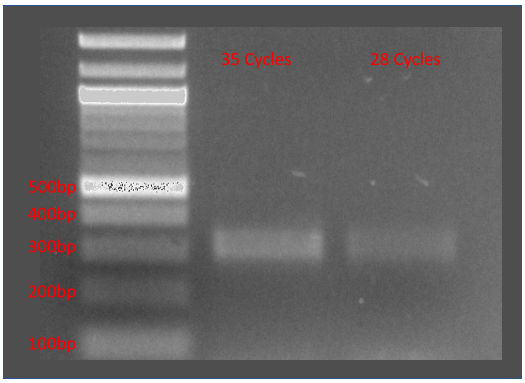


Figure 4.6: DNA gel showing specific amplification of rs12252 locus from nascent RNA. Amplification could be observed at 28 and 35 cycles, with greater band intensity at 35 cycles, as expected.

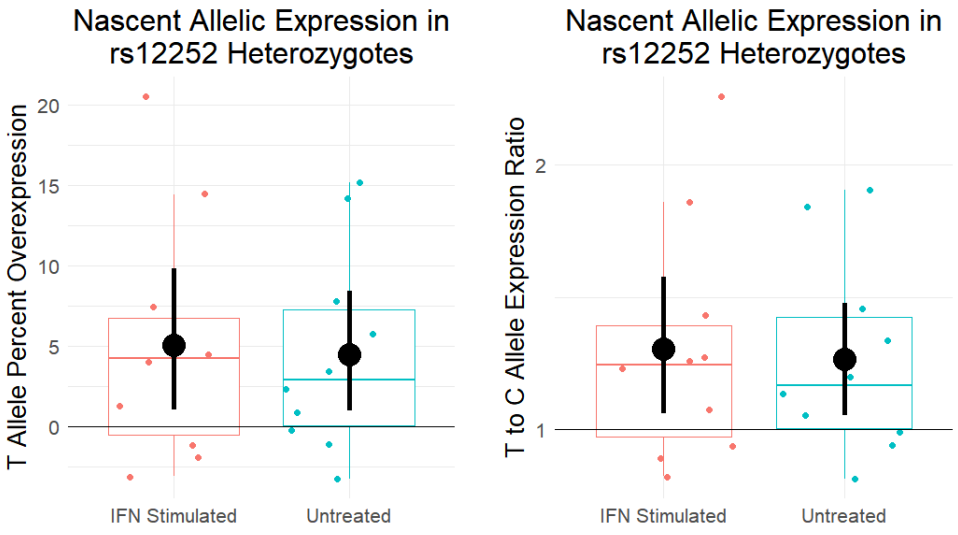


Figure 4.7: Nascent IFITM3 mRNA shows preferential allelic expression, again favouring the T allele over the C allele. This effect is not as strong as when observed in total RNA. Data is shown as percent expression (left) and T to C allele ratio (right). Black points represent mean and SEM.

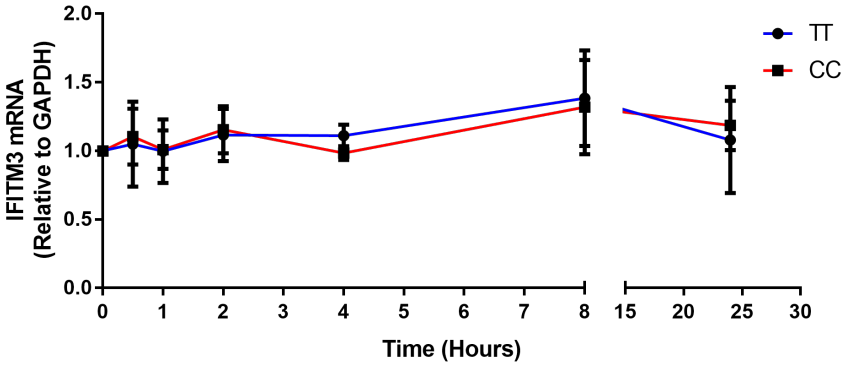


Figure 4.8: Time course of IFITM3 mRNA degradation using GAPDH mRNA as a control transcript. In each sample, IFITM3 expression relative to GAPDH was normalised to 1 at time 0. We observed no degradation of IFITM3 transcripts using this system, and no difference in relative expression to GAPDH over time with respect to rs12252 genotype.

Analysis of Degradation rate of *IFITM3* With Respect to rs12252 Allele

We also decided to investigate the degradation rate of *IFITM3* with respect to rs12252 genotype. We used an Actinomycin D (ActD) assay - this assay is based on the fact that cells treated with ActD have their transcription suppressed. We established a timecourse of 24 hours where we could observe the degradation of *IFITM3* transcripts present at time 0. After 24 hours, cells had died due to the toxic nature of ActD. We used a stable housekeeping transcript, *GAPDH* [314], as a reference to measure *IFITM3* degradation from.

As a control, we incubated samples with IFN-α2 for 8 hours under the presence of ActD. This control was designed to test the efficacy of ActD - if *IFITM3* was induced by IFN while being treated by ActD, it would suggest that ActD was not sufficiently suppressing transcription in the treated cells. After comparing *IFITM3* expression between the ActD treated and IFN + ActD treated cells at 8 hours, we saw a negligible increase in average *IFITM3* Ct value (0.238) in the ActD treated cells compared to the ActD + IFN treated cells (data not shown). This confirmed to us that the added ActD was sufficiently suppressing transcription.

After performing this assay we normalised all data points using GAPDH Ct

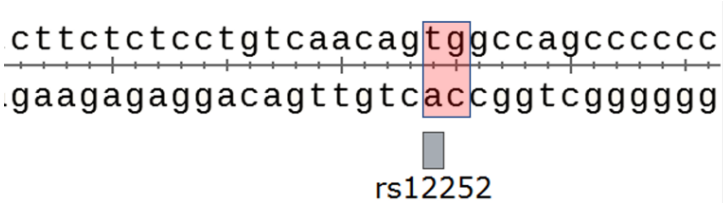
values. We then normalised this data against the 0 time point, setting this time point to 1 across all samples. We found that relative levels of *IFITM3* rose slightly over the timecourse. This suggests that *IFITM3* mRNA likely has a slightly longer half-life than *GAPDH* mRNA. It is likely that *IFITM3* mRNA is highly stable, and cannot be studied sufficiently using this system. We did not observe any differences in degradation rate as compared between TT and CC cell lines.

Analysis of Methylation State of rs12252

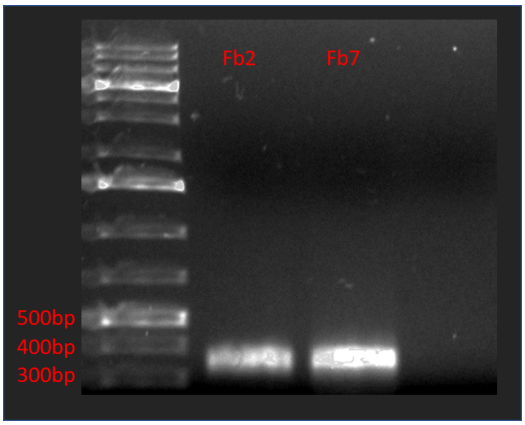
The T to C change incurred by rs12252 in *IFITM3* results in the creation of a CpG site, or a potential methylation site (fig. 4.9a). We hypothesised that the methylation of rs12252 may explain the apparent preference observed in heterozygous cell lines for the T allele.

We choose to look at the abundance of methylation of the rs12252 locus in primary skin fibroblasts with the CC genotype. This was based on the fact that we observed allelic expression differences in CT skin fibroblasts, suggesting that rs12252-linked mechanisms driving reduced expression of *IFTIM3* would be present in these cell lines. We performed bisulfite conversion on gDNA derived from the CC cell lines.

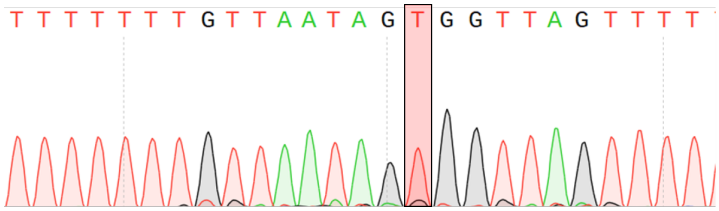
We developed a direct sequencing PCR protocol for amplification of the bisulfite treated rs12252 locus (fig. 4.9). We determined specificity of these primers via gel electrophoresis. We cloned amplicons from three different skin fibroblast lines into pCR4-TOPO vectors via TOPO-TA cloning. We performed 10-20 cloning reactions per cell line, and in all cases, were unable to find any evidence of methylation occurring at the rs12252 locus (data not shown). While this does not totally exclude the possibility that in some tissues methylation of this locus occurs, it is clear from this result that methylation of rs12252 does not govern the observed reduction of *IFITM3* expression in totality.



(a) Potential Methylation Site of rs12252

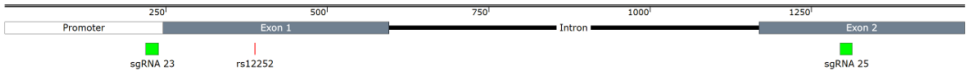


(b) Methylation Amplicon

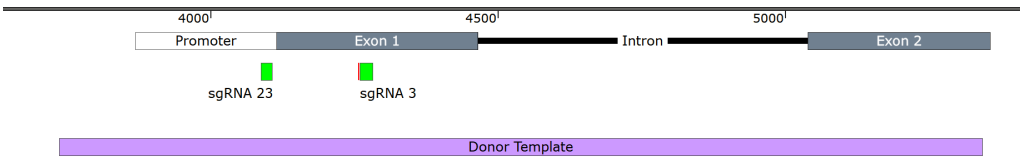


(c) Methylation Sequencing Example

Figure 4.9: ***a.*** *Figure demonstrating rs12252-C producing a viable methylation site.* ***b.*** *DNA Gel image showing specific amplification of the rs12252 locus within IFITM3 after bisulfite conversion. Bands are indicative of gDNA from two different fibroblasts used as a source.* ***c.*** *Figure showing an example sequencing trace where the methylation state of rs12252 can be determined. rs12252 locus is highlighted in red. In this case, rs12252 is not methylated.*



(a) Guide RNA Locations for *IFITM3* Knock-Out



(b) Guide RNA/Donor Construct Locations for rs12252 Site-Directed Mutagenesis

Figure 4.10: ***a.*** *DNA map showing the locations of guides sg23 and sg25 in reference to the IFITM3 locus. These sgRNAs were used to generate IFITM3 knock-out cell lines.* ***b.*** *DNA map showing locations of guide RNAs and donor construct used for site-directed mutagenesis of rs12252 in reference to IFITM3 locus.*

4.2.4 Analysis of rs12252 as an eQTL using CRISPR-Cas9

Editing of *IFITM3* locus with CRISPR-Cas9 Mutagenesis

We took advantage of the Cluster of Regularly Interspersed Short Palindromic Repeats (CRISPR) - CRISPR Associated Protein 9 (Cas9) system to investigate the effects of rs12252 on an isogenic background.

We also decided to knock *IFITM3* out on the same background using CRISPR-Cas9. We reasoned that knock-out cell lines would provide a valuable control in a wide range of experiments. We decided to generate these cell lines on a HEK293 and A549 cell background. Our rationale for using HEK293 cells was based on the fact that growing single-cell clones of these cells is relatively simple and well described. These cells also express *IFITM3* without IFN stimulation. A549 cells were chosen due their origins from lung carcinoma cells. These are immortalised alveolar epithelial cells, a cell type naturally infected by IAV. We genotyped both of these cell lines for rs12252; these cells both carry the TT genotype (data not shown).

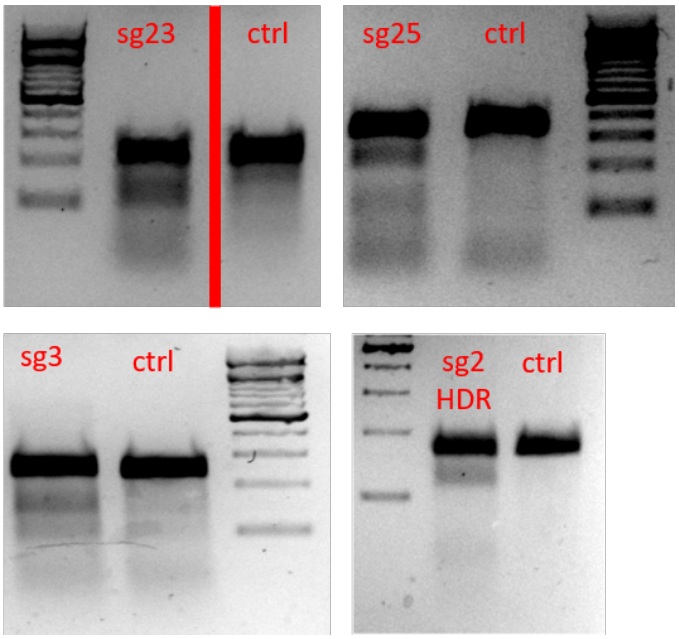


Figure 4.11: Surveyor assays showing cutting activity for all sgRNAs used. In all cases, application of sgRNA showed cutting of its constitutive locus. In each case a control reaction was run without sgRNA treatment - these lanes showed no cutting activity. All surveyor assays were performed by Dr. Philip Hublitz.

Guide RNA design, cloning and testing was performed by Dr. Philip Hublitz. To produce *IFITM3* KOs, a dual sgRNA strategy was used, whereby two different loci within *IFITM3* were targeted (sg23, sg25, Appendix B) (fig. 4.10a). For site directed mutagenesis, we also used a dual sgRNA strategy (sg23, sg3, Appendix B), to be cotransfected with a large 1606bp donor construct (fig. 4.10b). This construct was homologous to *IFITM3*, with the exception of the rs12252 locus, which was converted to a C, and the Protospacer Adjacent Motif (PAM) of sg3, which was converted to a EcoRV target site. The alignment of this donor template extended past the target sites of both sgRNAs. Endonuclease activity of sgRNAs was tested on a HEK293T background via surveyor assay (fig. 4.11).

We obtained three or more KO clones for both A549 and HEK293 cell lines (fig. 4.12, data not shown for A549 cells) via western blot and DNA sequencing. We observed differing indel mutations, depending on the cell lines. We were unable to successfully generate any positive hits for successful incorporation of rs12252-C on these cell backgrounds. After screening a wide array of different cell clones, we decided to change our CRISPR-Cas9 strategy.

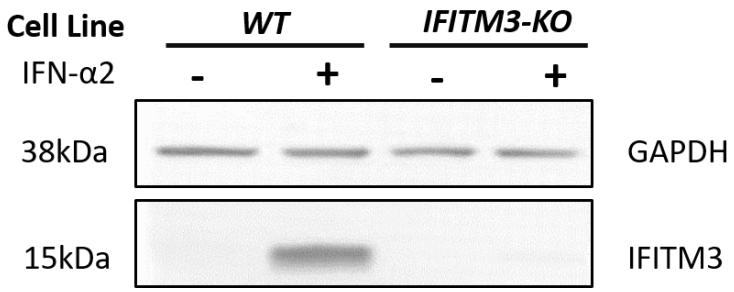


Figure 4.12: Western blot showing successful knock-out of IFITM3 from a HEK293 cell background. WT HEK293 cells and potential KO cells were stimulated with IFN-α2 before running as a western blot.

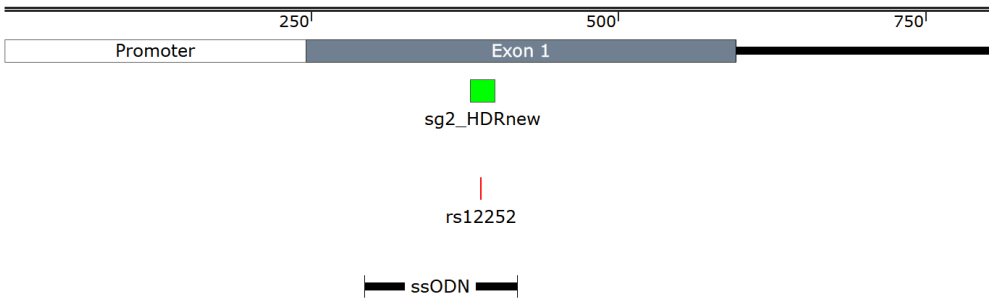
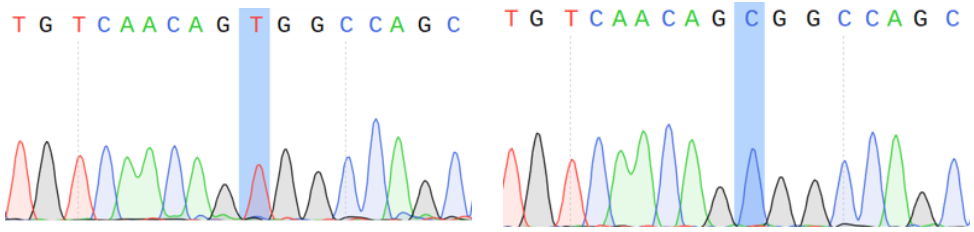


Figure 4.13: DNA map showing the location of sg2_HDR_v2 and ssODN construct in the context of the IFITM3 locus. sg2_HDR_v2 and ssODN were used in a revised strategy to introduce rs12252 to cell lines via site-directed mutagenesis.



(a) Wild-Type (b) rs12252 Knock-In

Figure 4.14: ***a and b.*** *Sequencing traces showing successful homozygous insertion of the C allele onto a HEK293 cell background. rs12252 SNP is shown highlighted in blue.*

After failing to perform site-directed mutagenesis of rs12252 with the CRISPR-Cas9 system, we reapproached our strategy. This started roughly 2 years after we originally decided to generate cell lines with CRISPR, and was based on advances in CRISPR techniques favouring HDR. We used only one sgRNA in this instance, developed with help from Dr. Philip Hublitz, (sg2_HDR_v2, Appendix B). This sgRNA was highly specific to the rs12252 locus (fig. 4.13), but showed 1 mismatch to a region of the short arm of chromosome 12. This locus is intergenic, and does not have strong Deoxyribonuclease (DNase)-I hypersensitivity. Additionally, this sgRNA had two mismatches to the *IFITM2* locus.

We used an ssODN template, which was synthesised commercially, and was designed to be thiolated at 5' and 3' ends to prevent degradation within the cell. The sgRNA was cloned into the px458 vector containing an eGFP marker. A surveyor assay was perfomred by Dr. Philip Hublitz, showing positive cutting activity of the sgRNA (fig. 4.11). In transfection, RS1 was added as an HDR-enhancing drug in light of reports that this drug was more effective than SCR7 [315].

Out of roughly 100 clones, one positive clone was found initially. There was, a degree of background in the sequencing trace used for screening. For this reason, we performed PCR over the rs12252 locus, and cloned the product into the pCR4-TOPO vector using TOPO-TA cloning. Ten of the resultant colonies were picked and sequenced. This showed a mix of T and C alleles at the rs12252 locus, indicating either heterozygosity for rs12252, or a population containing a mix of different clones. To address this, we performed single cell dilution of the clonal population. After growth of the resultant clones, we again sequenced over the rs12252 locus, this time

obtaining clean sequencing traces with clear homozygosity for the C allele (fig. 4.14). We additionally checked this clone for off-target effects of the sgRNA by sequencing over the potential off target locus on *IFITM2*. This returned to show no damage to the integrity of *IFITM2*, indicating that there were no IFITM2 related off target effects.

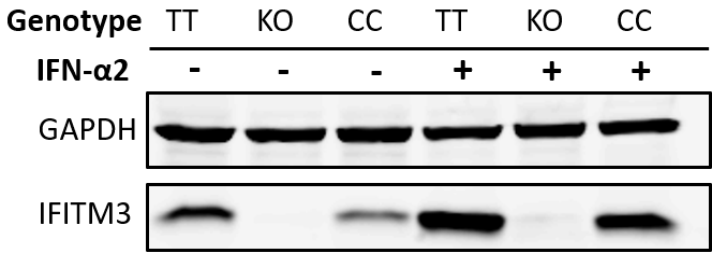
Analysis of Expression of Edited HEK293 Cells

After using the CRISPR-Cas9 system to introduce rs12252-C into the HEK293 cell background, we examined expression levels of *IFITM3* before and after IFN stimulation. We used western blot to probe the expression of IFITM3 protein in wild-type and mutant versions of the cell line. We performed three repeats of this western blot, examining *IFITM3* expression without treatment, and with 100U/ml IFN-α2 treatment for 24 hours. After western blotting, we performed image analysis using FIJI to quantify band intensity.

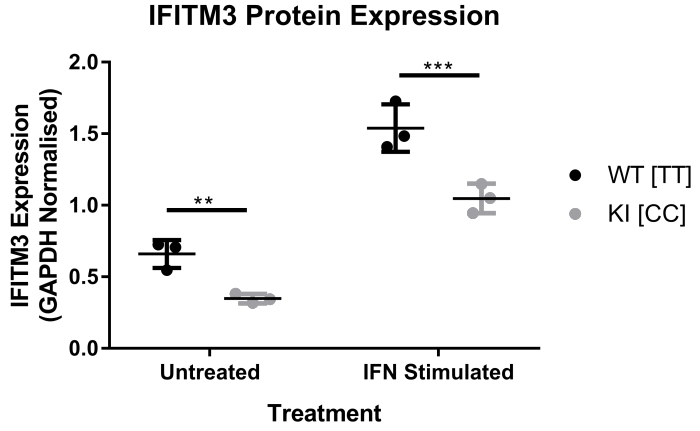
In untreated cells, the rs12252-C containing cells showed a reduction in IFITM3 expression compared to the wild type (fig. 4.15). This was statistically significant ($p = 0.0085$); the WT cells showed a mean 1.90 fold expression increase in IFITM3 when compared to the CC cells. In IFN treated cells, CC cells again showed a statistically significant reduction in IFITM3 expression ($p = 0.00060$). WT cells showed a mean fold-change increase of 1.47 compared to the CC cells.

We also used Quantitative PCR (qPCR) to compare mRNA levels of *IFITM3*. This analysis was performed with three technical replicates, and was performed three times total. In each case, the expression level of *IFITM3* mRNA in the CC HEK293 clone was reduced both in unstimulated and IFN stimulated cells. Differing Ct values for *GAPDH* meant that aggregated experiments did not reach statistical significance. However, in each experiment, the difference in expression was consistently statistically significant (via t-test). We also examined the fold change between CC and TT cells in each experiment. Here, we could observe highly consistent differential expression of *IFITM3* - the TT genotype was associated with a mean fold change in *IFITM3* expression of 1.57 (untreated) and 1.90 (IFN stimulated).

We also examined expression levels of *IFITM1* and *IFITM2* and *STAT1* in our edited HEK293 cells. We observed a modest but consistent

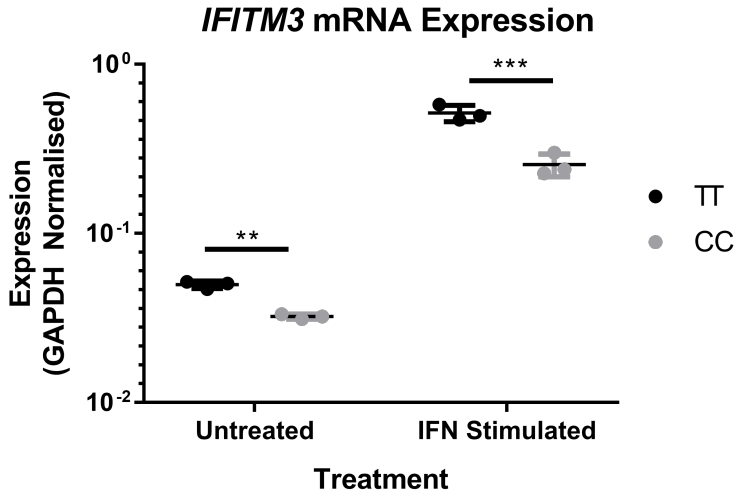


(a) IFITM3 Western Blot in HEK293 Cells

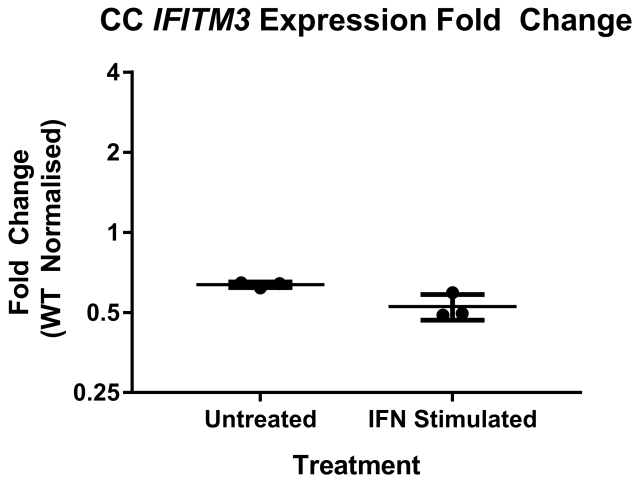


(b) IFITM3 Expression Quantitication in HEK293 Cells

Figure 4.15: ***a.*** *Example western blot showing IFITM3 expression reduction in CC HEK293 cells when compared to the WT counterparts. Western blot is shown with IFITM3-KO cells as a negative control.* ***b.*** *Quantification of IFITM3 expression in three repeats of the western blot shown above. IFITM3 expression is significantly reduced in CC cells at basal level and after IFN stimulation. Error bars show mean and standard deviation*



(a) *IFITM3* qPCR KI vs WT



(b) Aggregate Fold-Change in Over 3 Experiments

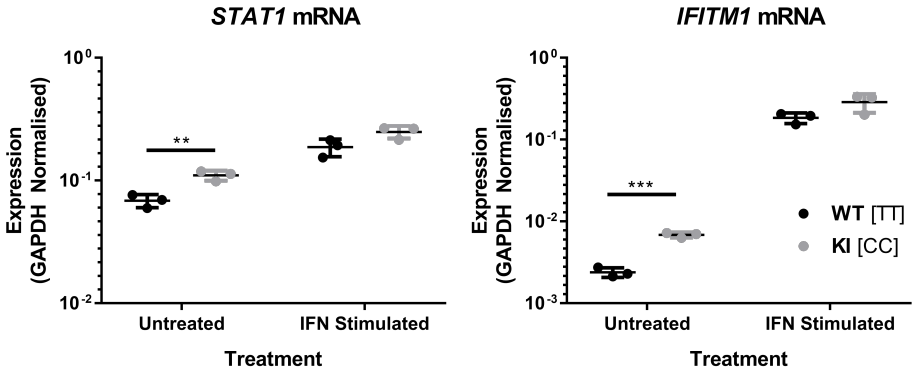
Figure 4.16: ***a.*** *Example run of IFITM3 qPCR. Technical replicates are shown individually. Difference in IFITM3 expression reaches statistical significance.* ***b.*** *Aggregate fold change in expression over three experimental qPCR repeats. WT expression is normalised to 1 in each condition. Error bars show mean and standard deviation for both figures.*

increase in expression of *IFITM1* and *STAT1* in the HEK293 cells with the CC genotype. Over three experiments, *IFITM1* expression was significantly differentially expressed consistently in untreated cells, but was not significantly differentially expressed in IFN-treated cells (via t-test). On average, CC cells were associated with mean fold changes of 2.27 and 1.62 in untreated and IFN treated cells respectively. Like *IFITM1*, *STAT1* was consistently expressed at a significantly higher level in untreated CC cells, but was not in IFN-treated cells. In this case, CC cells were associated with mean fold changes of 1.48 and 1.36 in untreated and IFN treated cells respectively (fig. 4.17a).

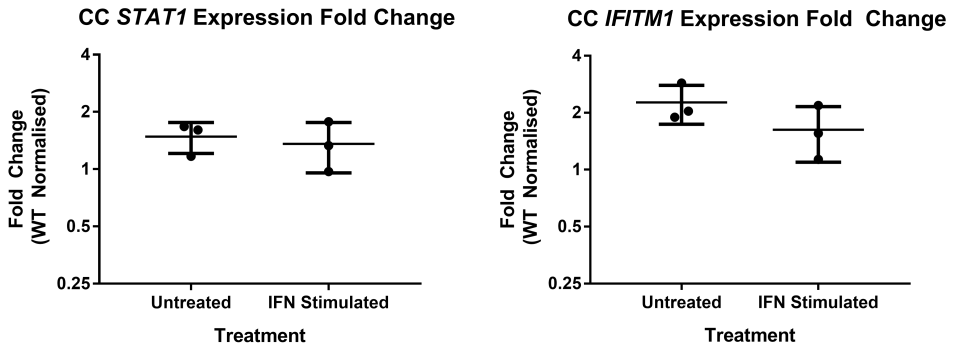
When examining *IFITM2*, we found only minor induction of this gene by type-I IFN, in contrast to the other 3 ISGs tested, although basal expression was higher in CC cells (data not shown). This observation was consistent between two *IFITM2* probe sets.

4.2.5 Analysis of rs12252 In Healthy Donor Cohort

Finally, to examine the effects of rs12252 on *IFITM3* expression, we used a large group of donors. Studies using this cohort were approved by the local ethics committee of Beijing You'An Hospital, and the University of Oxford Tropical Ethics Committee, under OXTREC 1001-13. These volunteers were generally aged between 18 and 40, with a single donor aged 63. The median age of the full cohort was 27. Donor age was uniformly distributed between all genotype groups, patient age was not associated with genotype in this cohort, as confirmed by multiple Kolmogorov–Smirnov tests between age distributions ($p = 1$ in all cases, after Bonferroni correction). All donors were healthy when samples were taken. Demographics are shown in more detail in table 4.2.



(a) *STAT1* and *IFITM1* qPCR ()KI vs WT)



(b) Aggregate Fold-Change Over 3 Experiments

Figure 4.17: ***a.*** *Example run of STAT1 and IFITM1 qPCR. Technical replicates are shown individually. Difference reaches statistical significance in untreated cells, but not in IFN treated cells.* ***b.*** *Aggregate fold change in expression over three experimental qPCR repeats. Here, WT expression is normalised to 1 in each condition. ISGs are found at higher expression levels in CC cells compared to TT. Error bars show mean and SD in all cases.*

Table 4.2 Healthy Donor Demographics

Genotype	Number	Percent Total
Male donors		
CC	18	24.3
CT	43	58.1
TT	13	17.6
Total	74	N/A
Female donors		
CC	61	30.5
CT	105	52.5
TT	34	17.0
Total	200	N/A
All donors		
CC	79	28.8
CT	148	66.1
TT	47	17.2
Total	274	N/A

Donor samples were obtained in the form of whole blood. RNA was isolated and sent to Beijing Genomics Institute (BGI) for sequencing. cDNA libraries were prepared from poly(A) enriched RNA. Final libraries contained paired end, 100bp reads. Sequencing was performed on a BGISEQ-500 platform. Final sequencing depth was 50-60 million reads per sample. Quality control is shown in Appendix D.2

Analysis of *IFITM3* Expression

IFITM3 expression was compared between CC and TT genotypes. *IFITM3* was differentially expressed with regards to rs12252, with higher expression associated with the CC genotype, as opposed to the hypothesised TT genotype (fig. 4.18). In this group, CC donors had, on average, a 1.5 fold higher level of *IFITM3* expression when compared to the TT donors. This difference was found to be highly significant, with a p-value of 6.8×10^{-5}.

In addition to testing for differences between homozygous groups, we constructed an additive model of expression, whereby, within each donor, each incidence of a T allele was equivalent to 1 point on a continuous scale (CC = 0, CT = 1, TT = 2). Here, we found that *IFITM3* expression was related to rs12252 genotype in an additive manner. *IFITM3* expression increased by a factor of 1.21 per C allele. This relationship was significant,

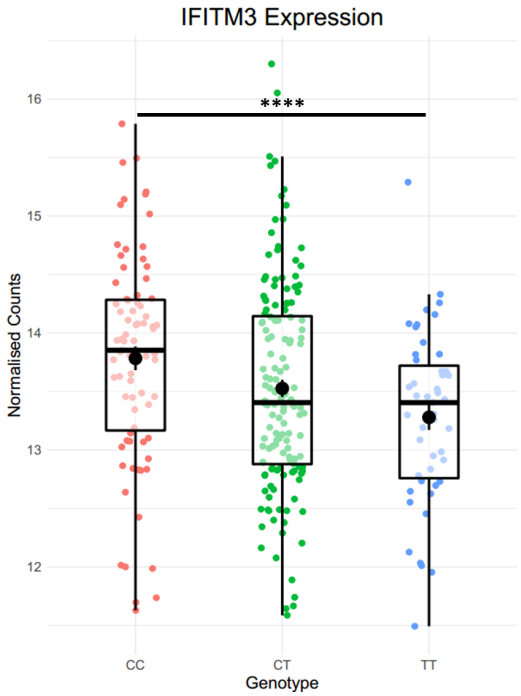
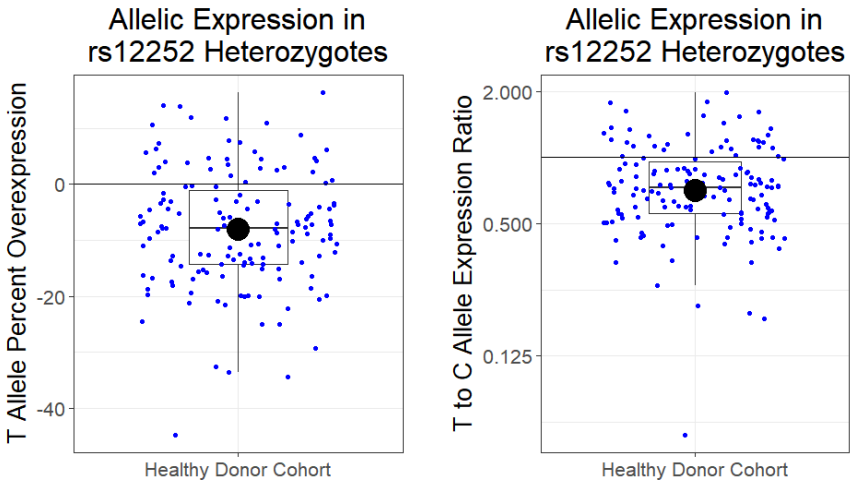


Figure 4.18: Average IFITM3 expression was found to be higher in CC donors than TT donors. IFITM3 expression is shown as normalised counts after application of a variance stabilising transformation. Black points represent mean and 95% SEM Differential expression is highly significant between homozygotes (via Wald test), and under an additive model of gene expression (via linear regression based model).



(a) T Allele Overexpression (b) T to C Ratio

*Figure 4.19: Dot plots showing allelic expression of rs12252 in the healthy donor cohort. Allelic expression is shown as percent T allele expression (**a**) and T to C allele ratio (**b**). Black points represent mean and SEM. Donors showed a clear allelic expression bias, with the majority of donors expressing higher levels of the C allele than the T allele. Spread of points on the x-axis is arbitrary.*

with a ($p = 0.00027$). CT donors had a median expression level roughly equal to TT donors - however, TT donor mean expression level was significantly lower than CT donor mean expression.

Allelic Expression Analysis of rs12252 Heterozygotes

We subsequently analysed the allelic expression of rs12252 in the 104 heterozygous donors to further investigate our previous findings with regards to allelic expression. In this case, we found that of the 113 heterozygous donors investigated, only 34 of them preferentially expressed the T allele, as opposed to 83 of them that preferentially expressed the C allele. The mean percentage T allele expression accounted for 41.9% of total *IFITM3* expression (fig. 4.19a); there is a 1.29 fold greater expression of the C allele on average (fig. 4.19b). These results were consistent with the rest of this cohort in that the C allele was associated with higher *IFITM3* expression.

Cis-eQTL Analysis of rs12252

We next used cis-eQTL analysis to first examine local gene expression effects of rs12252. We reasoned that rs12252 would be more likely to be associated with local gene expression dynamics than distal ones, given our findings related to rs12252 and *IFITM3* expression. With cis-eQTL analysis, we could use a less strict multiple-testing correction when compared to whole genome differential gene expression analysis, due to the smaller number of statistical tests being made. This analysis was performed with a maximum distance of 1Mbp. This amounted to a roughly 1.3Mbp window of inclusion for genes on chromosome 11, as rs12252 is 0.3Mbp from the end of chromosome 11. We accounted for the sex and age of the donors by including these parameters as covariates of the dataset. Overall, 81 different genes were used for testing. We used an initial p-value threshold of 0.1.

Table 4.3 Significant cis-eQTLs for rs12252

Gene Name	**p-value**	**p-value (Adjusted)**
AC136475.1	1.44×10^{-28}	1.01×10^{-26}
AC136475.5	1.73×10^{-9}	6.06×10^{-8}
AC136475.9	2.80×10^{-9}	6.54×10^{-8}
AC136475.8	1.02×10^{-7}	1.80×10^{-6}
IFITM3	0.00212	0.0297
IFITM1	0.00464	0.0541

In the analysis, we observed six significant eQTLs for rs12252, shown in table 4.3. The effect sizes of these eQTLs were not unidirectional - some genes were expressed at higher levels in CC donors, and some in TT donors (fig. 4.20). Notably, we observed eQTLs for *IFITM3* and *IFITM1*. However, *IFITM1* had an adjusted p-value of 0.0522 - failing to cross the threshold of 0.05. *IFITM1* was differentially expressed at higher levels in CC donors (fig. 4.21), with 1.22 times higher average expression in CC donors than TT donors on average. Under the additive model, we found that *IFITM1* expression was increased by a factor of 1.10 per C allele. This is a smaller magnitude of change than that observed with regards to *IFITM3*.

The other four significantly differentially expressed eQTLs were found to be Non-Coding RNA (ncRNA)s with a close genetic distance to rs12252 and *IFITM3*. These RNAs were either Long Intergenic Non-Coding RNA (lincRNA)s or antisense RNAs (via GENCODE 27, table 4.4) [316]. It is worth noting for the sake of clarity, that these transcripts are annotated

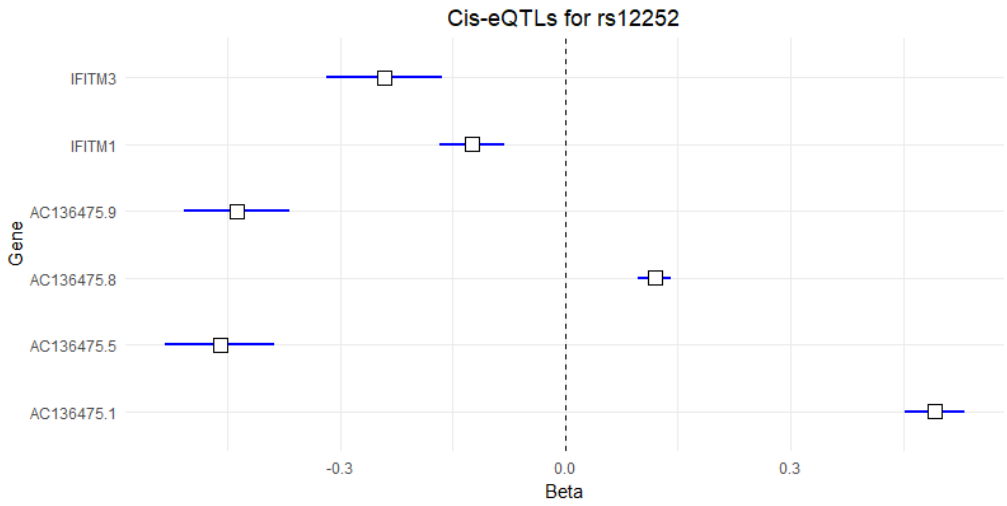


Figure 4.20: Forest plot showing effect sizes (beta) of rs12252 cis-eQTLs. Predicted effect sizes of all genes are shown with line denoting standard error. Genes with negative beta values show higher average expression in assoociated with the C allele, and genes with higher expression associated with the T allele have positve beta values.

with different names in the current and past versions of the human genome reference annotations. These RNAs are discussed further later in the chapter in section 4.2.5.

Table 4.4 Further ncRNA Information

ncRNA[1]	**ncRNA[2]**	**RNA type**	**DNA Strand[3]**
AC136475.1	RP11-326C3.11	antisense RNA	positive
AC136475.4	RP11-326C3.10	antisense RNA	positive
AC136475.5	RP11-326C3.12	lincRNA	positive
AC136475.9	RP11-326C3.15	lincRNA	positive
AC136475.8	RP11-326C3.14	lincRNA	negative

[1] Current Name
[2] Previous Name
[3] *IFITM3* is on negative strand

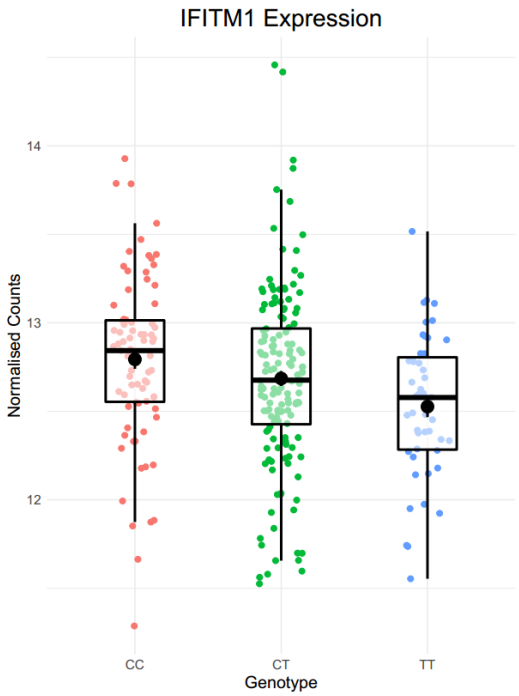


Figure 4.21: Average IFITM1 expression was found to be higher in CC donors than TT donors. IFITM1 expression is shown as normalised counts after application of a variance stabilising transformation. Black points represent mean and SEM. Differential expression is highly significant between homozygotes, and under an additive model of gene expression

Differential Gene Expression Analysis of Healthy Donor Cohort

We performed differential gene expression analysis to observe overall trends associated with rs12252 genotype. We used both the additive model of expression, and the homozygotes exclusively. We also performed trans-eQTL analysis to accompany our differential gene expression analysis. Notably, our trans-eQTL analysis did not find any significant eQTLs for rs12252. In the remaining two analyses, we found highly significant differential expression of the same ncRNAs we observed in our cis-eQTL analysis. Analysis is explored in more detail in Appendix E

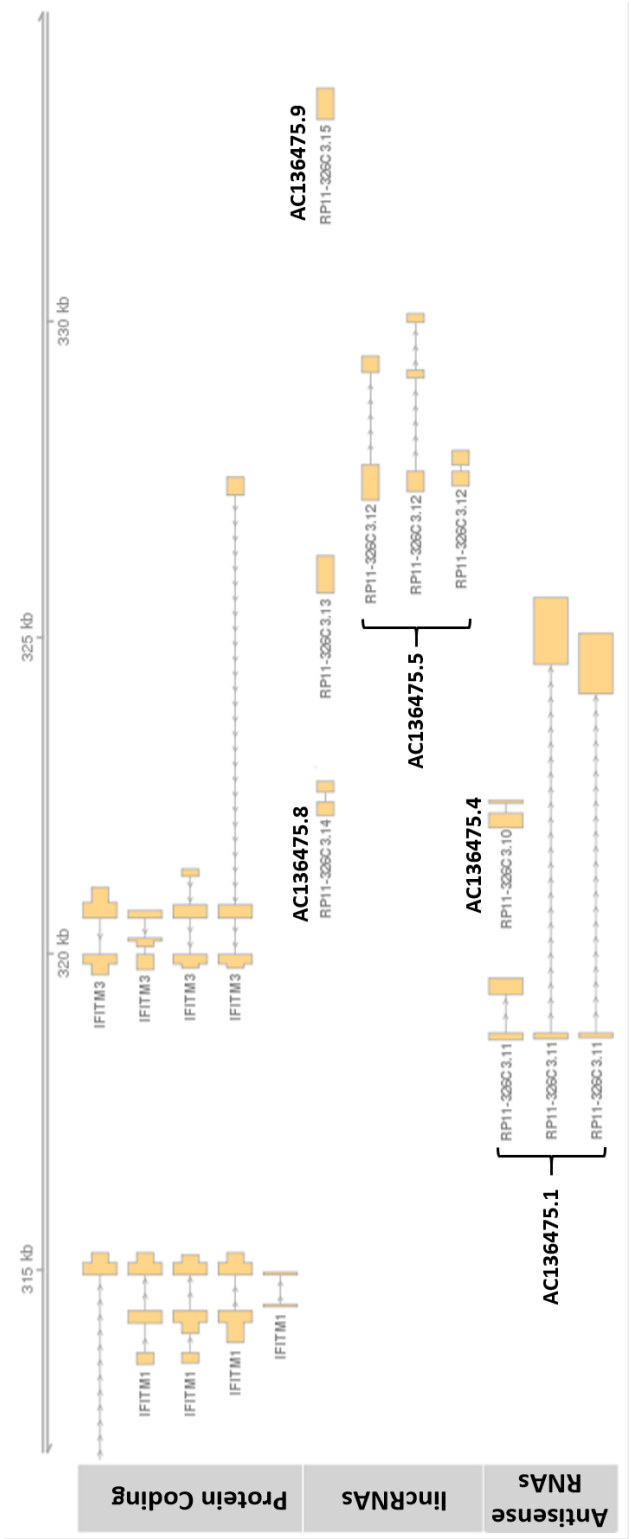


Figure 4.22: Map of IFITM3 and IFITM1 loci on chromosome 11 with surrounding ncRNAs flagged as eQTLs. ncRNAs are sorted by classification into lincRNA or antisense RNA. All transcripts are shown for all genes.

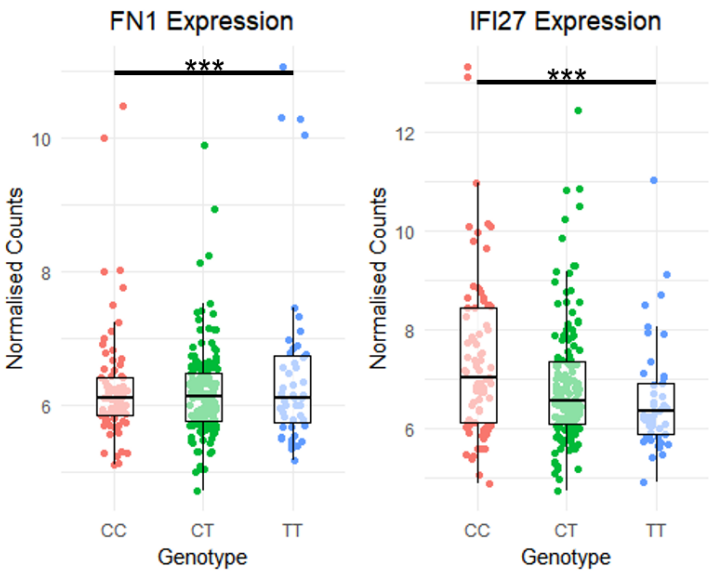


Figure 4.23: Differential expression of FN1 (left) and IFI27 (right). Expression is shown as normalised counts after application of a variance stabilising transformation. We observed higher expression of IFI27 in CC donors when compared to TT donors. Conversely, FN1 expression was biased in favour of TT donors, although this may be an outlier effect

We also observed differential expression of *Fibronectin 1 (FN1)* under both models. This test was likely confounded by a number of outlier samples that showed much higher expression levels of *FN1* than average - median expression among genotype groups was roughly the same. In addition, we observed significant differential expression of the *Interferon Alpha-Inducible Protein 27 (IFI27)* gene under the CC/TT model (fig. 4.23). CC donors showed, on average, a 2.5-fold higher expression levels of *IFI27* when compared to TT donors. This gene was not significantly differentially expressed under the additive model ($p = 0.180$)

Investigation of Differentially Expressed ncRNAs

When examining significantly differentially expressed genes with respect to rs12252 genotype, we observed that six of the eight most significantly differentially expressed genes were ncRNA within 15kb of the *IFITM3* locus using the CC vs TT model. This was in concordance with our cis-eQTL testing, which found four of these five ncRNAs to be eQTLs for rs12252.

More information regarding these ncRNAs is shown in table 4.4. Mapping locations of these features is shown in fig. 4.22. We found that the expression of these genes did not uniformly favour CC or TT donors; 3 of these were expressed at higher levels in CC donors, and 2 in TT donors. This is shown in more detail in table 4.5.

Table 4.5 Differentially Expressed ncRNAs

Transcript	**CC/TT**		**Additive**	
	Fold Change[1]	**p-value**	**Fold Change[12]**	**p-value**
AC136475.1	2.27	1.67×10^{-43}	1.52	9.09×10^{-35}
AC136475.4	0.183	2.27×10^{-7}	0.432	4.93×10^{-10}
AC136475.5	0.449	1.65×10^{-5}	0.678	2.29×10^{-5}
AC136475.9	0.441	1.65×10^{-5}	0.673	2.29×10^{-5}
AC136475.8	1.78	0.00235	1.34	0.00220

[1] Fold change favouring T allele
[2] Fold change per T allele

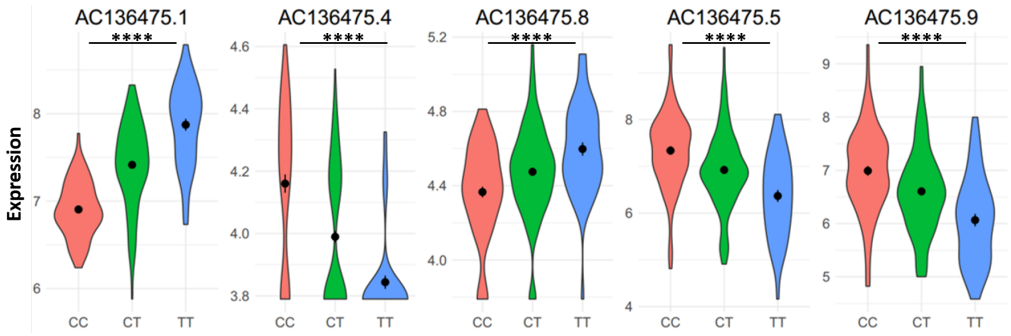


Figure 4.24: Differential expression of ncRNAs close to the IFITM3 locus. Expression patterns of these ncRNAs vary drastically depending on the rs12252 genotype. Expression is plotted in normalised counts after application of a variance stabilising transformation

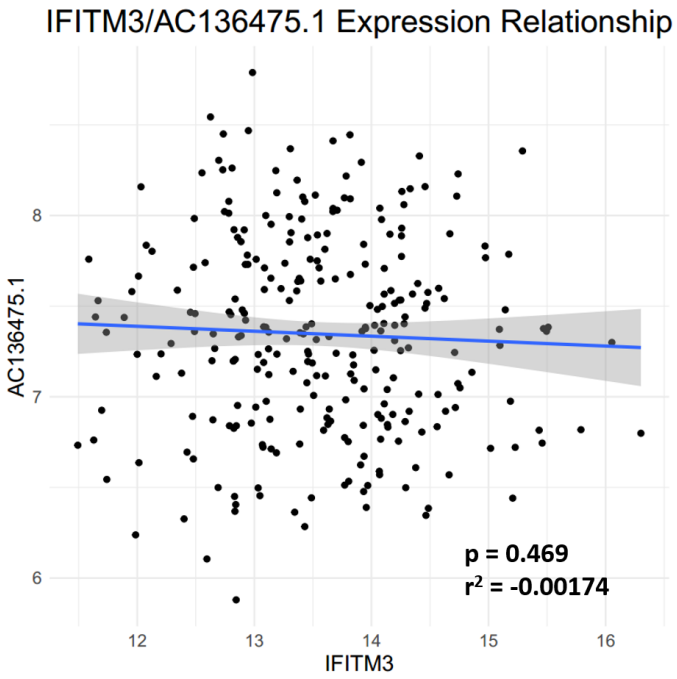


Figure 4.25: Scatter plot examining the relationship between IFITM3 and AC136475.1 expression. Expression is plotted in normalised counts after application of a variance stabilising transformation. There was no observed correlation between expression levels of the two genes.

The most significantly differentially expressed ncRNA of the dataset was AC136475.1. This RNA is transcribed on the positive strand over the *IFITM3* locus, which is located on the negative strand. The exonic portions of this gene do not overlap with the *IFITM3* locus. This transcript is associated with higher levels in TT donors than CC donors. This gene was expressed at a 2.27 fold higher level in TT donors than CC donors; this difference was found to be highly significant with a p-value of 1.67×10^{-43}. In addition, this gene was the most statistically significant fit for the additive model; it was expressed at 1.52 fold higher levels per T allele, and was associated with a p-value of 9.09×10^{-35}. We used linear regression analysis to explore the relationship between *IFITM3* and AC136475.1 expression; this relationship was statistically significant (fig. 4.25).

All aforementioned ncRNAs were expressed at low levels. With this in mind, we decided to visualise raw read alignment to these features, to investigate whether or not they showed real alignment to these features, or if they were a product of alignment noise. We overlayed coverage of 10 random CC donors

and 10 random TT donors relative to the annotated locus. These alignments were plotted as histograms of alignment depths. In visualising the *IFITM* locus on chromosome 11, alignments to non-coding RNAs are not visible - they are obscured by relatively higher alignment depth to *IFITM* genes (fig. 4.27).

Given this, we visualised all alignments to the five differentially expressed ncRNAs. In visualising AC136475.4/RP11-326C3.10 and AC136475.8/RP11-326C3.14 we observed extremely low read alignment - under 10 reads were aligned for every sample (fig. 4.27). This was consistent with their relatively low expression, even compared to the other ncRNAs. This alignment also did not appear to be specific to the annotated transcript.



Figure 4.27: Overlayed alignment of 10 CC and 10 TT samples to AC136475.4/RP11-326C3.10 and AC136475.8/RP11-326C3.14 in CC and TT donors. Darker shades of blue indicate greater agreement of alignment depth between samples. Alignment density is extremely low, with the greatest level of alignment near exon 1 of AC136475.8/RP11-326C3.14.

Looking at AC136475.5/RP11-326C3.12 and AC136475.9/RP11-326C3.15, we observed relatively high read alignment (fig. 4.28a). This alignment did not seem extremely specific, with the exception of alignments to the first exon of AC136475.5/RP11-326C3.12. This region showed a large number of reads that did not map to any specific feature, yet appeared to be distinct from background noise. We finally separately visualised the first and second exons of AC136475.1/RP11-326C3.11 - these exons are found on opposing sides of *IFITM3*. Here, we saw convincing alignment to the second

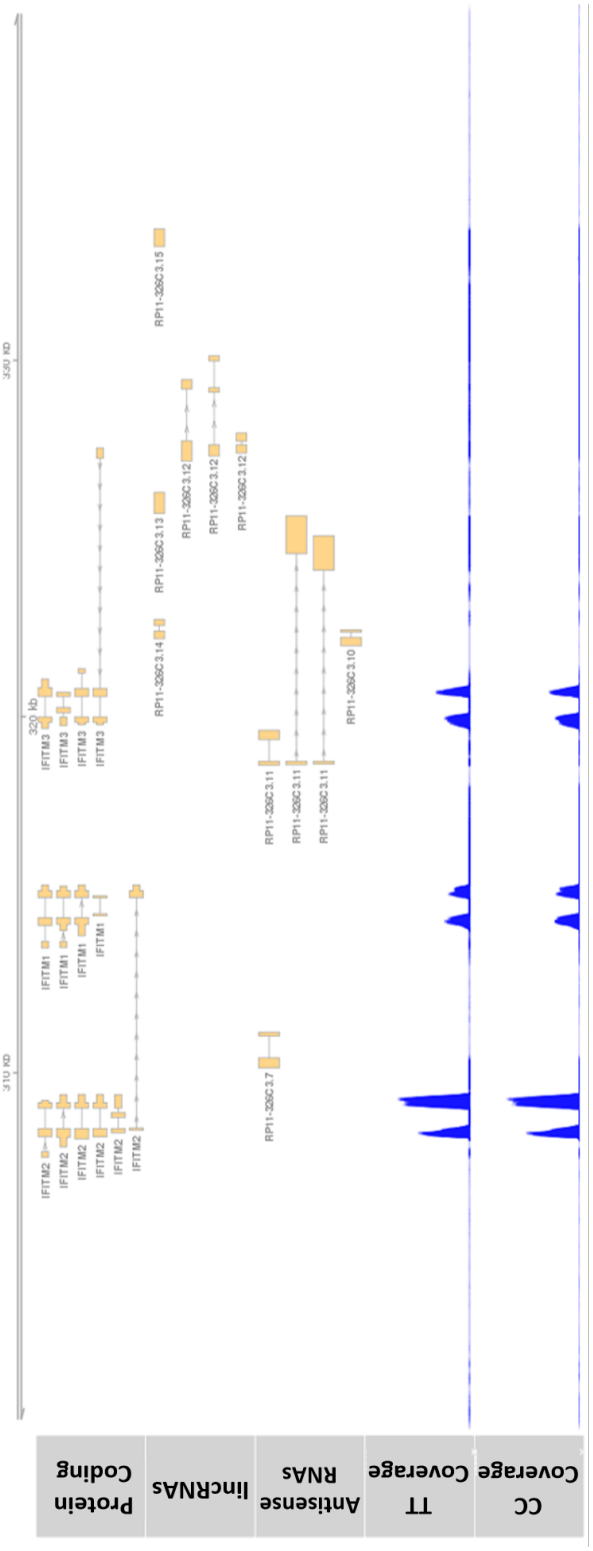


Figure 4.26: Overlayed alignment to the IFITM1, 2 and 3 loci of 10 CC and 10 TT donors. Darker shades of blue indicate greater agreement of alignment depth between samples. Degree of alignment is shown as a density plot. Alignment peaks overlayed with exons of IFITM genes. At this scale, we could not observe alignment to ncRNAs.

exon (fig. 4.28b). However, alignment to the first exon did not appear as specific, and did not occur to the same depth (fig. 4.28c). Overall, given this visualisation, we expect the reads assigned to AC136475.1/RP11-326C3.11 and AC136475.5/RP11-326C3.12 were the most likely to be accurate representations of real ncRNAs.

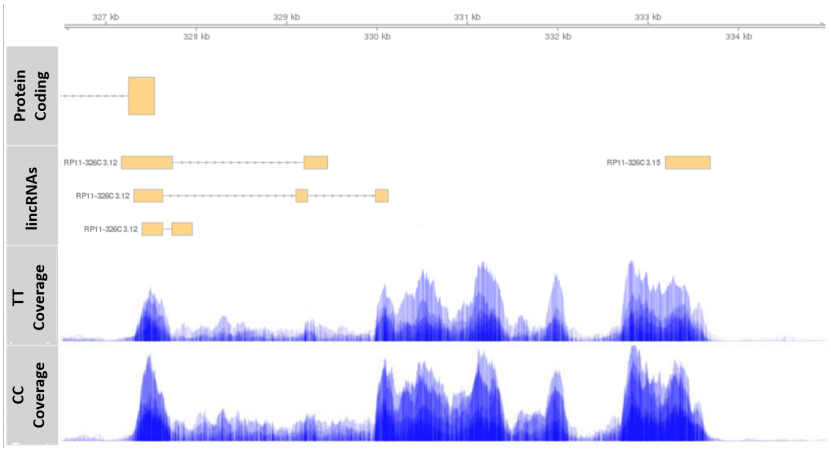
To further investigate the expression of AC136475.1/RP11-326C3.11, we used qPCR to amplify this ncRNA from our WT and CRISPR-Cas9 edited HEK293 cells. We used the reference sequence of AC136475.1/RP11-326C3.11 to design a custom TaqMan assay using the Thermo Fisher Scientific online webtool - this tool does not provide primer sequences, but shows locations of primer alignments. The designed primers span the first and second exons of this ncRNA, and ensure amplification is specific to the transcript itself. The size of the predicted amplicon was 122bp. We performed this qPCR using *GAPDH* as an internal control gene, and with untreated and IFN-α2 stimulated samples.

We observed low level amplification of AC136475.1/RP11-326C3.11, with a viable change in fluorescent signal beginning at roughly 30 cycles (fig. 4.29a). Ct values generally fell between 33 and 35 cycles. We did not observe differential expression with respect to either rs12252 genotype or IFN stimulation (fig. 4.29b). We verified the specificity of the custom primers for AC136475.1/RP11-326C3.11 via DNA gel electrophoresis - here we saw single band sizes corresponding to the predicted size of the amplicon (fig. 4.29c).

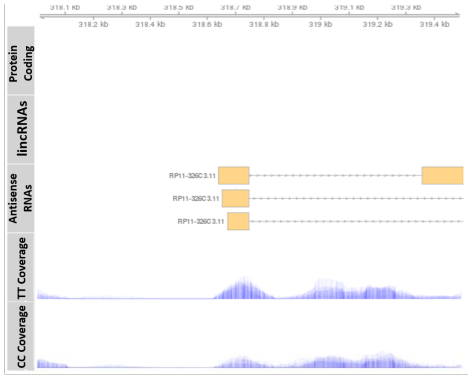
Gene Ontology Analysis of Healthy Donor Cohort

We subsequently performed GO analysis on this dataset (table 4.6). As mentioned in Appendix E, Principle Component Analysis (PCA) showed little sample-to-sample distance on average between donors in this dataset. For this reason, while performing GO analysis, we opted used a relaxed α of 0.05 (non-adjusted) as a threshold for inclusion into the gene subset tested for enrichment. While this gave us more power to detect trends in our dataset, this increased our risk of false-postitive GO enrichments. Given our hypothesis that the rs12252 genotype affected the expression levels of *IFITM3*, and potentially other *IFITM* genes, we removed these genes from our testing subset. GOs are explained in more detail in section 2.5.8.

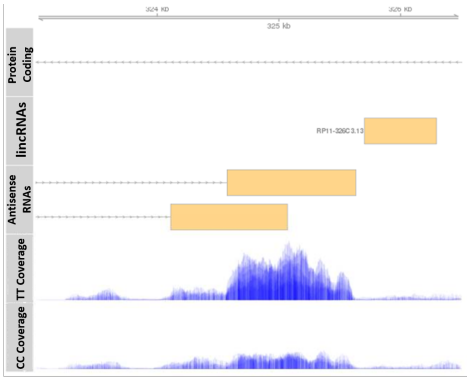
For biological processes, we observed enrichment of terms related to glutamine metabolism, dendrite development and multicellular processes



(a) AC136475.5/RP11-326C3.12 and AC136475.9/RP11-326C3.15

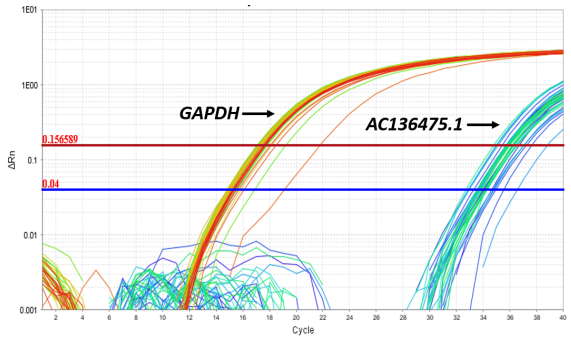


(b) AC136475.1/RP11-326C3.11 Exon 1

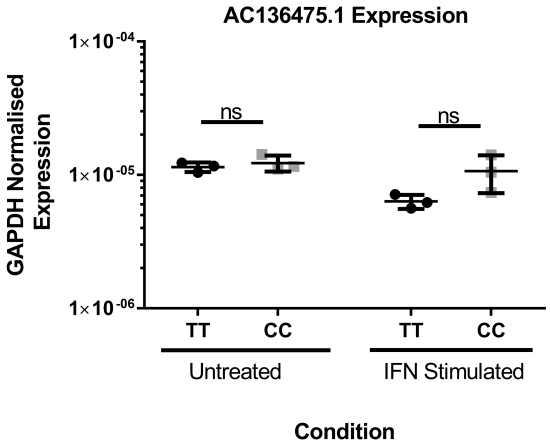


(c) AC136475.1/RP11-326C3.11 Exon 2

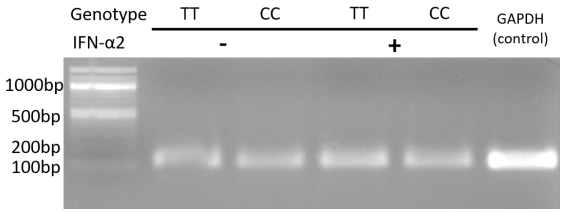
Figure 4.28: ***a.*** *Overlayed alignments of 10 CC and 10 TT samples to the AC136475.5/RP11-326C3.12 and AC136475.9/RP11-326C3.15 loci. Darker shades of blue indicate greater agreement of alignment depth between samples.* ***b.*** *Overlayed alignments of 10 CC and 10 TT samples to exon 1 of AC136475.1/RP11-326C3.11.* ***c.*** *Overlayed alignments of 10 CC and TT samples to exon 2 of AC136475.1/RP11-326C3.11.*



(a) Fluorescent Signal from AC136475.1 Primers

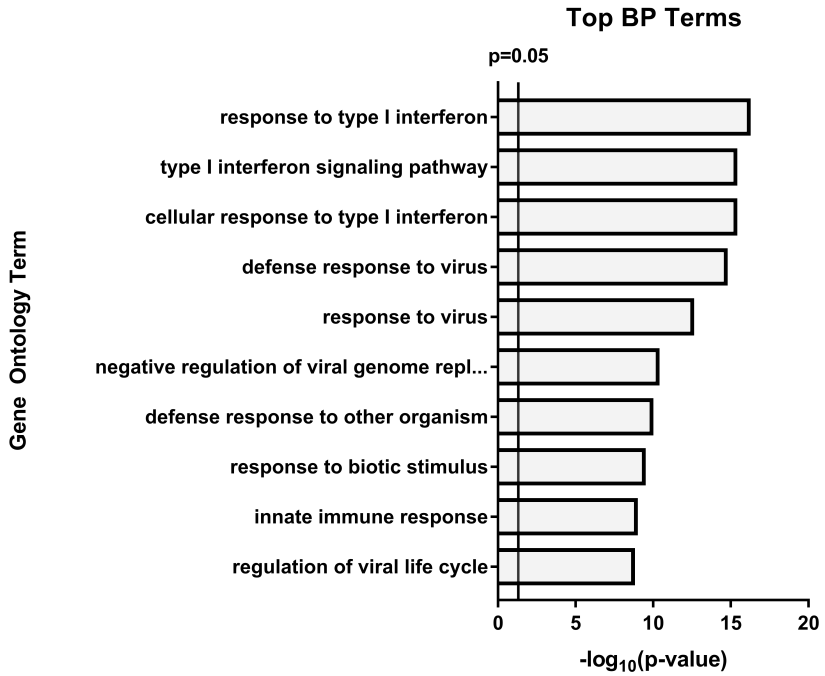


(b) Quantification of AC136475.1 qPCR

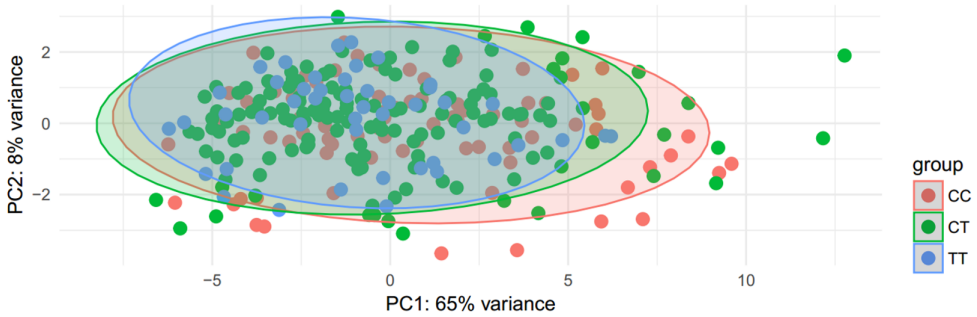


(c) Gel Electophoresis of qPCR Product

Figure 4.29: ***a.*** *Amplification AC136475.1 from cDNA of HEK293 cell lines. Graph shows change in fluorescent signal plotted against cycle number. Plot shows GAPDH and AC136475.1 amplification. AC136475.1 is amplified at high cycle count.* ***b.*** *Quantified expression of AC136475.1 in HEK293 cell lines. AC136475.1 expression is shown as expression normalised to GAPDH expression values. No expression effect is observed with respect rs12252 genotype or IFN stimulation.* ***c.*** *Gel electrophoresis of AC136475.1 qPCR products. DNA gel shows a singular band with amplification of a product agreeing with the predicted size of 122bp. A GAPDH qPCR product is shown as a positive control.*



(a) Significantly Enriched Biological Processes (CC)



(b) PCA Plot Based on Expression of Genes in GO:0034340

Figure 4.30: ***a.*** *Significantly enriched biological processes in CC donors when compared to TT donors. The x axis shows $-log_{10}(p - value)$ for each GO term found - these terms are all considered highly significant. Terms are highly associated with IFN response and ISGs.* ***b.*** *PCA plot generated using genes in to "response to type I interferon" GO term. Points show individual donors, and elipses show 95% confidence intervals. The rs12252-C allele has a mild association with differential ISG signature.*

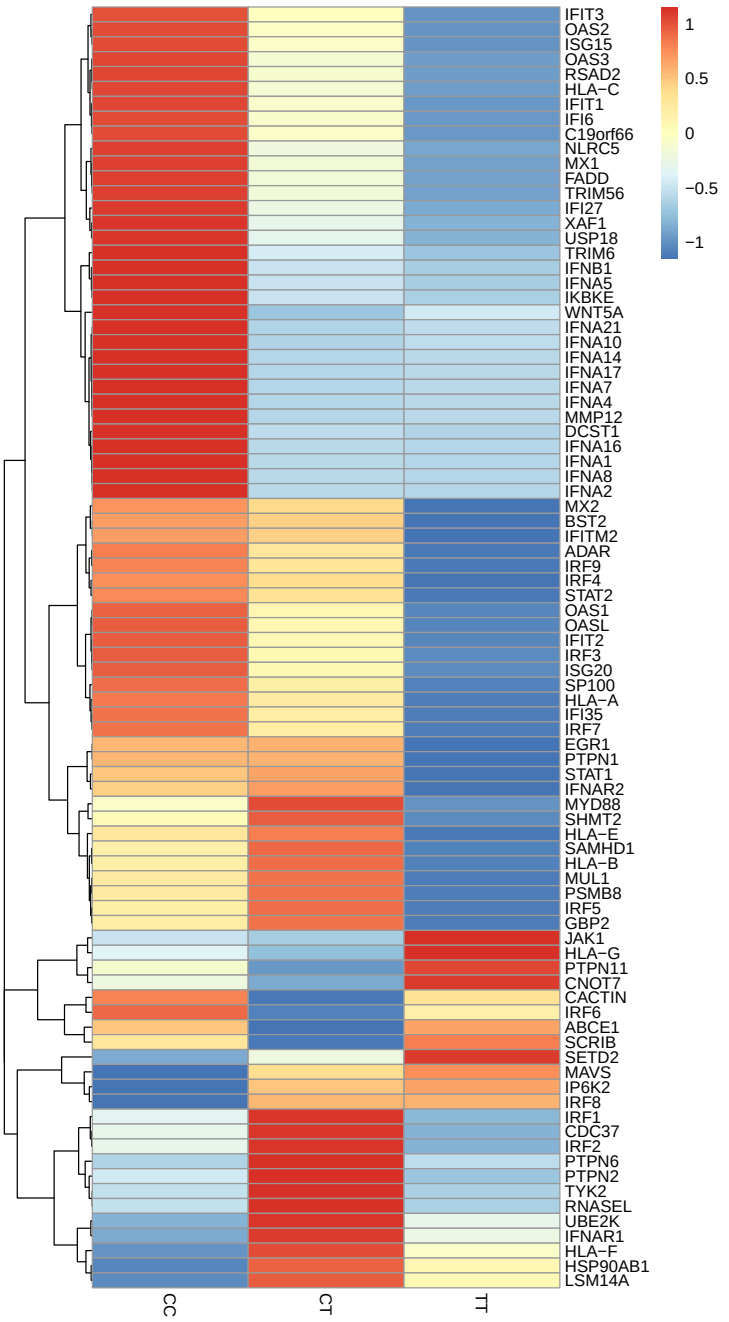


Figure 4.31: Heatmap of genes within the "response to type-I interferon" GO term. Expression scores were averaged between genotypes, and scaled by row. For this reason, differences in expression are somewhat exaggerated. The C allele of rs12252 was largely associated with high expression of genes in this panel, affirming previous findings. However, this is not uniformly true, with genes at the bottom of the chart showing more nuanced expression patterns.

in TT donors. The significance of these results was not high enough to warrant serious consideration given our generous testing parameters. However, when examining genes upregulated in CC donors, we observed highly significant enrichment of GOs related to type-I IFN responses. These terms included response to type I interferon (GO:0034340), type I interferon signalling pathway (GO:0060337), and cellular response to type I interferon (GO:0071357) (fig. 4.30a). These findings suggest the rs12252 genotype is related to baseline IFN signatures in individuals, albeit mildly. To further explore this, we obtained the list of 144 genes included in the most significantly enriched GO term, response to type I interferon. We excluded *IFITM1* and *IFITM3* genes from this ontology, alongside any genes that were not tested for by DESeq2. We used the remaining 84 genes for PCA based on this gene signature subset (fig. 4.30b). This showed an effect associated with the rs12252-C allele, supporting the notion of a higher basal IFN signature in CC donors. To visualise this on a gene-by-gene basis, we used mean normalised gene expression scores for each genotype to generate a heatmap using the same gene panel as was used for the PCA (fig. 4.31). Without averaging, a heatmap of 274 columns is not visually understandable. Here, we could see the same pattern as expressed with the PCA, although these effects are not uniformly visible across all genes in the panel.

Table 4.6 Top 3 Enriched GO Terms by Genotype/GO Subset

Gene Ontology	**p-value**
Biological Process (TT)	
glutamine family amino acid metabolic process	0.0012
regulation of dendrite development	0.0018
multicellular organismal process	0.0031
Biological Process (CC)	
response to type I interferon	7.1×10^{-17}
type I interferon signalling pathway	5.1×10^{-16}
cellular response to type I interferon	5.1×10^{-16}
Molecular Function (TT)	
receptor activity	0.00084
molecular transducer activity	0.00124
signaling receptor activity	0.00224
Molecular Function (CC)	
double-stranded RNA binding	1.6×10^{-7}
scavenger receptor activity	0.00041
single-stranded RNA binding	0.00058
Cellular Compartment (TT)	
apical part of cell	0.00038
apical plasma membrane	0.00041
anchored component of membrane	0.00276
Cellular Compartment (CC)	
condensed chromosome	7.7×10^{-6}
chromosome, centromeric region	1.1×10^{-5}
condensed chromosome, centromeric region	2.7×10^{-5}

In addition, we tested our dataset for enrichment of terms related to cellular compartment and molecular function. While we were able to observe enrichment of multiple GO terms here, none of these reached significance levels obtained when looking at upregulated biological processes in CC cells. A summary of all GO analysis findings is shown in table 4.6. We observed significantly more enrichment of GO terms in CC samples overall than TT samples.

Further Investigation of ISG Signatures in Healthy Donor Cohort

Previously, we had found that *IFITM* gene expression was highly correlated; this was supported by results those presented in the Allen *et al* study [219]. Based on this, we decided to examine the relationship between expression of *IFITM* genes. Here, we found that *IFITM1* and *IFITM3* expression were highly correlated - here, linear regression analysis gave an adjusted r^2 value

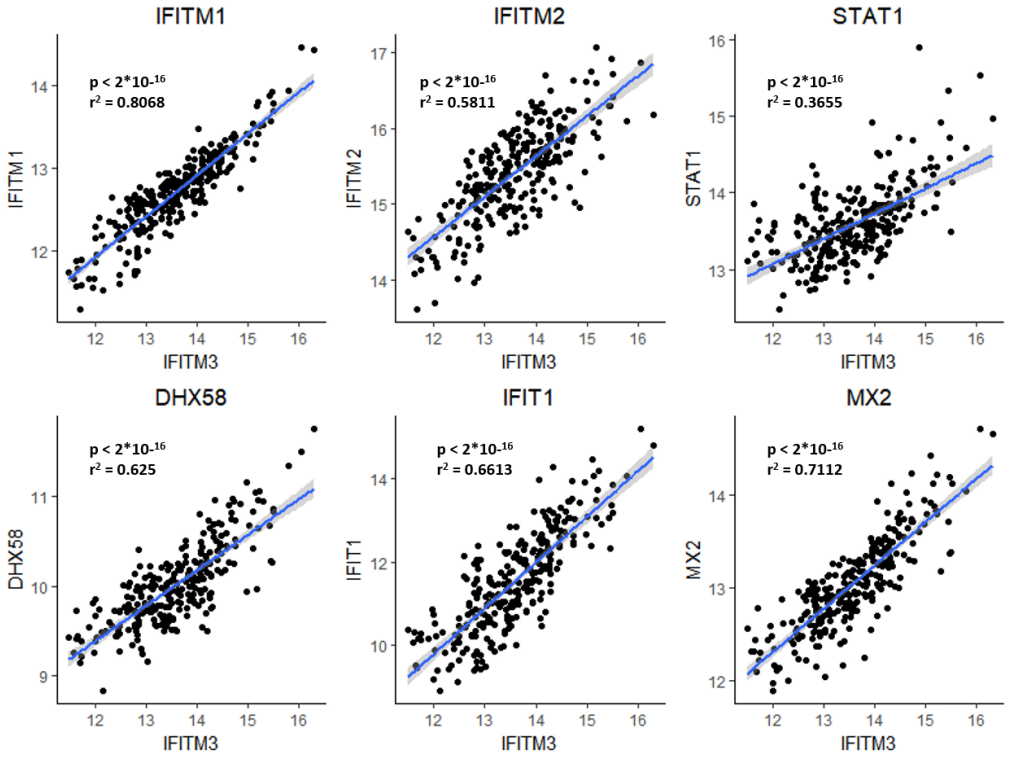


Figure 4.32: Correlations between IFITM3 expression and other ISGs, including IFITM1 and IFITM2. Expression is shown in normalised counts with a variance stabilising transformation applied. IFITM3 expression is highly correlated to IFITM1, IFITM2 and other ISG expressions.

of 0.81, and a p-value less than 2×10^{-16}. *IFITM2* and *IFITM3* were also correlated in terms of expression but less so - here, linear regression analysis gave an r^2 value of 0.55, and a p-value less than 2×10^{-16} (fig. 4.32). In addition, we examined the relationship between *IFITM3* expression and the expression of four randomly selected ISGs (STAT1, DHX58, IFIT1 and MX2) - the expression of these genes was highly correlated with *IFITM3*. We used a standard Bonferroni correction to control for false positives in this analysis.

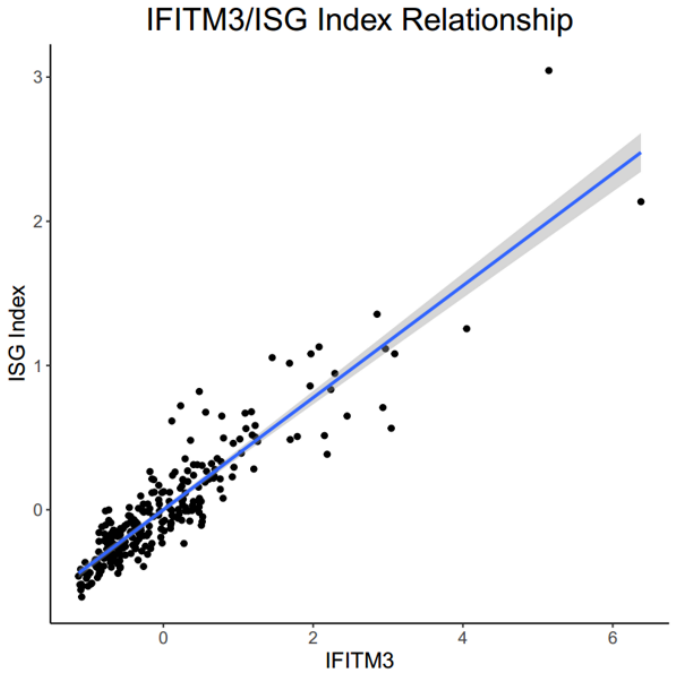


Figure 4.33: Linear regression showing relationship between ISG index and IFITM3 expression. IFITM3 expression is shown in terms of standard deviation of normalised counts from expression mean.

Having observed that the CC donors showed a weak IFN signature, and the expression of ISGs appeared to be highly correlated with each other, we separated the IFN signature of each donor from their *IFITM3* expression. Here, we hoped to show a more direct genotype effect on *IFITM3* expression separate from the ISG signature. For this analysis, developed an algorithm outlined in section 2.5.10. This methodology involves the calculation of an "ISG index" for each donor which represents, on a donor by donor basis, a measure of general ISG expression. Since this methodology interprets *IFITM3* expression in terms of SDs from the mean, we ensured that *IFITM3* expression was normally distributed. Using Shapiro-Wilk tests, we confirmed normality of *IFITM3* expression in genotype groups ($p = 3.4 \times 10^{-6}$ (CC);

$p = 7.5 \times 10^{-15}$ (CT); $p = 8.8 \times 10^{-6}$ (TT) and overall ($p < 2.2 \times 10^{-16}$)

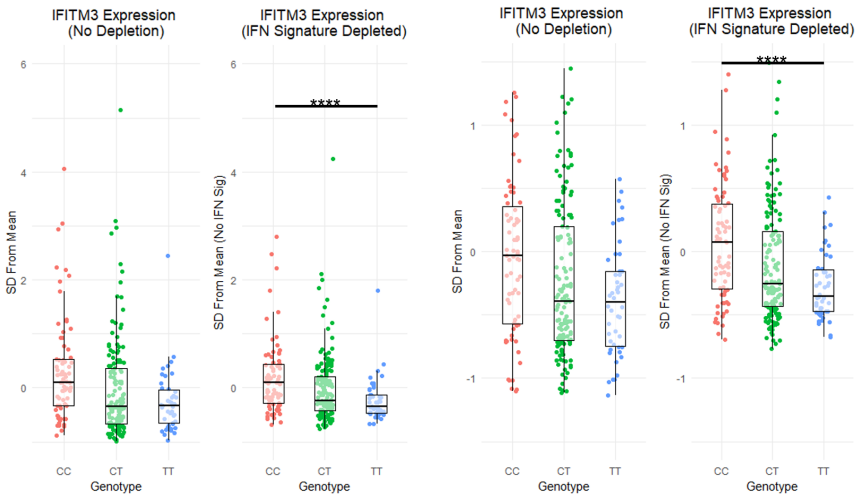
We ensured that our calculated ISG indecies showed correlations with *IFITM3* expression via linear regression (fig. 4.33). As expected, we observed strong correlations between our calculated ISG index and *IFITM3* expression ($p < 2.2 \times 10^{-16}$, $r^2 = 0.839$).

Table 4.7 IFITM3 Expression Statistics Before and After IFN Signature Depletion

Genotype	**Undepleted**		**Depleted**	
	mean	**SD**	**mean**	**SD**
CC	0.217	1.00	0.180	0.673
CT	-0.0133	1.07	-0.0237	0.652
TT	-0.323	0.600	-0.228	0.398
Total	0	1.00	0	0.635

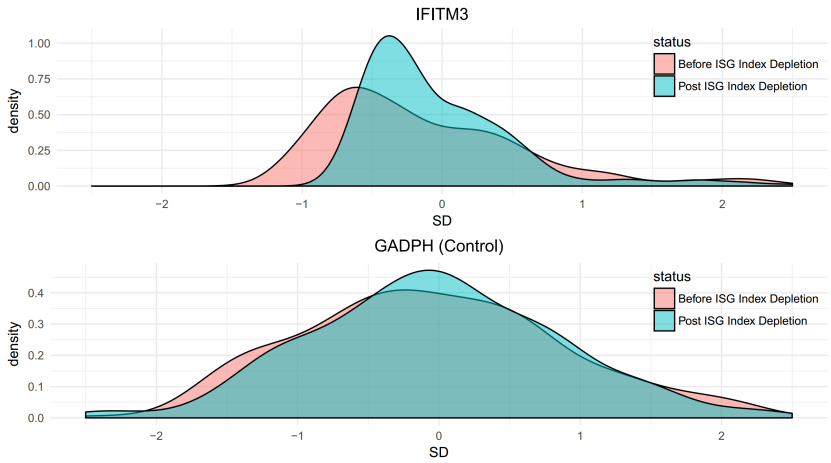
After we performed this analysis, we could still see a strong relationship between rs12252 genotype and *IFITM3* expression. This was highly significant between CC and TT donors via t-test ($p = 3.79 \times 10^{-5}$), and using an additive model ($p = 0.000341$) (figs. 4.34a and 4.34b). This constituted an increase in statistical significance when compared to tests performed before ISG signature depletion ($p = 2.39 \times 10^{-4}$ (CC/TT), $p = 0.00332$ (Additive)) Again, rs12252-C was associated with higher *IFITM3* expression.

This algorithm incurred weak changes on genotype specific mean *IFITM3* expression values - the difference in mean expression between CC and TT donors was reduced from 0.539 to 0.408 SDs. This was accompanied by a substantial decrease in the distribution of *IFITM3* expression (fig. 4.34c) - overall *IFITM3* expression SD was decreased from 1.00 to 0.635 - this constitutes a 2.48-fold decrease in variance. Via F-test, we found this change in SD to be highly significant ($p = 1.85 \times 10^{-13}$). As a control, we ran the same analysis with *GAPDH* - this algorithm had no significant effect on the overall dispersion of *GAPDH* expression ($p = 0.399$). Changes in SD are explored in more detail in table 4.7. When applied to other ISGs, this algorithm also reduced expression dispersion. In fact, application of this algorithm selectively reduced dispersion of ISGs, in comparison to control housekeeping genes which were largely unaffected by it (fig. 4.35). For housekeeping genes, we tested *GAPDH*, *ACTB*, *IPO8*, *HPRT1*, and *PGK1P2*. For ISGs, we tested *STAT1*, *IRF1*, *MDA5*, *IFI6*, *MX2*, *IFIT1*, *DDX58*, *ISG15*, and *IRF7*. However, when applied to these ISGs, this algorithm did not incur strong increases in significance of expression association with rs12252 genotype, like



(a) Full

(b) Zoomed



(c) Expression Distribution

Figure 4.34: ***a.*** *IFITM3 expression with and without incorporation of calculated ISG index - a estimate of underlying ISG signature. donors are stratified by rs12252 genotype. IFITM3 expression is displayed as SDs from mean IFITM3 expression, before and after loss of ISG signature. Our previously observed relationship between rs12252 and IFITM3 expression is preserved without ISG signature. rs12252-C is associated with significantly higher levels of IFITM3 expression.* ***b.*** *Plots in* ***a.*** *shown on a larger scale. Sample SDs are reduced.* ***c.*** *Density plots showing overall distribution of IFITM3 expression before and after loss of ISG signature. Loss of ISG signature significantly reduces overall distribution. Overall GADPH expression is largely unaffected by application of this algorithm*

we observed with *IFITM3* (fig. 4.36).

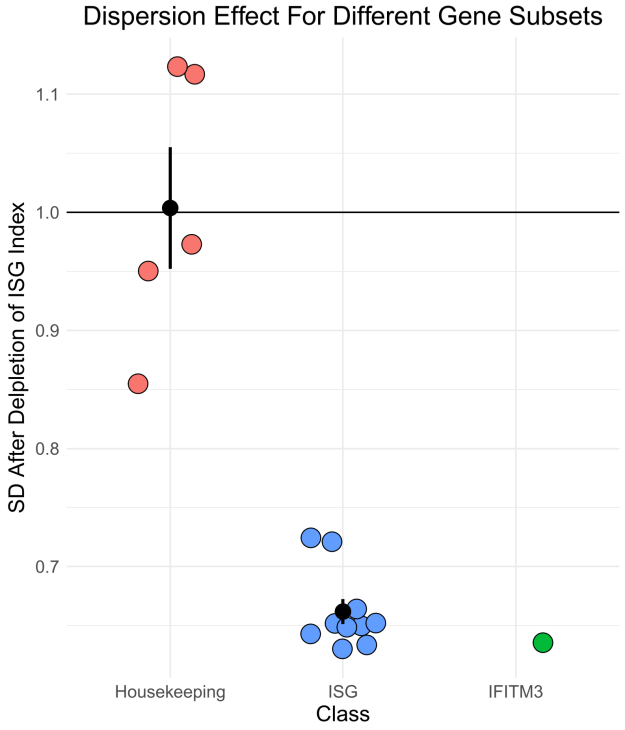


Figure 4.35: Effect of ISG signature depletion algorithm on different gene groups. As part of this analysis, all expressions of all genes were mean centred to 0, with an SD of 1. Depletion of ISG signature uniformly reduced the dispersion of ISG expression across a panel of 10 ISGs. In comparison, a panel of housekeeping genes was not affected uniformly, with post hoc SDs centred around 1.

Investigation of Gene Relationships in Healthy Donor Cohort

In the Allen *et al* 2017 study [219], it was observed that on a rs34481144 genotype basis, the expression of *IFITM3* was correlated with genes in its broader locus on chromosome 11. This was observed using PBMC samples from 45 individuals. In this study, these relationships were observed on subsets of rs34481144 genotypes, meaning a relatively small sample number was used. To follow up on these findings, we used our cohort to investigate whether these findings were reproducible in a larger cohort regardless of genotype. In this analysis, we analysed all local genes that were significantly correlated with *IFITM3* expression in the Allen *et al* study (Sirtuin 3 (SIRT3), Beta-1 4 N-acetylgalactosaminyltransferase (B4GALNT4), Single

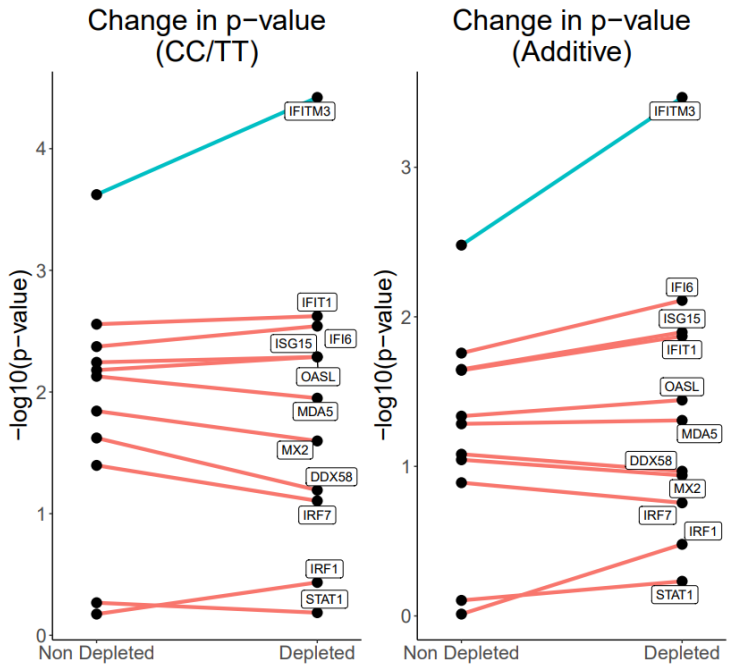


Figure 4.36: Effect of ISG signature depletion algorithm on associations between rs12252 and gene expression using a CC/TT model (left) and additive model (right). Statistical testing was performed using t-tests (CC/TT) and linear regression (additive). Depletion of ISG signature did not incur strong increases in p values with the exception of IFITM3.

Ig IL-1-related receptor (SIGIRR), Anoctamin 9 (ANO9), Leucine-rich repeat-containing protein 56 (LRRC56), Ras association domain family member 7 (RASSF7)). We used linear regression between *IFITM3* and all genes above to test this hypothesis. We used a standard Bonferroni correction to control for false positives in this analysis.

After performing this analysis, we observed significant correlation between *IFITM3* and *RASSF7*. This relationship was highly statistically significant (p=0.000733), but was accompanied by a relatively low adjusted r^2 value of 0.0494.

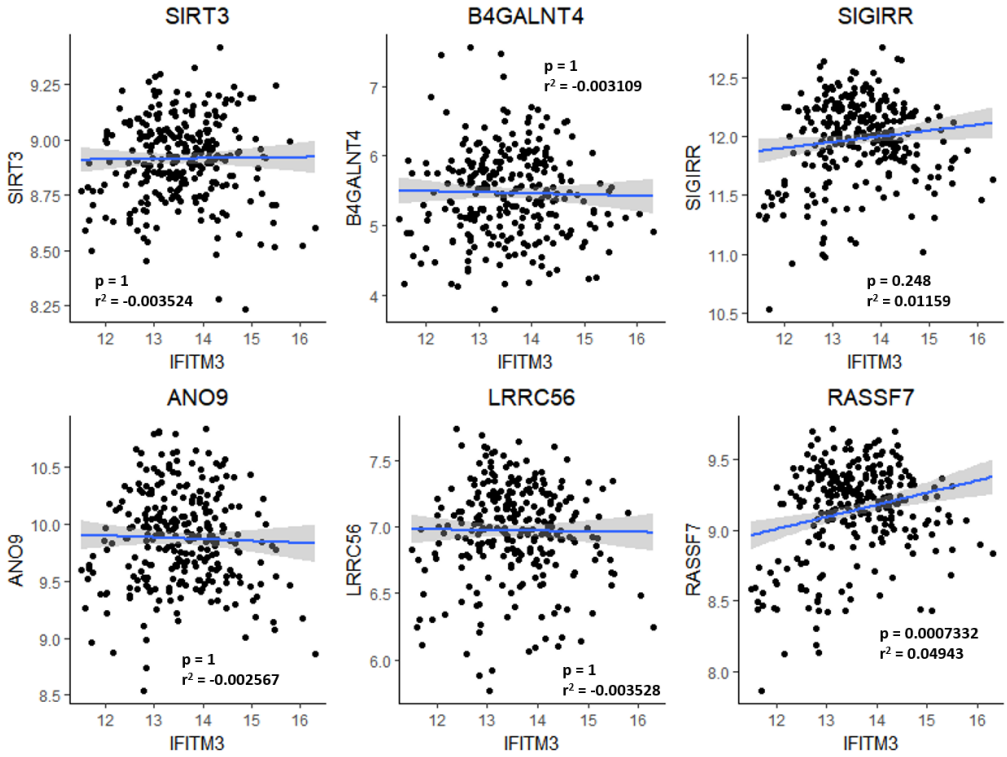


Figure 4.37: Correlations between IFITM3 expression and expression of genes near the IFITM3 locus. Genes were chosen based on significant expression relationships in the 2017 Allen et al study. Expression is shown in units of normalised counts, after application of a variance stabilising transformation. None of the above genes showed expression correlations with IFITM3 aside from RASSF7

4.3 Discussion

4.3.1 *IFITM3* Expression is Modulated by rs12252

The effects of rs12252 on the *IFITM3* gene are not well understood in the literature to this day. In chapter 3, we showed that the previous conventional hypothesis as to the effects of this SNP was likely false, leading us to conduct an investigation into its effects from the ground up. Over the course of this chapter, we have accumulated a wide range of evidence that rs12252 is a acts as a modulator of *IFITM3* expression. After originally observing this trend via allelic expression, we have investigated this relationship in a wide range of systems. As explored below, our data points to a complex relationship between rs12252 and *IFITM3* expression.

Over the course of data collection for this chapter, several studies have been published where the expression of *IFITM3* has been investigated in relation to rs12252, although none of these analyses produce evidence strong enough to draw any firm conclusions. In the 2017 Randolph *et al* study [249], rs12252 was evaluated as a risk factor for severe paediatric influenza virus infection, using a primarily Caucasian cohort. In this case, a statistical power calculation was performed, giving greater than 80% change to detect an association between rs12252 and *IFITM3* expression despite using only three CC homozygotes for this analysis. In this case, no association was found between the two factors. In the 2017 Allen *et al* study [219], the authors again attempted to identify associations between rs12252 and *IFITM3* expression, this time using monocyte-derived DCs from 534 healthy donors. These studies did not report donor demographics; however, there were only 3 different donors with the CC genotype. Despite the results of these studies, these investigations suffer low statistical power when compared to our Chinese healthy donor cohort.

Initially, we worked to expand our group of fibroblast cell line cohort to develop a useful *in vitro* method to investigate the effects of rs12252 on *IFITM3*. After obtaining three cell lines of each homozygous genotype, we profiled expression of IFITM3 at the protein and mRNA level. any effect of rs12252 on gene expression - it was unclear if this was related to a lack of effect or a lack of statistical power. This indicated to us that a large scale cohort expansion would be required to capture this effect. When looking at significant eQTLs for *IFITM3* and other genes, it was clear that

a large cohort with relatively even distribution of alleles would be required to investigate this effect using donor samples.

While fibroblasts were useful as an early system to investigate rs12252, we decided to limit our use of these cell lines due to concerns regarding cellular senescence. With increasing passage numbers, our fibroblasts showed decreasing proliferation, eventually becoming senescent. This is typical behaviour for primary, non-immortalised fibroblast cell lines [317]. One concern we had was the link between passage number and accumulating DNA damage. It has been shown that acquired DNA damage induces Cyclic GMP-AMP Synthase (cGAS) signalling in senescent cells, which leads to higher levels of IFN signalling with higher passage numbers. Based on this phenomenon, high expression of ISGs is considered a hallmark of cellular senescence [318, 319]. When we performed experiments with fibroblasts, we endeavoured to make comparisons between cell lines with similar passage numbers. Despite this, we limited the use of primary human fibroblasts going forward. We felt that this was unlikely to be an issue when using an allelic expression system, as we were not examining total *IFITM3* expression.

Due to our initial lack of statistical power in investigating rs12252 as an eQTL, we used allelic expression imbalance as a surrogate for real expression. Allelic expression imbalance has been shown to be a common phenomenon within the human genome. Different studies have reported that allelic expression imbalance occurs in anywhere between 10-68% of exonic SNPs in the human genome [320, 321]. In a 2008 study by Serre *et al* [322], incidents of allelic imbalance were been shown to be biologically relevant; expression effects observed in this manner can be recapitulated using alternative methods. In addition, it was shown that allelic expression is predictive of total gene expression in homozygous individuals, and can be predictive of epigenetic modifications. Finally, this study demonstrated that passage number did not significantly change allelic expression.

The use of allelic expression was born out of necessity, but yielded compelling results with regards to rs12252. Here, we could control cell-specific effects of individual samples by looking at the ratio between C and T alleles on an individual basis. This allowed us to overcome some of our observed sample-to-sample variation that confounded our expression studies using homozygous samples. In our targeted RNA-Sequencing experiment, control gDNA samples showed allelic expression imbalance in favour of the C allele, further supporting that our observations were real in this system.

Our initial investigation using allelic expression strongly suggested associations between the T allele and higher expression of *IFITM3*. While on a sample-to-sample basis, it was likely that individual ratios could be affected by chromosome specific factors that are unrelated to rs12252, the fact that T allele overexpression is highly consistent between samples indicates that the observed phenomenon is specifically related to rs12252 itself. We did see some evidence that the effects of rs12252 were modulated by IFN induction or active infection, although this would require further investigation to confirm, and is not supported by results from our CRISPR-edited cell lines. This runs in contrast to our pilot study, where we largely observed *IFITM3* expression differences at basal expression.

We subsequently used CRISPR-Cas9 to generate a final *in vitro* system to investigate the effects of rs12252 as an eQTL. Here, we generated cell lines incorporating the rs12252-C allele on an isogenic background to observe the effects of this SNP in isolation.

It is well documented that the CRISPR-Cas9 generally incurs off-targeting effects when applied to cell lines, many of which are not accurately captured using the *in silico* analysis we performed when designing the sgRNAs used in these experiments [323]. Guide RNA design was performed relative to the human reference genome, which does not accurately represent the genome of the HEK293 or A549 cell lines we used. For this reason, even fully accurate *in silico* analysis could not capture all potential target sites for our sgRNA within the genome of our targetted cell lines. We did not observe any differences in growth rate or cell morphology between WT HEK293 cells and the rs12252-C Knock-In (KI) lines - this does not rule out any large scale, off-target changes.

To confirm the association between rs12252 and *IFITM3* in an *in vitro* model, a greater number of HEK293 cell clones with rs12252-C incorporated would be ideal. We were unable to generate more than one of these clones despite screening over 100 clones. However, obtaining a larger number of these lines is an important endeavour moving forward. Recently, programmable adenine base editors have been designed using a sgRNA scaffold, which may allow for generation of the rs12252 SNP on an isogenic background more efficiently with less potential for genome aberration [324].

When using CRISPR-Cas9 edited HEK293 cells as a model to study rs12252, we found evidence that this SNP, in isolation, was sufficient to influence the expression of *IFITM3*. In this case, the fact that expression differences are

observed in both untreated and IFN stimulated conditions again suggests that the mechanism of expression reduction is not related to IFN induction. Additionally, the fact that rs12252 was enough to reduce the expression of *IFITM3* in isolation is supportive of the idea that rs12252 is the causative SNP for the observed reduction in *IFITM3* expression. Due to the nature of human samples, it was never possible to fully isolate the effects of rs12252 and potential causative SNPs in LD with it via viral patient cohorts. However, this data clearly supports the hypothesis that rs12252 itself is driving the differential expression of the *IFITM3* gene.

Despite the shortfalls of the CRISPR-Cas9 system, the data obtained using these cell lines, taken in summation with other datasets from this chapter indicate that this effect observed in these cells is more than likely a real effect of rs12252-C. Moving forward, this system will prove invaluable in investigating the effects of rs12252 on *IFITM3* expression and cell physiology.

Our large healthy donor cohort was likely our strongest tool to identify eQTL effects of rs12252. With 274 donor samples, and relatively even distribution of rs12252 genotypes, we had a great deal of statistical power to test rs12252 as an eQTL. This cohort was additionally advantaged by the fact that donors were similar in age, reducing the effects of age-related differential gene expression. Whether or not donors are smokers is considered to be an informative phenotype in blood transcriptional signature, and would have been useful as a covariate [325]. This information was not obtained, and we could not consider it in our modelling. However, with our large cohort size, it is unlikely that this would drastically influence the strong eQTL effects we observed.

A final drawback of this cohort is the fact that RNA was obtained from whole blood instead of from a more uniform cell source. This is a common tissue for eQTL databases - GTEx also uses whole blood as a tissue. Using whole blood as a source of RNA, allowed us to store blood for long periods of time - this was essential for gathering the samples needed for this cohort. In addition, we were able to obtain a higher RNA quality than possible with sorted and frozen cells. Publicly available data from the BioGPS database [326] suggests that *IFITM3* expression differs between different blood cell types; this is in agreement with findings from our group as well. If specific cell types were overrepresented between genotypes in this dataset, this pattern could influence *IFITM3* expression, and distort SNP-specific effects we observed. We consider this possibility unlikely: this dataset contains enough samples to overcome random sample-to-sample variation. If genotype did influence the

abundance of specific cell types within this dataset, we could expect to see cell specific markers differentially expressed between genotype groups. With the exception of *FN1*, addressed below, this was not the case. We would also not expect to see agreement between allelic and real expression in this case.

We observed differential expression of *FN1* in this dataset. Via the GTEx database [242], the expression of this gene within the whole blood dataset is likely to be entirely derived from fibroblasts. While this does raise some concerns about differential cell populations in this sample, it should be noted that this gene had low expression, and its differential expression was caused by outlier effects. The detection of this gene was most likely incurred by a low level fibroblast contamination of whole blood samples from the phlebotomy needle used in collection. Outlier samples likely had a larger number of contaminating fibroblasts than average - this could have been caused by suboptimal phlebotomy in a small number of donors. Regardless, these fibroblasts would be a minority population of cells, and would have negligible effects on overall gene expression.

We had little to no concerns with the validity of the data obtained from this healthy donor cohort, and have a great deal of confidence in the results from this dataset.

In the healthy donor dataset, we observed differential expression of *IFITM3* with respect to rs12252. The fact that *IFITM3* was differentially expressed in favour of the C allele was the opposite to the effect that we expected to observe in this dataset. This dataset provides the strongest piece of evidence that rs12252 is an eQTL for *IFITM3* in an EAS background. This is supported by allelic expression imbalance observed in favour of the C allele. We do not expect this result to be spurious with relation to previously obtained results.

The hypothesis that low *IFITM3* expression is related to worse viral disease outcomes is supported by the recent Allen *et al* study [219]. Here, strong evidence was presented that the rs34481144 SNP modulated *IFITM3* expression, and thus was associated with more severe outcomes in IAV infection. Given our findings with regards to rs12252, it is likely that other *IFITM3* eQTLs could be determinants of severity in IAV infection. This evidence is also supported by a wealth of *in vitro* investigations showing relationships between cell line *IFITM3* expression and susceptibility to IAV infection. It has also been shown that high levels of IFITM3 expression in resident memory T cells was associated with greater survival of these cells

during IAV infection [327]. Taken in summation, it does not seem likely that this observed higher expression of *IFITM3* in CC donors would explain the phenotype associated with the CC genotype.

Despite our initial, and limited observations that rs12252 was associated with reduced basal expression of *IFITM3*, it has become clear that the relationship between rs12252 and control of *IFITM3* expression is more complex than initially thought. While our initial findings pointed towards an association between rs12252-C and reduced expression of *IFITM3*, our data taken from our large healthy donor cohort indicates that the C allele may have the reverse effect on expression of *IFITM3*. It may be that rs122252 modulates *IFITM3* in a tissue specific, bidirectional manner, or that the lower expression of *IFITM3* associated with rs12252-C occurs in donors not captured by our healthy donor cohort - this could be a phenomenon best observed in older patients for example.

The idea of rs12252 may be acting as an eQTL where the directionality of expression is tissue specific would be unusual if true. In the GTEx study [242, 328], the vast majority of eQTLs significant for multiple tissues were shared directionality of expression between tissues. While there were instances where multi-tissue eQTLs showed opposing effects, the authors suggested these effects may have been related to LD patterns. Here, SNPs in LD with the eQTL SNP in question would have masked the effects of the queried SNP through eQTL effects of their own. It is possible that this would not account for all SNPs and that rs12252 is in strong LD with a causative SNP that modulates the expression of *IFITM3* within this dataset. This hypothetical SNP also may not have been captured with the 1000 genomes database.

A 2012 Fu *et al* study [329] also observed tissue dependent, opposite eQTL effects while performing larger eQTL analysis. They found that of the tissue-dependent eQTL, 4.4% of these were opposite effect eQTLs. In this study, it was determined that this incidence was much greater than would be expected by chance, as controlled by comparing two different datasets from the same tissue source. For this reason, it seems unlikely that SNPs in LD can account for all observed cases of opposite effect eQTLs. Interestingly, the authors observed that of all SNP classifications, synonymous SNPs were the most likely to cause unusual tissue specific effects. The authors did not observe any opposite effect eQTLs for *IFITM3*. A Hauberg *et al* 2017 study [330] also identified a number of tissue dependent opposite effect eQTLs in a meta-analysis of 57 GWASs. Overall, there is substantial evidence that some eQTLs do exert opposing effects that are tissue specific, and it is certainly

possible that rs12252 could be one. Insofar as we are aware, there have been no studies linking a tissue-dependent, opposite effect eQTL to a mechanism.

As mentioned with regards to our CRISPR-Cas9 edited cell lines, it is highly unlikely, given our evidence, that rs12252 affects the translation efficiency of *IFITM3* - we observe trends at the protein level, which would supersede this hypothetical interaction. In a 2008 Tats *et al.* study, it shown that a number of codon pairs are evolutionarily preferred or avoided; it is likely that translational bias is one of the main drivers of this phenomenon [331]. We cross referenced the serine codon containing rs12252, alongside both of its flanking codons, to the database prepared in this paper: we found no evidence here that rs12252 altered its codon to incur lower translational efficiency. With this in mind, we are confident that rs12252 modulates *IFITM3* expression directly at the mRNA level, and that any additional protein level effects would be incidental.

Overall, we have accumulated strong evidence that rs12252 is an eQTL for *IFITM3* in individuals with an East Asian background. In addition, our CRISPR-Cas9 studies, and the strong association we observe between rs12252 and *IFITM3* expression indicate that rs12252 is the likely causative SNP for this relationship. With this in mind, it is important to consider that we currently have the strongest base of evidence for rs12252 acting as an eQTL from our healthy donor cohort, and further studies are required to confirm the idea that the T allele is associated with high levels of *IFITM3* expression in some tissues. Moving forward, it would be useful to look at other tissue types to attempt to generate strong evidence of further rs12252-*IFITM3* eQTL effects.

4.3.2 Mechanism of rs12252 Effect is Still Poorly Understood

In addition to investigating the likelihood that rs12252 was an eQTL for *IFITM3*, we performed some preliminary investigation into potential mechanisms by which rs12252 may modulate the expression of *IFITM3*. These investigations largely centred on the effects of rs12252 on *IFITM3* mRNA, without consideration of the specific mechanisms by which these effects would have occured. With our previous results, we had strong evidence that the rs12252 mechanism of action was related to mRNA expression. We assumed therefore, that this SNP affected the transcriptional rate, or

degradation rate of *IFITM3* mRNA. We hoped that, in identifying which of these mechanisms was affected, we could form the basis for further mechanistic studies.

We first investigated the transcriptional rate of *IFITM3*, using the targetted RNA-Sequencing approach we used previously to investigate allelic expression imbalance in rs12252 heterozygotes. In this experiment, we again observed allelic expression imbalance in favour of the T allele; this effect was not as strong as the effect observed using total RNA. In this experiment, we observed a greater C allele PCR bias when compared to the total RNA experiment. This greater bias is likely a compounding of the effect seen with total RNA, increased with further PCR cycles and may explain why this effect was muted in comparison to the previous effect.

Ideally, we would like to confirm the observed reduction in transcriptional rate associated with the C allele using real expression as opposed to allelic expression imbalance. For this reason, we intend to re-perform nascent mRNA pulldown studies using CRISPR-Cas9 edited cell lines when we have obtained a large enough number of clones to confirm real expression reduction in HEK293 cells.

Our investigation of *IFITM3* mRNA stability in both CC and TT cells involved a transcriptional inhibition assay in primary human fibroblasts. This assay used ActD, a transcriptional inhibitor, to suppress all transcription. It has been shown, in numerous studies, that ActD has effects cellular physiology, and can change the half-life of numerous mRNA transcripts. This could potentially confound any findings observed with this assay [332, 333]. Other assays used to measure mRNA half life generally involve short term induction of genes with artificial, inducible promoter elements [334]. Given that the mechanism of action of the rs12252 SNP has not been discerned, we have consistently chosen to not use artificial gene constructs to investigate its effects. We believe it is likely that the genomic context of rs12252 is important to its mechanism of action - this context would be lost with an artificial promoter.

With these studies, we were unable to determine the half life of *IFITM3* mRNA of either genotype, largely due to the limitations of the ActD assay. We saw that *IFITM3* mRNA was highly stable; as a result, this assay was unable to capture the degradation of *IFITM3* transcripts before cellular death. As a result, we were unable to observe any degradation of *IFITM3* with the systems we were given.

It may be possible to use pulse/chase labelling to examine the degradation rate of *IFITM3* mRNA with an appropriate cell system [335]. This would effectively involve extending mRNA labelling experiments performed in our transcriptional rate studies, with the Click-It kit. Here, we could remove our labelling reagent after a given time, and extract cellular mRNA at a series of time points, constructing a decay curve for labelled *IFITM3* mRNA without profoundly affecting intracellular conditions. As mentioned previously, this would be a useful follow-up experiment to perform using CRISPR-Cas9 edited cell lines.

We additionally explored methylation of rs12252 as a potential mechanism of transcriptional silencing associated with the downregulation of *IFITM3*. While methylation based transcriptional silencing is most commonly associated with gene promoters than the coding sequence of the gene itself. While literature related to exon 1 methylation in relation to transcriptional silencing is sparse, a study by Brenet *et al* found that, on a whole genome basis, methylation present on the first exon of genes is generally more strongly correlated with silencing of gene expression than methylation in any other part of the gene, including the promoter region and transcriptional start site [336]. It should be noted that there is no proposed mechanism by which this suppression may be occurring.

In addition to this body of evidence, a study by Edwards *et al* showed that in general, the majority of CpG sites in the genome are methylated, and that this is the "default-state" of CpG sites within the genome [337]. Based on this information, we reasoned that methylation of rs12252 was not an unlikely mechanism to explain the differential allelic expression observed previously.

We were unable to observe any methylation using our fibroblast cell system, although this does not rule out methylation of rs12252 existing in a tissue specific manner. However, the fact that we observed differential allelic expression using the same cell system indicated to us that methylation of rs12252 could not explain the observed reduction in expression associated with the C allele. However, it is not implausible that methylation of this locus could affect expression combinatorially with our hypothesised mechanism of *IFITM3* differential expression via rs12252.

Regardless of mechanism, it is still unusual that a synonymous coding SNP would be linked to the mRNA expression of its own gene. A number of studies have linked codon usage in the 5' coding sequence of mRNA (where rs12252 resides) to translation efficiency [338]. However, as mentioned

above, we have strong evidence of this mechanism occuring pre-translation. With our observations that rs12252 may act as a bidirectional eQTL, it is important to consider that its mechanisms of action may be tissue specific. As stated previously, we have not seen any evidence of any mechanisms explaining opposite-effect eQTLs. For this reason, any observations hinting at mechanisms underlying rs12252 based modulation of *IFITM3* expression may not be broadly applicable to all tissues.

In scientific literature, there are several instances where mRNA stability has been linked to synonymous coding SNPs mechanistically. In a 2004 Capon *et al* study, synonymous SNP in the human *CDSN* gene was shown to increase the stability of *CDSN* mRNA - this SNP was shown to incur decreased binding affinity for an unknown RNA-binding protein [339]. It was assumed that this interaction facilitated increased degradation of the *CDSN* mRNA. It is plausible that rs12252 may modulate *IFITM3* expression in this way. While we were unable to observe any relationship between rs12252 and *IFITM3* degradation, this was largely a result of experimental limitations, and does not rule it out as a mechanism. It would be possible to further investigate mRNA-protein interactions by similarly carrying out UV cross-linking analysis with regards to the *IFITM3* gene.

Overall however, our results point in favour of rs12252 affecting the transcriptional rate of *IFITM3* rather than the degradation of the mRNA. While this requires further study to confirm, we speculate that this is likely to be the result of differential TF binding affinity at the rs12252 site - this would implicate the rs12252 site as an enhancer element within a coding sequence. While exonic enhancers are relatively understudied, there is a wealth of evidence for their existence within mammalian genomes.

A 2012 study by Birnbaum *et al* found evidence of exonic DNA regions acting as enhancer elements on mouse and zebrafish backgrounds; these could influence the expression of nearby genes [340]. Additional studies by Ritter *et al.* [341] and Tümpel *et al* have shown cases where exonic enhancers can affect the expression of the specific gene they reside in [342]. In fact, the former study found that for any given exonic enhancer, it was more likely to affect the expression of its own gene than other genes. In the latter study, tissue specific enchancer activity of the *Hoxa2* gene was discovered in a mouse background. We would expect *IFITM3* to be a poor candidate gene for containing exonic enhancers, given that its locus is under positive selection [240]. However, the 2012 Ritter *et al.* study found no evidence that conservation levels were an indication of exonic enhancers.

In addition to these studies, there is also direct evidence of transcription factor binding within exons. A highly comprehensive genome wide analysis by Stergachis *et al* used a combination of DNase-I footprinting and TF specific Chromatin Immunoprecipiation (CHIP)-Seq to obtain a comprehensive overview of coding sequence enhancers [343]. In this study, it was shown that DNase-I footprints within exons were relatively rare when compared to regulatory regions. However, it was seen that roughly 15% of coding exons overlapped with DNase-I sites, and roughly 14% of coding bases were present within a DNase-I footprint. Footprint density was significantly enriched within the first exon (where rs12252 is located) when compared to other exons. The fact that these coding sequence enhancers appear to be so common lends strong support to the idea that this could be the mode by which *IFITM3* expression is modulated by rs12252. Taken as a whole, the idea that rs12252 affects the activity of an exonic enhancer for *IFITM3* appears to be a viable hypothesis, and one worth investigating in the future.

Based on our findings, and taking scientific literature into context, we believe the most effective course of further study into the rs12252 mechanism would involve an expansion of available CRISPR clones. With these clones, we could use a combination of nascent mRNA pulldown assays and DNase-I fingerprinting to identify potential differential enhancer usage at the *IFITM3* locus.

4.3.3 ISG Signature is Related to rs12252

Over the course of this investigation, we have found that the effects of rs12252 may extend beyond modulation of *IFITM3* expression. We have found evidence that rs12252 may be related to the expression of *IFITM* genes, and ISG signature in general.

When performing differential gene expression analysis, we observed differential expression of *IFI27* with respect to rs12252 under our CC/TT model. Increased expression of this gene is associated with rs12252-C. This gene is an ISG; this differential expression is in concordance with the increased IFN signature observed in CC donors. It is worth noting that this differential expression is significant in comparison to most ISGs, implicating it as a gene potentially more specifically related to rs12252. One caveat to this finding is that this gene is expressed at extremely low levels in whole blood, both in our dataset, and the GTEx dataset - this casts some doubt

on the relevance and validity of this finding. It may be that this gene, like numerous other ISGs, is expressed at extremely low levels before stimulation with IFN. It has been shown that low expression can lead to imprecision in differentially expressed gene detection [344]. While our pipeline does account for genes with low expression levels, it is possible that this filtering threshold was not high enough in this case.

While its role as an ISG is well understood, research relating to the function of *IFI27* is sparse. A 2015 study by Hsieh *et al* found that *IFI27* knockdown incurred aberrant expression levels of cyclin proteins, and eventually caused S-phase arrest in keratinocytes [345]. However, it is still unclear whether or not expression levels of this gene would associate with cell cycle progression.

Interestingly, our eQTL analysis found that there was a mild eQTL effect of rs12252 on *IFITM1* expression. We do not have the same level of confidence in this result as that observed with *IFITM3*. It's worth mentioning that the evidence for this finding is partially contradicted by our findings using CRISPR clones. However, it is interesting to consider that both *IFITM3* and *IFITMI* expression may be regulated by rs12252. Notably, we have observed associations between the C allele of rs12252 and the incidence of liver cirrhosis in Hepatitis C Virus (HCV) patients - it is the general consensus within the literature that *IFITM1* is the most important *IFITM* gene involved in the restriction of HCV. In fact, evidence for *IFITM3* restriction of *HCV* is somewhat mixed. We have made similar observations relating to rs12252-C and HIV-I progression [256]. Again, reports regarding which *IFITM* gene is most effective at restricting HIV-I infection are conflicting. Overall, it may be worth considering that these associations between rs12252 and viral infections may be mediated by both *IFITM3* and *IFITM1* expression.

When performing GO analysis on our healthy donor cohort, we found enrichment of a number of terms associated with TT and CC donors. However, as a result of the properties of this dataset, it should again be noted that our testing parameters were more forgiving than usual. For this reason, we do not have strong levels of confidence in the majority of these enriched terms. Interestingly, CC donors were associated with much more significant levels of enrichment than TT donors.

In performing GO analysis, our most enriched terms by far were related to type I IFN signalling in CC donors. This was underscored by a higher expression of terms within the "response to type I interferon" ontology on average. It is worth noting that this analysis did not show distinct ISG

expression levels bounded by rs12252 genotype; donor distributions were largely overlapping. However, this analysis did show a clear association between rs12252 genotype and clustering location. While this is an interesting result, it is difficult to determine whether or not this ISG signature is a real observable effect of rs12252 genotype, or if this observation is incidental.

Previously, an inverse relationship between IFITM3 expression and IFN signalling via IRF3 has been reported [239]. IRF3 is only considered essential for induced IFN signalling, and not the tonic IFN signalling observed in this cohort [137]. However, a recent report has suggested that, *in vitro*, cells deficient in IRF3 show a reduced level of tonic IFN signalling [346]. However, this does not clearly link to our observations in our healthy donor cohort, where high *IFITM3* expression was correlated with high, not low ISG expression.

In addition to this, a minor ISG signature associated with the CC genotype was observed in our CRISPR-Cas9 cells, albeit accompanied by a marked reduction in *IFITM3* expression. However, given that we do not have a large number of cell clones, speculation about this finding is likely somewhat spurious. Due to the fact that this analysis was done based on a single HEK293 cell clone, it is unclear whether this represents a difference in basal ISG expression as a result of the CC genotype, or whether it is an effect of the clone itself. In the context of the reduced *IFITM3* expression seen in the CC HEK293 cells however, the fact that there may be a stronger ISG signature in these cells serves to strengthen the idea that this reduced expression is an effect of the rs12252 SNP, and is not caused as the result of an abberant ISG signature in these cell clones. To investigate this further, it is incumbent upon us to produce more cell CC genotype HEK293 cells to find the cause of this phenomenon.

The finding that rs12252-C may be associated with an increased IFN signature would be unexpected given the fact that rs12252 is associated with poor outcomes in viral infection. It has been shown that a high level of basal ISG expression contributes to resistance to influenza infection in bronchial epithelial cells [347], and Simian Immunodeficency Virus (SIV) infection in rhesus macaques [348]. With this in mind, it seemed more likely that rs12252 would be associated with a reduced ISG signature. We do not currently have an explanation for the disconnect between these observations - it may be possible, again, that this is a tissue specific effect.

In our RNA-Sequencing experiment in chapter 3, and in the 2017 Allen *et al* study, distinct correlations were observed between expression levels of *IFITM1*, *IFITM2* and *IFITM3* [219]. When considering this data, it was never clarified whether this relationship was specific to the *IFITM* genes, or whether this was a product of the aforementioned ISG signature. In our large healthy donor cohort, we observed extremely strong relationships between *IFITM3* and other *IFITM* genes. However, this relationship did not appear stronger than the relationship betwen *IFITM3* expression and the expression of other ISGs. Of all genes tested, *IFITM1* expression correlated most to *IFITM3*, but the r^2 value associated with this correlation was not substantially higher than with respect to other ISGs. In addition, the correlation observed between *IFITM3* and *IFITM2* was weaker than that of many ISGs investigated. With this in mind, we believe it is more likely that observed relationships between *IFITM* gene expression levels are not specific to the *IFITM* genes. We believe that this consistent ISG signature is not unexpected - tonic IFN signalling is thought to be mediated exclusively through IFN-β [349, 350]. With this in mind, it seems likely that this relationship can be traced back to the amount of constitutive IFN-β produced in each donor. Given the fact that irregular IFN homeostasis has been associated with succeptibility to viral infection, cancer, and autoimmunity, it is possible rs12252 could be associated with disease phenotypes [140, 351, 352].

Based on our previous findings regarding expression relationships between ISG expression levels, we endeavoured to reduce a generalised ISG signal from *IFITM3* expression. We largely chose to do this to investigate whether differential *IFITM3* expression, with respect to rs12252, occurred regardless of the increased ISG signature associated with CC donors. We found that by estimating average ISG signature on a donor-by-donor basis effectively reduced the dispersion of *IFITM3* expression while retaining the rs12252 effect. This algorithm increased the significance of *IFITM3* differential expression as mediated by rs12252. In comparison, randomly selected ISGs did not show this increase in significance, further supporting the idea that the rs12252/*IFITM3* eQTL relationship is real and specific to *IFITM3*. By using *GAPDH*, a gene generally not considered to be affected by IFN [353], as a control, we showed that this algorithm was capable of specifically reducing the dispersion levels of ISGs. It is worth noting that this method is only an approximation of the effects of tonic IFN signalling on *IFITM3* expression - this factor cannot be properly controlled without real experimental evidence. It is likely that a real measurement of this influence would further decrease the dispersion of *IFITM3* expression in comparison to our algorithm. However,

we feel that this algorithm in its current state serves to strongly refute the idea that our observed eQTL effects of rs12252 on *IFITM3* are not secondary to an incidental ISG signature favouring the C allele.

Overall, our data obtained over this study supports a potential link between rs12252 genotype and ISG signature, that may be worth investigating further. However, it remains clear that the eQTL effects of rs12252 and *IFITM3* are not distorted by this trend.

4.3.4 rs12252 is Associated with Local ncRNA Expression

We finally investigated the effects of rs12252 on genes locally expressed around the *IFITM3* locus.

In the 2017 Allen *et al* study [219], it was reported that, on a genotype-by-genotype basis for rs34481144, the expression level of *IFITM3* was correlated with the expression levels of surrounding genes. It was suggested that this was the result of differential CTCF binding based on this SNP causing aberrations in a topologically associated domain boundaries [354]. It is presumed that this changed the chromatin formation, and regulatory elements present in this locus, thereby creating this observed association. We do not have genotype information for this SNP in our cohort, but, as mentioned previously, it is found at extremely low levels in EAS populations (MAF = 0.01 (1000 genomes [241])). With our data, we could confirm, with greater statistical power than the Allen study, that without reference to genotype, *IFITM3* local genes expression levels were not associated, aside from a mild correlation observed with regards to *IFITM3* and *RASSF7*. There are no reports of any SNPs common to EAS populations affecting CTCF, or other topological associated domain boundary proteins in the *IFITM3* locus, and if these SNPs were present, it may be that we would have to examine this correlation in a subset of our donor population to observe any effect of this hypothetical SNP. With our data, it seems likely that expression-level relationships do not occur naturally at this locus, and that this effect is highly genotype specific. If the chromatin remodelling hypothesis is correct, our finding serves to strengthen the idea that this is a SNP-specific effect associated with chromatin remodelling.

Perhaps the most unexpected observation made with respect to our large

healthy donor cohort was the discovery of highly differentially regulated ncRNAs within the rs12252 locus. Here, we identified five ncRNAs that were differentially regulated by rs12252. These five ncRNAs make up the bulk of all ncRNA in this locus within ensembl reference annotations, and hence counted by us. The differential regulation of these ncRNAs was, as mentioned previously, highly significant with respect to all other genes within the genome. Taken at face value, these ncRNAs suggest stark differences in transcriptional activity surrounding the rs12252 locus associated with the T and C alleles.

To judge the likelihood that these transcripts are truly expressed as part of the human transcriptome, it is important to understand the evidence that led to these annotations in the first place. GENCODE 27 annotations categorise all transcripts into TSLs [316]. Briefly, this is a scale ranging from 1 (TSL1) to 5 (TSL5), where TSL1 has the highest level of evidence, and TSL5 has the lowest [355]. In a number of cases, most notably when a transcript only has a single exon, no TSL level is assigned (TSLNA). Within each feature, each isoform has an individual TSL, meaning not all isoforms are as likely as one another to exist. We have outlined the TSLs for each of these ncRNAs in table 4.8.

Table 4.8 ncRNAs TSLs

Transcript	**TSL** [1]
AC136475.1	TSL1
AC136475.4	TSL5
AC136475.5	TSL3
AC136475.9	TSLNA[2]
AC136475.8	TSL5

[1] Highest TSL ncRNA
[2] TSL not analysed

Overall, the most exists for the AC136475.1 and AC136475.5 transcripts. AC136475.5 is supported by a single Expressed Sequence Tag (EST); these are generally older sequences generated by random amplification and sequencing of poly(A) tails and 3' UTRs of an mRNA pool. Again, the presence of only a single supportive EST suggests that the likelihood of this transcript existing is not extremely high. On the other hand, AC136475.1 is supported by multiple ESTs and mRNA sequences. In addition, the mRNA sequences specifically support the splicing model presented for this transcript via exon-exon junctions. It is worth noting that this ncRNA only

attained TSL1 for one of its three isoforms, ENST00000602429.1. Further investigation of the sequences that have provided evidence for AC136475.1 show that EST and mRNA sequences have been derived from placenta tissue, embryonic stem cells, and brain tissue [356].

To further investigate these transcripts, we viewed the sequence alignment to each of these regions. Firstly, two of these ncRNAs, AC136475.4 and AC136475.8, showed extremely low levels of expression. These alignments were indistinguishable from base levels of background alignment. In contrast, the expression levels of the remaining three ncRNAs - AC136475.1, AC136475.5, and AC136475.9 - were substantially higher, albeit still low relative to the average gene. These low expression levels are not unexpected - it is well established that, in general, ncRNAs are expressed at significantly lower levels on average than mRNA [357]. However, when visualising read alignment to these features, it was again unclear whether or not these aligned reads truly represented real production of the annotated transcripts, or if alignment was incidental. Again, we had the most confidence in our observed alignments to AC136475.1 - the most differentially expressed gene in this dataset.

To further investigate the existence of AC136475.1, we used qPCR to investigate its expression in our CRISPR-Cas9 system used for the study of rs12252. Here, we could observe, specific amplification of a PCR product corresponding to this ncRNA, consistent with the low expression levels observed for this transcript using RNA-Seq. The fact that the qPCR primers used to amplify this product spanned an exon-exon junction suggests that our observed alignments do correspond directly to the predicted ncRNA transcript. In this experiment, we did not observe differential expression of this transcript with respect to rs12252 genotype or IFN stimulation - it should be noted that amplification of this transcript occurred at high cycle counts, where qPCR is less sensitive. If true, this finding could indicate that the association between rs12252 genotype and AC136475.1 expression is due to linkage effects, or is tissue specific. Overall, this finding lends support to the idea that this transcript is truely expressed in a physiological context.

In addition to examining genome annotations, we briefly investigated expression levels of these ncRNA in the GTEx database [328]. Like with our samples, it is likely that some of these expression levels were counted as a result of alignment not specific to this feature. Here, there was little difference in expression levels on a tissue to tissue basis for AC136475.1 and AC136475.8. No expression was discovered for AC136475.4. However,

expression of AC136475.5 and AC136475.9 was highly specific to whole blood, and to some extent, spleen tissue. With our hypothesis that rs12252 differentially regulates *IFITM3* expression on a tissue-by-tissue basis, it is interesting to consider that these ncRNAs are potentially differentially expressed, in a tissue specific manner, with relation to rs12252. It is possible that this finding may hint at our hypothesised tissue specific mechanism of rs12252.

Despite the fact that alignments to these annotated features do not give strong evidence for the production of these specific ncRNAs, we still observe high differential alignment to these features by rs12252 genotype; an unlikely occurrence if this finding was strictly the result of technical noise. STAR, our genome alignment software of choice has a false positive alignment rate of well under 1%, suggesting that the majority of noise is not technical, but biological in nature [275]. Transcriptional noise is poorly understood, especially with regards to RNA-Seq alignment. There is a general agreement that alignment noise is biological in nature, the common causes of this variation are still a source of debate. It may be that our observed noise is derived from random RNA polymerase binding, unprocessed transcripts or undiscovered genomic features [358]. It is worth noting that the library preparation of this sample set involved poly(A) selection, meaning these observed would contain a poly(A) tail. This is not at all unusual for ncRNAs; it is more likely than not for ncRNA to have a poly(A) tail [214]. Overall, we argue that the source of this differential ncRNA/noise expression is potentially indicative of differential activity of this locus, regardless of its accuracy in relation to the annotated genome.

Taken as a whole, our evidence corroborated with evidence available on genomic databases suggests that AC136475.1 expression is likely to be real and highly associated with the rs12252-T compared to the rs12252-C allele. With this in mind, we considered whether or not AC136475.1 could explain the differential expression of *IFITM3* with respect to rs12252 we observed. AC136475.1 is annotated as an antisense RNA. It is generally accepted that antisense RNAs function to control gene expression, although this has not been specifically demonstrated with respect to AC136475.1. If AC136475.1 was capable of restricting *IFITM3* expression, this would make it a Cis-Natural Antisense Transcript (cis-NAT) - acting in cis rather than trans.

One of the mechanism of cis-NAT interference is somewhat plausible in the case of *IFITM3* - this is the transcriptional collision model. Here, two

RNAP-II complexes transcribing in opposite directions on opposite DNA strands collide, terminating transcription of the two genes. Experimental evidence has shown that this occurs naturally *in vitro* and *in vivo*, and collision events have been observed using microscopy in *E. coli* [359]. When collision events occur, polymerases are ubiquitinated and degraded [360]. This model could apply to *IFITM3* and AC136475.1.

In the case of our healthy donor cohort, the expression levels of AC136475.1 showed no correlation with *IFITM3* expression, casting doubt on the idea that this transcript is capable of mechanistically affecting *IFITM3* expression. With this in mind, we currently believe that this transcript's differential expression with respect to rs12252 does not explain the differential expression of *IFITM3*, although it may give an indication of differing transcriptional control in the region.

4.3.5 Concluding Remarks

Over the course of this investigation, we have obtained a wide range of evidence that rs12252 is an eQTL for *IFITM3* through the use of *in vitro* and *in vivo* systems. While it seems clear that this association is present, the mechanism by which rs12252 is associated with *IFITM3* is elusive. In addition, eQTL effects do not appear to be consistent between sample types - we believe that the effects of rs12252 are likely to be dependent on context in which they are observed. Further investigation is required to fully understand where this SNP is associated with higher or lower expression.

In addition, we have discovered potential links between rs12252, *IFITM3* expression, *IFITM1* expression, and basal ISG signatures. It is interesting to consider that this SNP may affect a more complex network of gene expression than the gene it is located within. Again, further investigation of the relationship between rs12252 and ISG networks could help to fully understand the context by which this SNP is associated with severe viral infection.

Chapter 5

Identification of Alternative eQTLs for *IFTIM3*

5.1 Introduction

In the 2013 Shen *et al* and 2017 Allen *et al* papers, substantial evidence was put forward that differential expression of the *IFITM3* gene was a determinant of susceptibility to severe influenza infection [211, 219].

In chapter 4, we showed that rs12252 acts as an eQTL for *IFITM3*. Despite this, further understanding of its potentially bidirectional nature are required to fully understand this. These findings support the idea that *IFITM3* expression, is a determinant in the susceptibility of patients to severe influenza. Taken as a whole, these findings led us to speculate that the expression of *IFITM3*, or perhaps the sum of a number of genetic variants, could have more predictive power in terms of severe influenza susceptibility than a single variant.

Recently, it has become clear that eQTLs are important factors in considering genetic risk for disease. Often, eQTLs have been directly called, or are highly associated by LD with variants discovered by GWAS - less than 5% of all GWAS hits confer changes in gene protein sequence [361]. In a landmark 2017 study by Marigorta *et al*, GWAS and eQTL data was integrated in an attempt to link gene expression to disease risk in Crohn's disease patients [362]. This study identified a number of candidate risk SNPs which were

strongly associated with, or were eQTLs for disease linked genes. The authors found that, for these genes, a score based on the expression of these genes was more predictive of disease risk than the genetic variants themselves. This finding has strong implications for the interface between disease risk and genetics.

Despite the idea that transcriptional risk is strongly predictive of disease, obtaining accurate expression data for genes like *IFITM3* would be difficult in the context of acute infection, given their variable expression under the influence of IFN and other cytokines associated with disease. With these considerations in mind, we postulated that identifying a panel of eQTLs for *IFITM3*, rather than rs12252 specifically, could be an important step in establishing a model for *IFITM3*-related risk for severe influenza. We decided that our large healthy donor cohort, used in chapter 4, would provide an ideal opportunity to further characterise the relationship between expression of *IFITM3* and genetic factors. This approach could potentially provide an expanded panel of predictive SNPs for *IFITM3* expression, which could be used in tandem for a more accurate picture of genetic risk for severe influenza.

In this short chapter, we perform preliminary exploratory analysis of novel eQTLs for *IFITM3* specific to East Asian populations. It is our hope that this could lead to novel clinical investigations under a new, multi-SNP, model of risk for severe influenza.

5.1.1 Project Aims

The aims of this project were as follows:

1. Use SNP typing, in combination with RNA-Seq data to identify new potential eQTLs for *IFITM3*
2. Investigate any discovered eQTLs to determine their potential for associations to severe disease.

5.2 Results

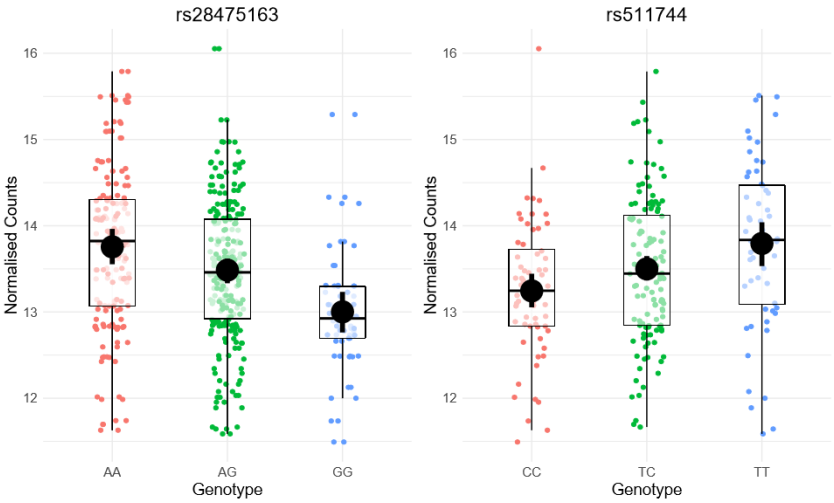
5.2.1 eQTL Analysis for *IFITM3*

Using samples from our healthy donor cohort, we performed whole-genome SNP typing using the Axiom Medicine Research Array. This conferred typing of over 900,000 SNPs. SNP typing was done with matched DNA samples sent from Beijing You'An Hospital to BGI. SNP array typing was performed for 236 of the 274 individuals in the healthy donor cohort. We used the 236 samples with both expression and genotype data for this analysis. Within this cohort were 176 female donors and 60 male donors. The median age was 27, and the mean age was 28.5.

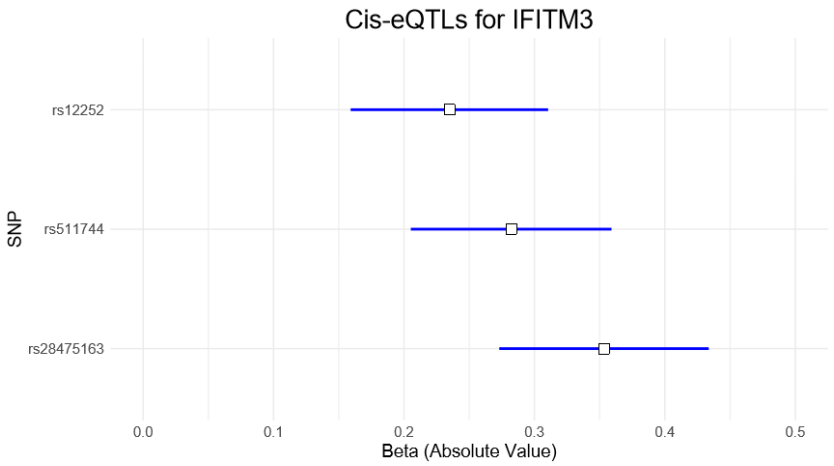
With this data, we used MatrixEQTL to identify novel eQTLs for *IFITM3*. We limited the analysis to only *IFITM3* expression, and looked for genes associated in both cis and trans. Like with section 4.2.5, we used an α of 0.1. For cis-eQTLs, we limited the maximum genomic distance to 1Mbp. We accounted for donor age and sex by inputting these as covariates. Performing this analysis, we found evidence for two cis-eQTLs and no significant trans-eQTLs. One of the two cis-eQTLs was associated with *IFITM3* expression at a higher significance level. These SNPs were rs28475163 ($p = 0.00840$) and rs511744 ($p = 0.0771$) (fig. 5.1a).

In 1000 genomes, rs511744 is relatively evenly distributed among world populations, whereas allele frequencies for rs28475163 are somewhat variable between populations. For both SNPs, EAS allele frequencies are almost evenly distributed, with 56% (rs28475163) and 54% (rs511744) MAFs. The population allele frequencies of both SNPs are shown in table 5.1. In our cohort, we had 79 AA donors, 122 AG donors, and 33 GG donors for rs28475163. We had 60 CC donors, 120 TC donors, and 56 TT donors for rs511744.

The rs28475163 SNP has A (reference) and G (alternative) alleles. Higher expression is associated with the A allele. rs511744 SNP has C (reference) and T (alternative) alleles. Higher expression is associated with the T allele. The rs28475163 SNP is in close proximity to *IFITM3* and rs12252, located at position 327143 on chromosome 11, 6.3kb from rs12252. Conversely, rs511744 is located at position 219089 on chromosome 11, placing it roughly 100kb from rs12252.



(a) Effect Sizes of *IFITM3* eQTLs



(b) Scatter Chart of Beta Values

Figure 5.1: ***a.*** *Expression of IFITM3 as delimited by the rs28475163 and rs511744 eQTLs. IFITM3 expression is shown in variance stabilised normalised counts. Both SNPs are significantly associated with the expression of IFITM3.* ***b.*** *Comparison between effect sizes (beta) of the three identified IFITM3 eQTLs. While beta estimated shared similar standard errors, rs28475163 was considered the strongest eQTL, and rs12252 the weakest.*

Table 5.1 Global Allele Frequencies of rs28475163 and rs511744

Population	rs28475163		rs511744	
	G Frequency	**A Frequency**	**T Frequency**	**C Frequency**
AFR	0.974	0.026	0.225	0.775
AMR	0.705	0.295	0.334	0.666
EAS	0.437	0.563	0.539	0.461
EUR	0.739	0.261	0.296	0.704
SAS	0.782	0.218	0.264	0.736
All	0.744	0.256	0.325	0.675

We compared these SNPs to rs12252, to contextualise their associated beta coefficients on *IFITM3* expression. We found that the beta coefficients for each SNP had similar standard errors. However, the effect sizes of rs28475163 (0.353) and rs511744 (0.282) were higher than the effect size for rs12252 (0.235) (fig. 5.1b). This means, in the context of this cohort, the genotype for these SNPs had a greater degree of influence over *IFITM3* expression than rs12252.

To investigate whether these SNPs worked cooperatively to influence the expression of *IFITM3*, we generated a linear model, where we examined the relationship between the total number of high expression alleles present in each donor against their expression of *IFITM3*. As stated previously, the high expression alleles were rs12252-C, rs28475163-A, and rs511744-T; each patient therefore had between 0 and 6 high expression alleles. We found that there was a significant association between number of high expression alleles and *IFITM3* expression ($p = 1.16 \times 10^{-}6$, fig. 5.2).

Despite the fact that we discovered no significant trans-eQTLs (with a p-value threshold of 0.05), we hypothesised that, when taken together, genotype for non-significant trans-eQTLs could predict *IFITM3* expression. To test this, we built a model of expression similar to the linear model in fig. 5.2. In this case, we used the top ten most significant trans-eQTLs for *IFITM3*, shown in table 5.2. When performing regression analysis comparing the number of high-expression trans-eQTL alleles against *IFITM3* expression, we showed an extremely strong relationship between genotype and expression ($p < 1 \times 10^{-16}$) (fig. 5.3).

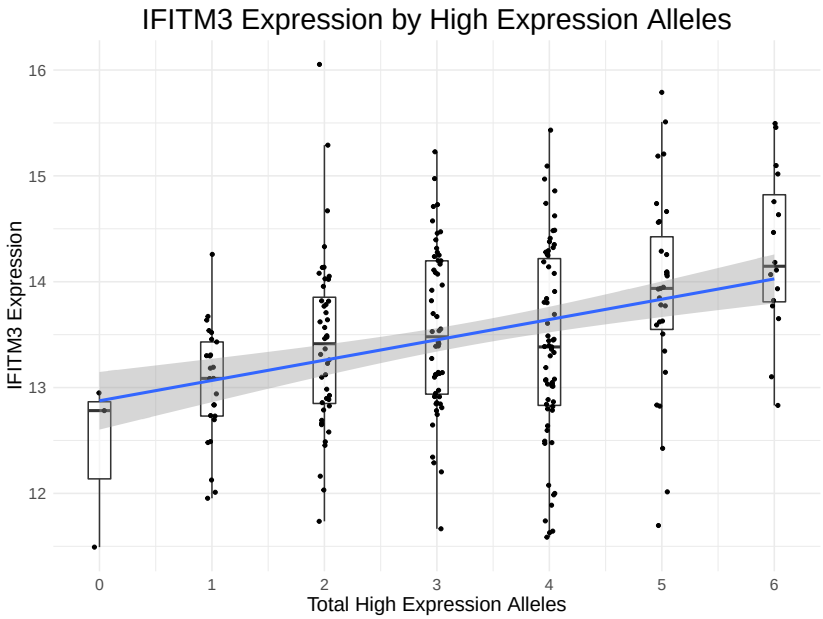


Figure 5.2: Expression of IFITM3 plotted against the total number of high expression alleles. IFITM3 exprssion is shown in variance stabilised counts. IFITM3 expression significantly increases with the total high expression alleles. Blue line represents linear regression line, with grey shading representing 95% CI.

Table 5.2 trans-eQTLs for IFITM3

SNP	**Locus**[1]	**p-value**	**p-value (Adjusted)**
rs1504920	12:118221470	7.87×10^{-7}	0.404
rs35419852	9:133717745	1.05×10^{-6}	0.404
rs1530541	12:118229731	2.21×10^{-6}	0.404
rs62124854	2:23225246	2.57×10^{-6}	0.404
rs3743338	15:36984409	2.83×10^{-6}	0.404
rs1710770	1:106032557	3.05×10^{-6}	0.404
rs3755955	4:994414	3.55×10^{-6}	0.404
rs5992978	22:18530208	3.58×10^{-6}	0.404
rs138890527	4:90173621	5.74×10^{-6}	0.494
rs114216050	1:187306879	6.15×10^{-6}	0.494

[1] chromosome:locus

We also found that rs28475163 was an eQTL for a number of other genes, namely ncRNAs proximal to the *IFITM3* locus. This SNP is an eQTL for four transcripts other than *IFITM3*, as outlined in table 5.3. Interestingly, three of these four ncRNAs were also flagged with regards to rs12252. Like with *IFITM3*, rs28475163 shows an even stronger association with

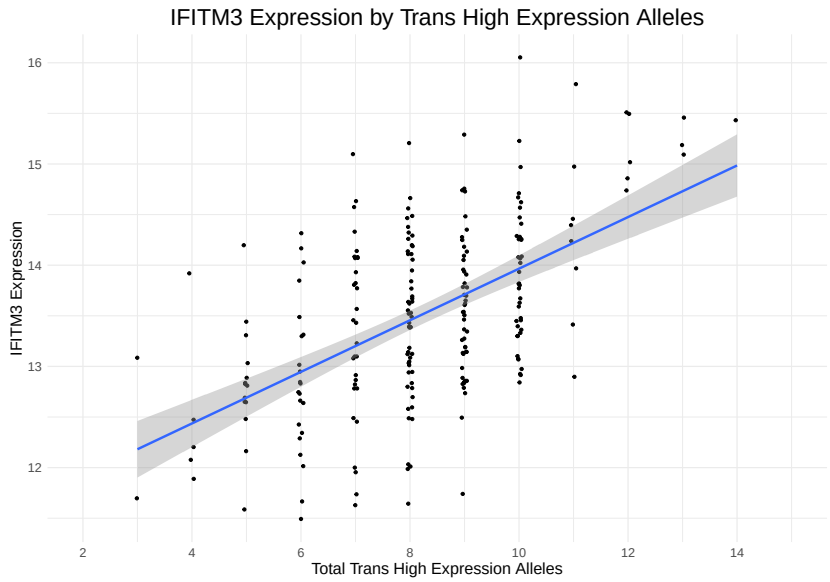


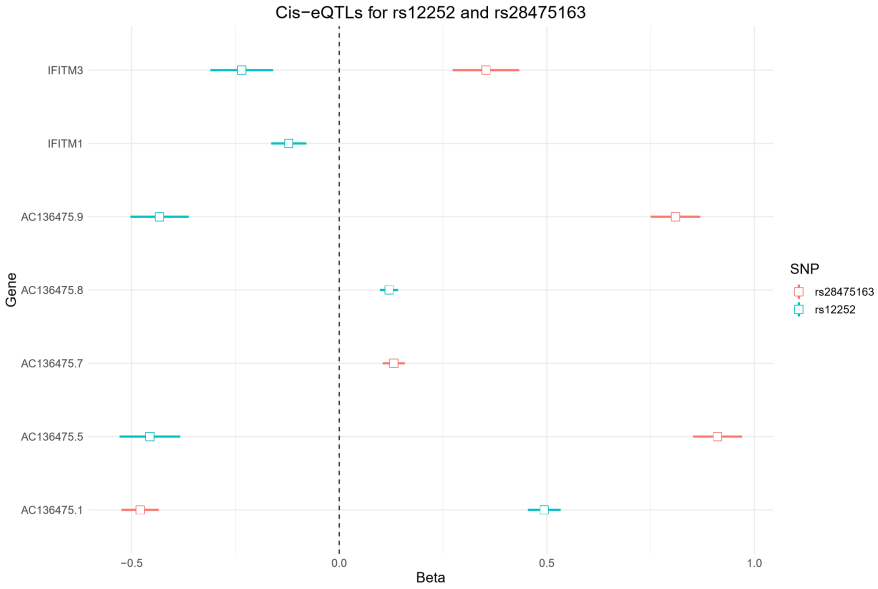
Figure 5.3: IFITM3 expression of each donor, plotted against the number of high-expression associated trans-eQTL alleles for IFITM3. Relationship between the two factors, as modelled by linear regression is highly significant. Blue line represents linear regression line, with grey shading representing 95% CI.

expression of these transcripts than rs12252 (fig. 5.4a). It is also worth noting that there is complete parity in the transcripts modulated by rs12252 and rs28475163 - for each transcript, rs28475163-A confers the same directionality of expression as the rs12252-C allele (fig. 5.4b).

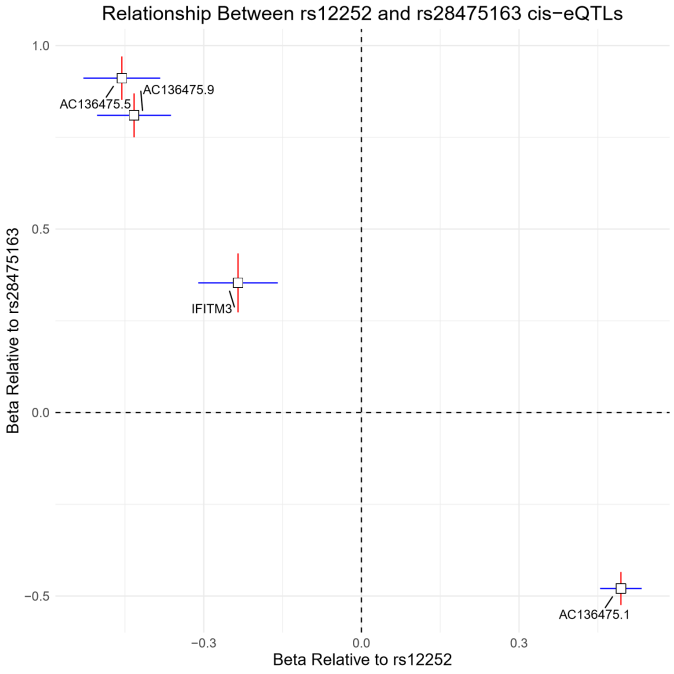
Table 5.3 Significant cis-eQTLs for rs28475163

Gene Name	**p-value**	**p-value (Adjusted)**
AC136475.5	2.05×10^{-37}	3.42×10^{-33}
AC136475.9	3.41×10^{-31}	3.70×10^{-27}
AC136475.1	8.99×10^{-22}	4.59×10^{-18}
AC136475.7	1.49×10^{-6}	9.26×10^{-4}
IFITM3	1.65×10^{-5}	0.00840

Subsequently, we checked whether the eQTLs found by this analysis were also present in the GTEx database. We found that, in GTEx, rs28475163 was an eQTL for *IFITM3* in whole blood. Again, high expression in this case was linked to the A allele (fig. 5.5).



(a) eQTL Forest plot



(b) Scatter Chart of Beta Values

Figure 5.4: ***a.*** *Forest plot showing comparison betweeen rs12252 and rs28475163 modulation of transcripts where significant. Beta values are in proportional to the effect size of the SNP. Bars show standard error of beta measurements.* ***b.*** *Scatter plot of effects of transcripts modulated by both rs12252 and rs28475163. Standard errors are shown in blue (rs12252) and red (rs28475163). Both SNPs modulate ncRNAs in similar ways.*

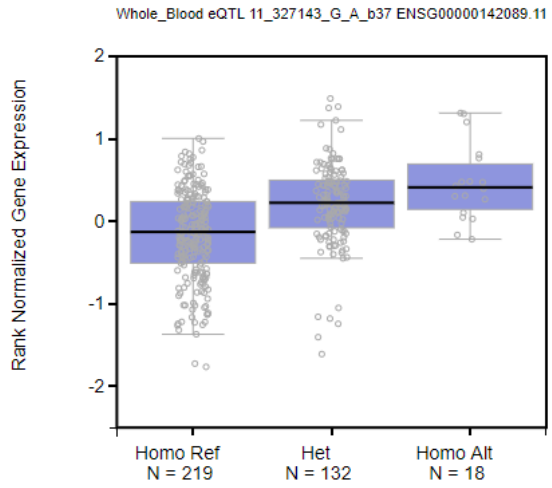


Figure 5.5: The rs28475163 is an eQTL for IFITM3 within the GTEx dataset, in whole blood. In this case, the ref allele is G, and the alt allele is A. Higher expression of IFITM3 is associated with the A allele, as with our dataset.

5.2.2 Genomic Context of *IFITM3* eQTLs

We next explored the genomic locus of the two *IFITM3* eQTLs to better their relationship to *IFITM3* expression. It is important to recognise that this is merely exploratory, and any potential functional effects of theses SNPs would need to be backed up with experimental evidence.

The rs511744 SNP is located, as previously mentioned, relatively distally from *IFITM3*, at locus 219,089 of chromosome 11. This SNP is located within an intron of the *SIRT3* gene. It does not overlap any regulatory features, such as enhancers or promoters, called by the ensembl regulatory build, or H3K27Ac markers.

The rs28475163 SNP on the other hand, is located in close proximity to *IFITM3*, although it does not reside in the canonical promoter. It is present at locus 327,143 of chromosome 11. This variant does not overlap with any protein coding genes or ncRNAs, but does overlap with H3K27Ac markers. This SNP also falls within an ensembl regulatory feature, termed ENSR00000035602, described as a promoter flanking region implicated in the binding of a number of transcription factors, most notably Upstream Stimulatory Factor 1 (USF1) and PU-1. rs28475163 is directly adjacent to TF regulatory motifs. In the ensembl regulatory build, this feature is considered

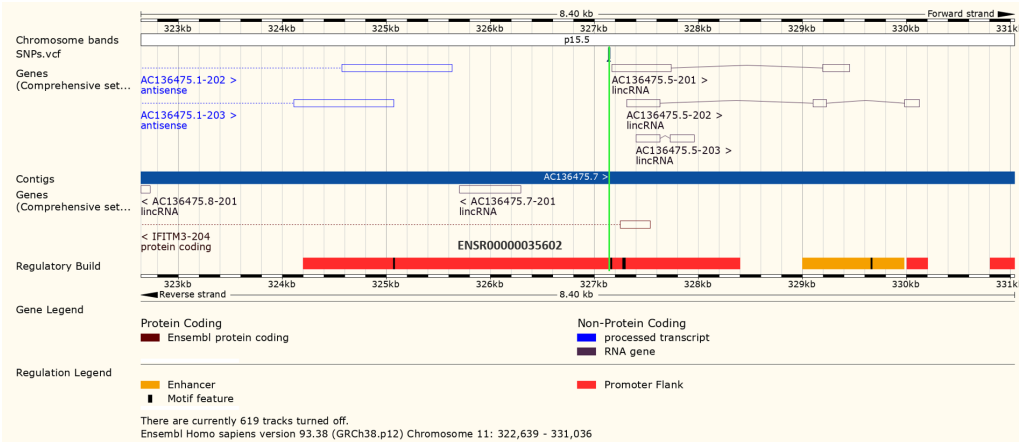


Figure 5.6: Annotation of rs28475163 (green line) on the ensembl genome browser. SNP does not overlap any ncRNAs, although it does lie in close proximity. The ENSR00000035602 regulatory region is shown in red. Motif features are shown in black. The rs28475163 SNP does not overlap with these motifs, although it is adjacent.

inactive in 41 cell types, and active in 27 cell types. Active cell types include many immune cells, including NK cells, T cells, B cells, monocytes, macrophages and neutrophils. This is consistent with our findings in whole blood. Inactive cell lines include lung cells and A549 cells. Evidence for this was accumulated by ensembl, from the ENCODE, BLUEPRINT, and Roadmap Epigenomics studies [276, 363–366].

5.2.3 Population Genetics of *IFITM3* eQTLs

We next investigated any potential LD between any of these three SNPs. We did this using a combination of 1000 genomes and our SNP array. Calculations from 1000 genomes are based on haplotype data, where genotypes are chromosome specific. This was not the case for our SNP array; we performed LD calculations based only on genotype. rs511744 was not found to be in LD with either of the other SNPs, both via 1000 genomes and our SNP array. However, rs12252 and rs28475163 were in weak LD, likely due to their close genetic distance. Alleles associated with high expression of *IFITM3* were correlated, (rs12252-C with rs28475163-A). These statistics are shown in table 5.4.

To investigate whether the effects of these SNPs were potentially attributable

Table 5.4 LD Statistics Between rs12252 and rs28475163

Cohort	r^2	**D'**
EAS[1]	0.285	0.574
CHB[1]	0.336	0.647
Healthy Donor Cohort	0.303	0.604

[1] Via 1000 genomes

to LD, we used both the 1000 genomes database and the SNP array to investigate the potential proxy SNPs. Firstly, we used the 1000 genomes database to identify proxy variants for rs28475163 and rs511744. In doing this, we used all (pooled) cohorts, the EAS cohort, and the CHB cohort from 1000 genomes. As with rs12252, we used an r^2 value of 0.7 as our association threshold.

We first performed this analysis for rs511744. Here, we found that this SNP had two proxy variants, shared between all pooled cohorts and the EAS cohorts. This was not true for the CHB cohort. These SNPs, outlined in table 5.5, are both located within 7kbp downstream of rs511744.

Table 5.5 Proxy Variants for rs511744

SNP	r^2 **(All)**	r^2 **(EAS)**	r^2 **(CHB)**
rs3825075	0.811	0.743	0.676
rs12574034	0.795	0.740	0.676

Subsequently, we found that rs28475163 was strongly associated with eight SNPs in all cohorts, sixteen variants in EAS, and sixteen variants in CHB. The sixteen high-LD variants in EAS and CHB were shared - these were the same SNPs. The eight SNPs in high LD within all cohorts were all within these sixteen variants. All of these variants were found within a roughly 10kb block upstream of the canonical *IFITM3* promoter. This is indicative of a strong haplotype within this region associated with rs28475163. Nine of these variants in the EAS/CHB cohorts, and four within all cohort distinctly overlapped with the ENSR00000035602 regulatory feature, alongside rs28475163 (fig. 5.7). In addition to this, one of these SNPs, in high LD with all cohorts, overlapped with enhancer region ENSR00000035603.

We also used our SNP typing data to calculate LD statistics for rs28475163, rs12252, and rs511744. We did not find any variants in high LD with rs511744 or rs28475163 using this analysis. With rs12252, we identified one proxy variant (fig. 5.8). This SNP was rs7948108; this variant was previously found

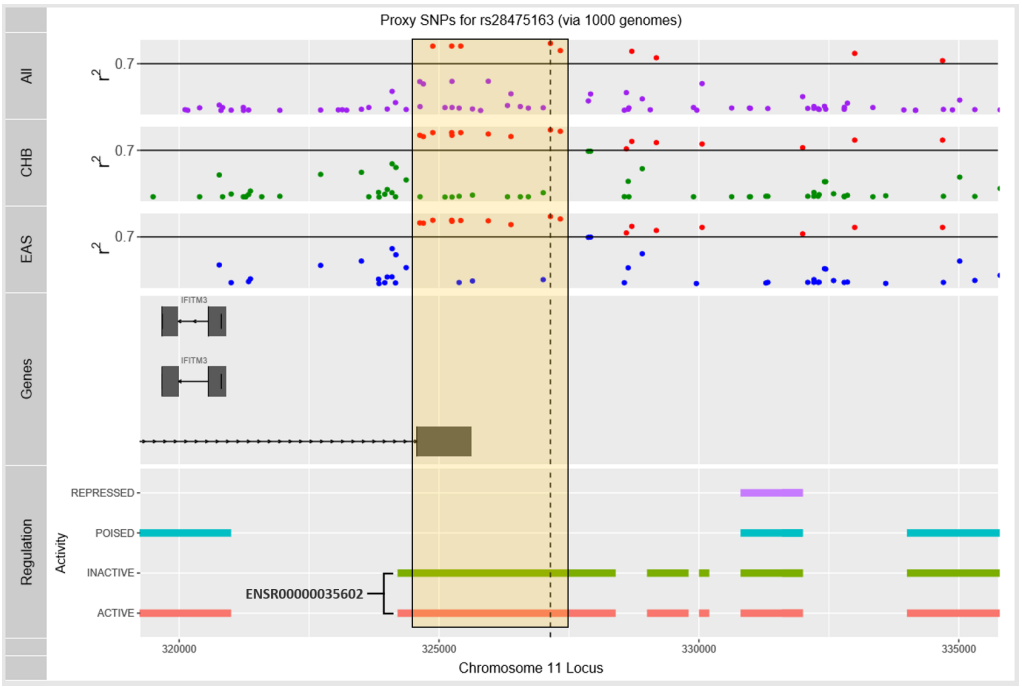


Figure 5.7: High LD SNPs for rs28475163 (dashed line), shown in relation to gene and regulatory annotations in the current genome build. The top three tracks show LD statstics relative to all 1000 genomes cohorts, the CHB cohort and the EAS cohort. The next track (descending) shows gene annotations via GENCODE 27. The Final track shows ensembl regulatory features, coloured in reference to activity associated with them. High LD SNPs overlap distinctly with the ENSR00000035602 regulatory feature, as shown via the orange box.

and discussed in section 3.2.5. Within our healthy donor cohort, rs12252 and rs7948108 had an r^2 value of 0.773.

5.3 Discussion

5.3.1 *IFITM3* Expression is Associated With Two Novel SNPs

To date, three different SNPs have been associated with *IFITM3* expression and susceptibility to severe infection. In this chapter, following from our findings that rs12252 is an eQTL for *IFITM3*, we attempted to identify

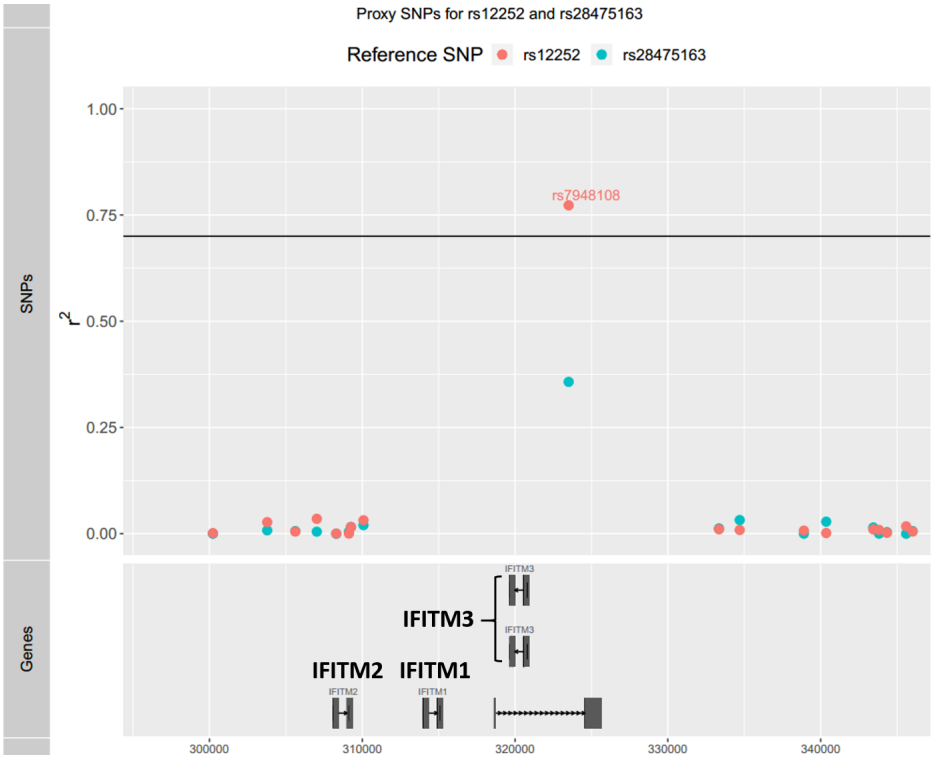


Figure 5.8: Calculated LD values for SNPs in the IFITM3 genomic neighbourhood, relative to both rs12252 and rs28475163 via SNP array. The top track plots shows, for each SNP tested, the calculated r^2 value plotted against the genomic locus of each SNP. One SNP was found to have an r^2 value greater than 0.7 in reference to rs12252. This SNP (rs7948108) has been called before using the 1000 genomes database. Gene mapping as plotted on the bottom track.

novel *IFITM3* eQTL within East Asian populations. In this study, we have shown, using a combination of RNA-Sequencing and SNP array data, that rs28475163 and rs511744 are eQTLs for *IFITM3*, and potentially modulate its expression.

In this study we were limited by the fact that our SNP array tested for roughly 900,000 SNPs, meaning many SNPs called by 1000 genomes or otherwise were overlooked. It is not possible to capture all SNPs without full-genome sequencing. For us, this strategy limited the SNP density in areas proximal to *IFTIM3*. Given this, it may be beneficial in the future to perform targeted genome sequencing around the *IFITM3* locus to capture other potential cis-eQTL variants, and potentially uncover new LD associations. However, we would not argue that this would be have been a more useful expenditure - the use of SNP array technology opens this dataset up to much

broader uses outside the context of *IFITM3*, and enabled us to investigate the potential of *IFITM3* trans-eQTLs.

Another notable limitation of his cohort is that it was performed using whole blood, and not a tissue more suitable for severe influenza. However, as mentioned previously, we were limited by samples that were available. Studies have shown that eQTLs are often consistent across tissues - a 2010 mouse study by van Nas *et al* found that cis-eQTLs are reproducible in the majority of tissues [367]. While lung tissue would have perhaps been the ideal tissue to use, contributions of blood cell subsets to the control of influenza infection should not be discounted. Both the 2017 Allen *et al* and 2014 Infusini *et al* studies have argued for *IFITM3* expression in immune cell subsets acting as determinants of infection severity [219, 368].

After identification of our three eQTLs for *IFITM3*, it was apparent that the newly discovered eQTLs both had stronger effect sizes than rs12252. This does not necessarily imply that these SNPs are stronger predictors of susceptibility to severe influenza infection than rs12252.

In this study, we generated two simple predictive models for *IFITM3* expression. In both these cases, the number of high expression alleles, of eQTLs both in cis and trans with *IFITM3* was strongly associated with *IFITM3* expression. These models could form a basis from which to test infection cohorts for SNPs associated with severe influenza infection. These models suggest independent modulation of *IFITM3* expression by each individual SNP. It is important to recognise that testing these models of cumulative genetic risk in the same cohort that the eQTLs were derived from means that their positive results are somewhat self-fulfilling. Ideally, in the future, these models would be used with unbiased data to test their true validity.

To our knowledge, models like this, with allele scores have been used many times in evaluation between genetic markers and phenotype [369]. However, these scores are generally generated for a panel of risk genes, rather than a single gene as we have done. There is some evidence that expression levels of other host genes could indicate risk of severe infection [370]. This evidence however, is extremely limited compared to that pertaining to *IFITM3*. In the future, using this dataset, we could work towards constructing a hypothetical model that incorporated a larger number of variants associated with expression of other genes.

Interestingly, rs28475163 has previously been flagged in a 2016 GWAS study by Astle *et al* [261]. Here, rs28475163-A was found to be associated with reduced percentages of Eosinophils within granulocytes and white blood cells, and increased percentages of neutrophils within granulocytes. It is possible that this genotype affected the composition of whole blood used for analysis, which may have influenced our results. This effect is likely to be small - in this study, the beta coefficient values of these increases/decreases were relatively minor (below 0.04 in all cases), and therefore likely did not impact the overall whole blood signature in a meaningful way. This could be investigated in the future by looking for rs28475163 associated differential expression of neutrophil and granulocyte markers.

When looking at the *IFITM3* cis-eQTLs on a singular level, we have a greater degree of confidence in the reproducibility of the effects of rs28475163 than rs511744. This is based on the fact that rs28475163 was more strongly associated with *IFITM3* expression, and is closer to the locus of *IFITM3*. On top of this, it is located on a defined regulatory feature, and was also flagged in the GTEx study, lending more confidence to this finding. However, it is still worth investigating potential effects of rs511744 in the future, given its potential to expand our knowledge of *IFITM3* expression modulating variants.

5.3.2 The rs28475163 SNP is Associated with Regulatory Features and Differential ncRNA Expression

In searching for rs28475163 proxy SNPs, we discovered a number of proxy variants in high-LD via 1000 genomes. These were conserved within East Asian populations. This well defined haplotype means, for the purposes of uncovering a mechanism, the effects of rs28475163 could be attributable to any number of SNPs within this haplotype. There is no way to disentangle these effects in this dataset; further studies using CRISPR-Cas9 mediated genome editing would be required to evaluate each SNP individually.

We also found strong evidence that rs28475163, and a number of SNPs in LD with it overlap a regulatory featured annotated as ENSR00000035602 within the ensembl regulatory build. This feature was generated by accumulated data from three studies, and has been shown to be active in 27 cell lines of the 68 possible. This evidence gives us a great deal of confidence in the

fact that it was annotated accurately. However, the details of its regulation are still somewhat unclear - it is not understood what effect this promoter flanking region has on *IFITM3* expression, if at all. However, it has been shown that PU-1, one of the TFs associated with this regulatory marker can induce *IFITM3* expression [371]. With this in mind, a potential mechanism of action could involve rs28475163, or SNP in high LD with it, altering a PU-1 binding site within this regulatory region, resulting in higher/lower expression of *IFITM3*. Genome editing of rs28475163 with CRISPR-Cas9 could again be used to generate an *in vitro* model to investigate this.

It is of some concern that ENSR00000035602 was not considered active in lung or A549 cells. It is likely that *IFITM3* expression levels would be important in the lung tissue in the context of influenza infection. However, as mentioned previously, the role of this feature in the context of immune cells should not be discounted in the context of IAV infection. In addition, it is possible that this regulatory feature would be active in the context of cytokine stimulation and infection within the lung.

It is also interesting to note that rs28475163 was associated with differential expression of ncRNAs near the *IFITM3* locus. These ncRNAs were discussed extensively in section 4.3.4; it is unclear what these ncRNAs indicate from a mechanistic perspective. The fact that rs12252-C and rs28475163-A regulate the expression of these ncRNAs with the same directionality is also notable. Under the hypothesis that differential ncRNA expression is indicative of differential regional transcriptional activity, this observation may indicate mechanistic similarities between rs12252-C and rs28475163-A. It is unlikely that these would have arisen via LD; the SNPs are not strongly linked. As with rs12252, it is difficult to judge the significance of this observation without additional experimental evidence.

5.3.3 Concluding Remarks

Looking forward, we have shown that genotypes of *IFITM3* eQTLs can be used in combination to model *IFITM3* expression more accurately than single SNPs. We have also accumulated evidence that the rs28475163 and rs511744 variants are eQTLs for *IFITM3*. Based on their genetic contexts, it seems more likely that rs28475163 would be a marker for risk of severe influenza, given its stronger mechanistic evidence, closer location to *IFITM3*, and higher significance of association.

Looking forward, we would like to use our previous cohorts, namely our severe influenza cohort from the 2009 H1N1 pandemic, and acute HIV-1 cohort, to determine whether these SNPs are risk factors for severe influenza. This information could be combined with previous rs12252 genotyping to investigate whether these SNPs could confer risk in an additive manner. Determination of novel risk variants for severe influenza infection would be an important step in getting a more holistic view of genetic risk in this context. Looking further forward, if we can determine an association between these novel eQTLs and severity of influenza infection, we can proceed with mechanistic studies, especially with regards to rs28475163.

Finally, in the future, we can use a more holistic genetic risk score to construct a preliminary model for influenza risk, incorporating a larger number of genes and eQTLs. This could potentially be used as a model for genetic risk moving forward.

Chapter 6

The Effects of rs12252 on Virus Infection

6.1 Introduction

As our final project we chose to investigate the effects of rs12252 on IAV infection. While results from chapter 3 and chapter 4 have shed some light on potential mechansisms by which rs12252 modulates *IFITM3*, there is a disconnect between these effects and its association with severe infection. In this chapter, we attempt to bridge this divide.

The effects of *IFITM3* with respect to infection have been well documented across a wide range of viruses. Studies using the *ifitm3*-KO mice are in some ways, a good point of reference for the effects of rs12252. While this SNP is not equivalent to a KO, our past work has shown that this SNP likely has eQTL effects for *IFITM3*; it is possible that effects of low *IFITM3* expression can be captured in a exaggerated form using a knockout. While rs12252 cannot be recapitulated in mice, the *ifitm3*-KO model is advantageous in that, unlike in humans, it allows for in-depth profiling of the effects of infection.

There have been numerous studies that use the *ifitm3*-KO mouse model in the context of IAV infection. Everitt *et al* and Bailey *et al* both showed higher death rates, higher mouse weight loss, and greater virus replication in *ifitm3*-KO mice when compared to WT mice [200, 372]. In the Everitt *et*

al study, KO mice had increased lung damage, and the virus had penetrated deeper into the lungs than in the WT mice. In the lung environment, KO mice also had higher neutrophil counts and lower NK cell and CD8+ T cell counts than WT mice. Inflammatory cytokines were also found at higher levels in the lung in KO mice.

In a 2015 Infusini *et al* study, it was observed that in WT mice, Ifitm3 expression was inducible via IAV infection [368]. In addition, high expression levels of Ifitm3 on DCs was shown to be conducive to the initiation of a timely CD8+ T cell response. This occurred despite the fact that DCs were more readily infected by IAV in *ifitm3* KO mice - this was thought to be linked to increased DC death and impaired migration associated with loss of Ifitm3.

It is important not to fully equate results from these studies to the effects of rs12252-C given that functional *IFITM3* expression is tunable via IFN. IAV replication in IAV patients would incur strong upregulation of *IFITM3* itself via the IFN response regardless of rs12252 genotype. While properly modelling these effects is not particularly feasible in the complex innate immune response network, it is worth noting that the elastic expression of *IFITM3* would be expected to introduce nuance into its effects on the immune response.

Before the start of this project, a pilot experiment was peformed by our laboratory group (Dr. Lili Wang and Dr. Shokouh Makvandi-Nejad) (fig. 6.1) using monocyte-derived DCs from healthy donors. These DCs were infected with S-Flu for 16hrs, and RNA was extracted from infected cells for RNA-Sequencing. The experiment used four CC donors and three TT donors. Monocyte-derived DCs were generated from blood monocytes, as detailed in section 2.3.4. In performing RNA-Seq analysis, we discovered a strong increase in IFN-α and IFN-β transcripts associated with rs12252-C. This was not statistically significant (*IFNB*: p = 0.194, *IFNA2*: p = 0.300)). Nevertheless, this experiment supported the idea that rs12252-C was associated with increased production of IFN, despite the limited number of donors.

We are not aware of any studies, aside from this experiment, that have given insights into the role of *IFITM3* genetics in the context of viral infection beyond clinical outcomes to infection. While the effects of the truncated IFITM3 protein have been explored with respect to viral infection using *in vitro* systems, our work in chapter 3 and the work of others has shown that these studies are not relevant to the physiological effects of the rs12252 SNP.

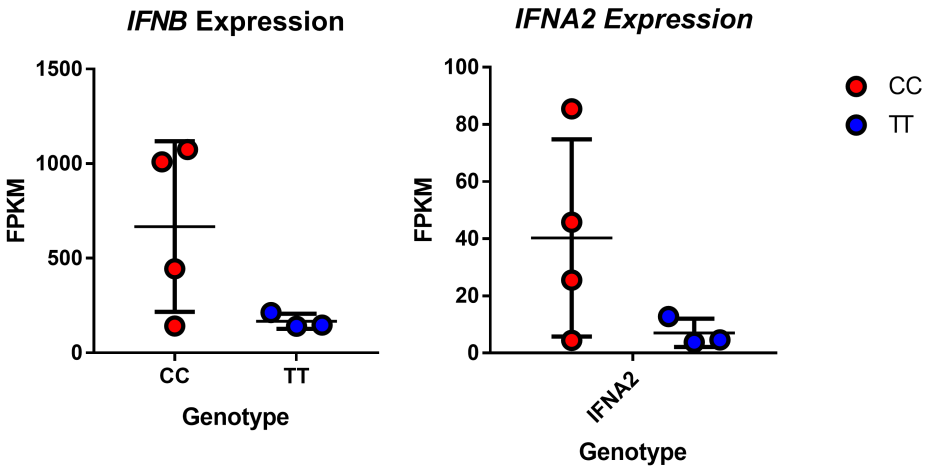


Figure 6.1: Expression of IFN transcripts in flu-infected monocyte-derived DCs upon infection with S-Flu. CC DCs showed higher expression of IFN transcripts than TT DCs on average. Theses differences were not considered statistically significant (IFNB: $p = 0.194$, IFNA2: $p = 0.300$). Data generated by Dr. Shokouh Makvandi-Nejad and Dr. Lili Wang

With this in mind, we aimed to use a combination of patient samples and *in vitro* models to attempt to understand the of rs12252 during infection.

In doing this, we benefited from patient samples obtained from Prof. Yonghong Zhang and Prof. Yan Zhao of Beijing You'An hospital, who contributed heavily to the 2013 Zhang *et al* study. Samples from this study were used in the following chapter.

6.1.1 Project Aims

The aims of this project were as follows:

1. Investigate the effects of rs12252 genotype in the context of IAV infection using transcriptomics
2. Use *in-vitro* systems to understand any observed transcriptional signatures

3. Link the predicted molecular effects of rs12252 genotype with its effects on IAV infection

6.2 Results

6.2.1 Examination of Blood Transcriptional Profile of Pandemic H1N1 Patients

PBMCs from H1N1pdm2009 patients were obtained from frozen cell samples, where RNA was extracted. Full explanation of cohort is shown in (Appendix A.1) We attempted to obtain RNA from patients with both mild and severe influenza, but were unable to obtain high quality RNA from patients with severe influenza infection. For this reason, we exclusively submitted RNA from patients with mild influenza. Six TT patients, 6 CT patients, and 5 CC patients with mild IAV infection were used. RNA samples were submitted to BGI in Beijing, China for sequencing. RNA-Sequencing was performed on a HiSeq 4000 platform. Reads were prepared using standard Illumina protocols, giving 100bp, paired end reads, with 50-60 million reads per sample. Quality Control is shown in section D.1.1

Analysis of *IFITM3* Expression

We firstly looked at differences in *IFITM3* expression levels between all patient groups. As with our large healthy donor cohort (section 4.2.5), we used two models of gene expression - a model comparing homozygous patients only, and an additive model of genotype.

A clear difference in *IFITM3* expression was seen between genotype groups, with CC patients having the highest expression and TT patients having the lowest expression (fig. 6.2). In the CC/TT model, CC patients expressed *IFITM3* at 2.58-fold higher levels than TT patients. This was statistically significant ($p = 0.00234$). Under the additive model, on average, patients expressed *IFITM3* at 2.27-fold higher levels per C allele. This was statistically significant ($p = 0.0102$). Based on this, we postulated that the C allele was associated with a stronger ISG signature in this dataset.

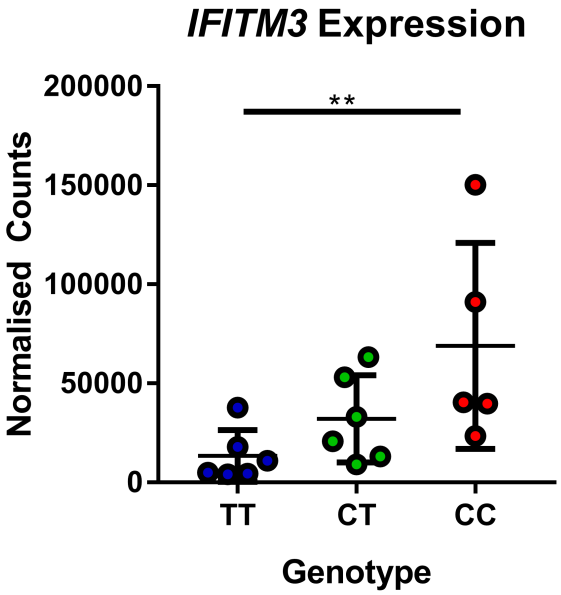


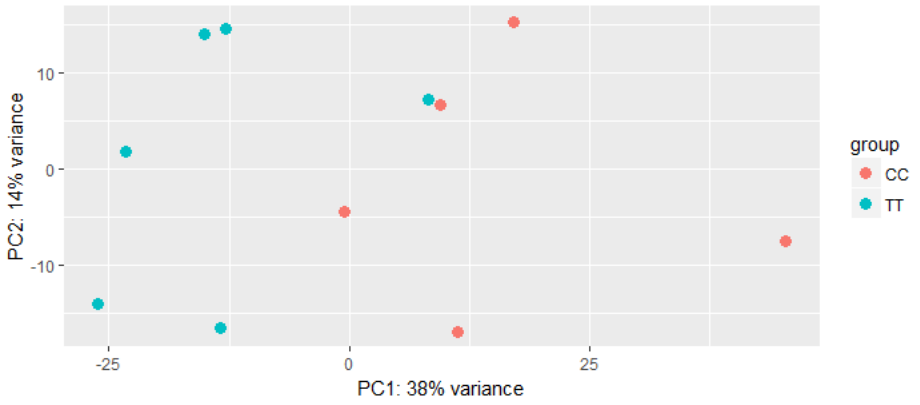
Figure 6.2: IFITM3 expression in all patient genotypes. CC patients expressed the highest levels of IFITM3 when compared to CT and TT patients. This difference was considered significant under both the CC/TT and additive model.

Differential Gene Expression Analysis of H1N1pdm2009 cohort

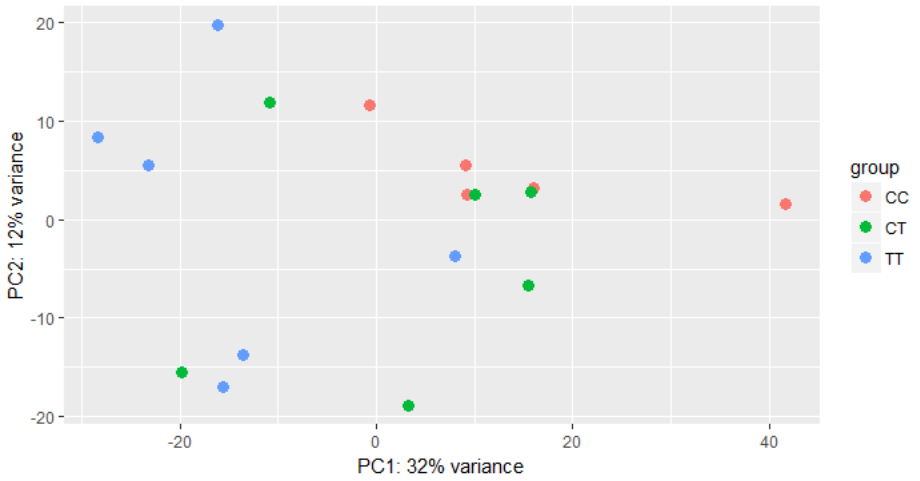
Differential gene expression analysis was performed with a CC/TT comparison and an additive model. All further p-values discussed have been adjusted for multiple comparisons across all genes via the Benjamini-Hochberg method. Initial diagnositics are shown in Appendix F.

PCA was performed twice, with all patients, and with just CC/TT patients (fig. 6.3). Obtained PCA plots showed stark clustering by genotype on the PC1 parameter, with patients roughly clustering within their own respective genotype. CC and TT patients are relatively separated, while CT patients seemingly cluster at the intersection between the two other genotype groups. These results were unexpected considering the fact that patient samples have a large amount of natural heterogeneity, and the fact that patients did not show symptomatic differences over the course of infection.

Under both statistical models, 18574 (CC/TT) and 21239 (Additive) genomic features underwent statistical testing for differential gene expression out of a possible total of 60620. Of the all genomic features tested, a total of 447



(a) PCA Plot Showing CC and TT Samples



(b) PCA Plot Showing All Samples

Figure 6.3: ***a.*** *PCA plot showing CC and TT genotypes. CC and TT patients are relatively differentiated, especially with regards to PC1, the component with the most difference attributed to it.* ***b.*** *All genotypes plotted on a PCA plot. The CT patients overlap in the centre of the TT and CC patients with regards to PC1, suggesting an intermediate gene signature.*

(CC/TT) and 374 (Additive) genes were considered differentially expressed; with an alpha of 0.05 after multiple testing adjustment. This represents differential expression of roughly 2.5% (CC/TT) and 1.8% of all genes tested.

We first examined expression levels of other *IFITM* genes, looking at *IFITM1* and *IFITM2* expression (fig. 6.4). The expression levels of these genes followed the same pattern as the one seen with *IFITM3*, where, on average, CC patients expressed the highest levels of these genes, and TT patients expressed the lowest levels. *IFITM2* expression was roughly 2.1-fold greater in CC patients compared to TT patients, and 1.54-fold increased per C allele under the additive model. These differences were statistically significant under the additive model, but not the CC/TT model (CC/TT: p = 0.0501, Additive: p = 0.0426). *IFITM1* expression was 2.61-fold greater in CC patients compared to TT patients, and 1.78-fold increased per C allele under the additive model. These differences were statistically significant (CC/TT: p = 0.00632, Additive: p = 0.0107).

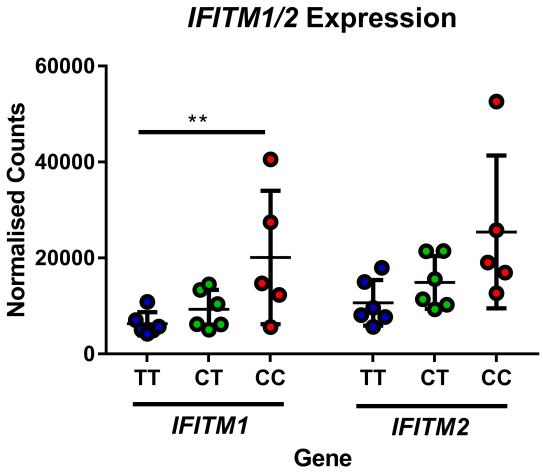


Figure 6.4: IFITM1 and IFITM2 expression patterns in H1N1pdm2009 patients. CC patients expressed the highest levels of IFITM genes in all cases.

We hypothesised that the difference in *IFITM* expression was related to a higher levels of cytokine signalling, specifically IFN, in CC patients compared to TT patients,. To investigate this, we firstly profiled the expression levels of the *STAT1*, *STAT2*, and *IRF9* TFs. As mentioned in section 1.4.2, these are the components of ISGF3 and STAT1 homodimers (fig. 6.5a).

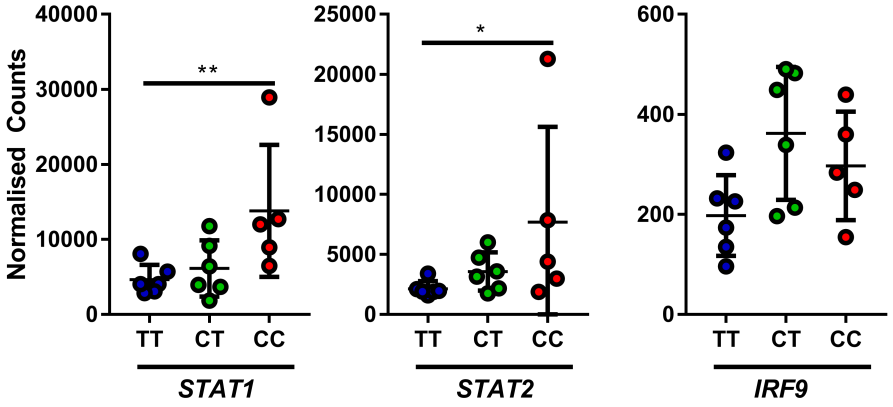
We found that both *STAT1* and *STAT2* were expressed at higher levels in CC patients compared to TT patients. *STAT1* was expressed at a 2.54-fold

higher level in CC patients compared to TT patients, and a 1.71-fold higher level per C allele under the additive model. This was statistically significant (CC/TT: p = 0.004307, Additive: p = 0.0299). *STAT2* was expressed at a 2.71-fold higher level in CC patients compared to TT patients, and a 1.89-fold higher level per C allele under the additive model. This was statistically significant (CC/TT: p = 0.0108, Additive: p = 0.0125). *IRF9* was not differentially expressed.

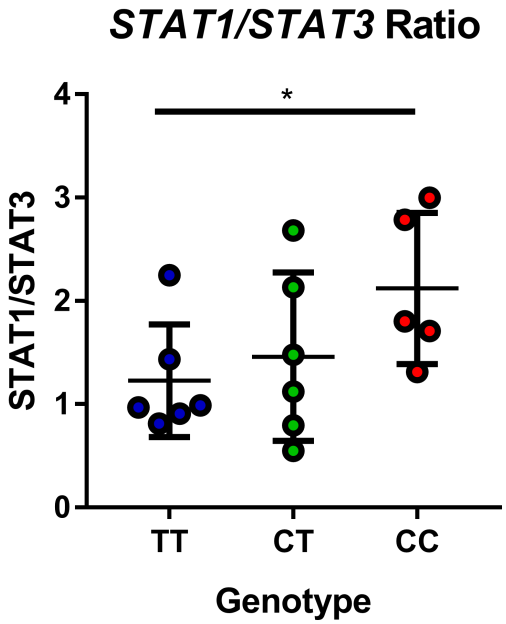
Additionally, we calculated the ratio of *STAT1* and *STAT3* expression levels (fig. 6.5b). This was based on arguments made by Ho *et al* in their study discussing the role of *STAT3* as a negative regulator of IFN responses [125]. Here, higher levels of *STAT1* expression relative to *STAT3* represents an environment that promotes the expression of interferon inducible genes. We found that CC patients had a 1.73-fold higher *STAT1* to *STAT3* ratio than TT patients. This was statistically significant via t-test (p = 0.045).

We next looked at ISG expression more systemically. We used a panel of ISGs obtained directly from the response to type-I IFN ontology (GO:0034340). 68/144 of these ISGs were tested by DESeq2, and thus were used as a panel for these analyses. We used this panel to generate a modified volcano plot using both CC/TT and additive models. In this plot, average expression of almost every ISG was higher in CC patients than TT patients under both CC/TT and additive models, with a large portion of the panel reaching statistical significance (fig. 6.6).

We also plotted a heatmap of ISG expression using the same genes (section 6.2.1). We used hierarchical clustering to sort patients in an unbiased manner based on their ISG expression profile. This clustering divided patients into three groups; the first was a single CC patient classified as an outgroup based on extremely high levels of ISG expression. The remaining patients were roughly sorted into one of two groups. The low expression group was populated by five TT patients, three CT patients and a CC patient, while the high expression group contained one TT patient, three CT patients and three CC patients. These systemic approaches to determining IFN signalling differences between the two patients groups suggest a clear trend of higher IFN signalling associated with rs12252-C



(a) *STAT1*, *STAT2* and *IRF9* Expression



(b) *STAT1*/*STAT3* Ratio

Figure 6.5: ***a.*** *Expression levels of STAT1, STAT2, and IRF9. The expression levels of the STAT genes is skewed heavily toward the CC patients, with both reaching statistical significance under both CC/TT and additive models. IRF9 was not found to be differentially expressed.* ***b.*** *Ratio of STAT1 to STAT3 expression as measured in normalised counts. Ratio is higher on average in the CC patients; this was found to be statistically significant between CC and TT patients.*

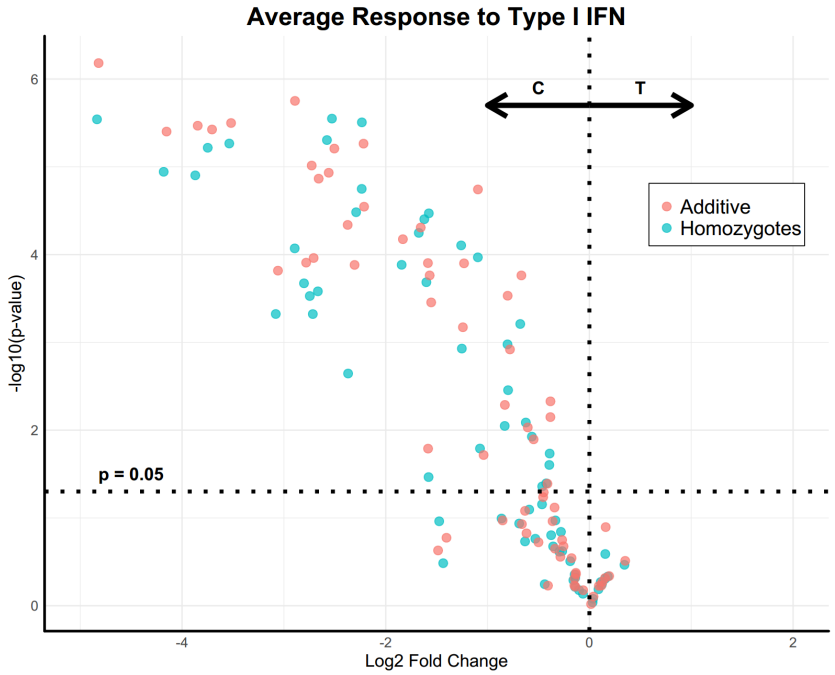


Figure 6.6: Volcano plot showing average expression of a panel of ISGs. $log_2 fold - changes$ of all genes were plotted against p-values for both the additive and CC/TT models. There is a heavy skew towards higher expression in the CC patients. In this plot, $log_2(fold - changes)$ were doubled in the additive model to obtain predicted differences in expression between homozygotes (normally this model shows $log_2(fold - change)$ per C allele instance large portion of the genes within the ISG panel were significantly differentially expressed.

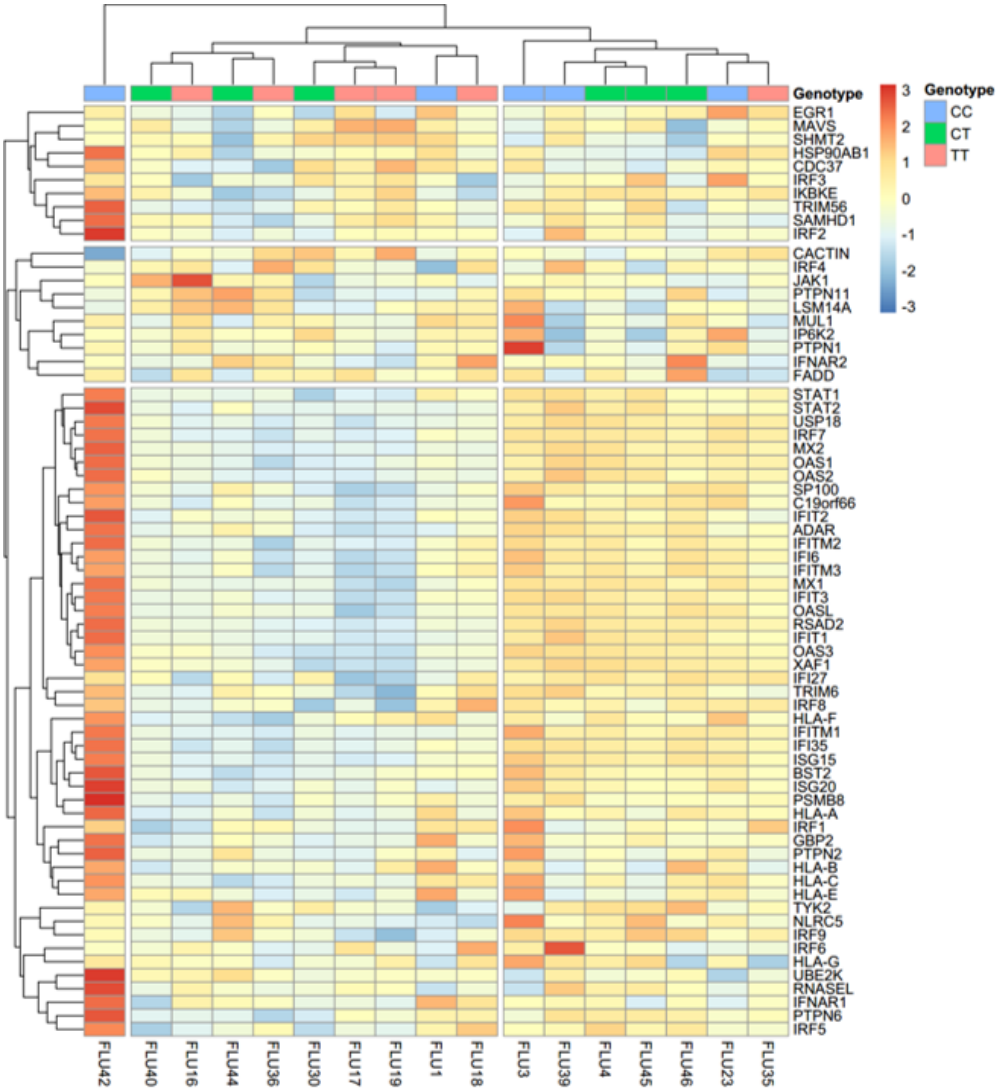


Figure 6.7: Heatmap prepared using a panel of ISGs applied to all samples. Heatmap was scaled by gene; expression levels are not comparable. Clustering tree was cut into three groups for both patients and genes. Samples were differentiated with respect to rs12252 genotype; the C allele was much more prevalent in high expression groups. One CC patient (far left) was considered an outgroup for its high ISG expression.

Gene Ontology Analysis

We subsequently used GO analysis to identify transcriptional signatures present in the aforementioned data. We used genes that were differentially expressed with a adjusted p-value less than 0.1, and a fold change greater than 2 between CC and TT patients. With these parameters, we obtained a list of 493 genes that were upregulated in CC patients and 73 in TT patients out of 19810 available genes.

Table 6.1 Top 3 Enriched GO Terms by Genotype/GO Subset

Gene Ontology	**p-value**
Biological Process (TT)	
synapse assembly	4.0×10^{-5}
positive regulation of synapse assembly	8.4×10^{-5}
nephron tubule morphogenesis	0.00018
Biological Process (CC)	
innate immune response	1.1×10^{-30}
defense response	3.4×10^{-30}
immune system process	6.8×10^{-29}
Molecular Function (TT)	
signaling receptor activity	0.016
cytokine binding	0.018
G-protein coupled receptor activity	0.019
Molecular Function (CC)	
signal transducer activity	2.4×10^{-6}
receptor activity	7.5×10^{-6}
molecular transducer activity	7.5×10^{-6}
Cellular Compartment (TT)	
intrinsic component of membrane	0.0028
integral component of membrane	0.0059
extracellular space	0.0064
Cellular Compartment (CC)	
nucleosome	1.0×10^{-11}
vesicle	2.1×10^{-11}
DNA packaging complex	6.0×10^{-11}

Upon obtaining results, TT patients terms were drastically under-enriched compared to their CC counterparts (table 6.1). The most enriched ontologies with respect to the TT patients were related to synapses. With CC patients, we found that the most significant enrichments occurred in the biological process group by a wide margin. These ontologies were related to the

innate immune response. Enriched molecular functions were related to signal transduction.

We further investigated the most significantly enriched biological process terms in CC patients. We plotted the 25 most significantly enriched biological processes (fig. 6.8) - this gave further confirmation that the enriched processes were highly related to the innate immune response and interferon signalling. With this result, it was apparent that the CC patients were again, heavily associated with higher levels of IFN signalling.

Given our observed enrichment of GO terms related to innate immunity and IFN signalling, we investigated how explanatory these GO terms were with respect to all differentially expressed genes. We clustered CC and TT patients based on the 447 significantly differentially expressed genes between CC and TT (adjusted $p < 0.05$) in a heatmap. Here, one CC patient was considered an outgroup; this was the same patient that was considered an outgroup in the heatmap produced based on the ISG panel. The remaining patients were, aside from one TT patient, segregated into genotype groups. This is not surprising given the sorting method was based on genes differentially expressed between the genotype groups. We identified four enriched gene ontologies of interest: Innate Immune Response (GO:0045087), Defense Response to Virus (GO:0051607), Response to Type-I Interferon (GO:0034340), and Response to Interferon Gamma (GO:0034341). Terms from these ontologies were superimposed on the aforementioned heatmap (fig. 6.9); these aligned with aspects of the heatmap where, on average, CC patients showed higher expression compared to the TT patients, showing that, in a large part, differential expression between TT and CC patients was concerned with innate immune signalling.

Further Investigation of IFN Signature via Interferome Database

We used the Interferome database (explained in section C.4.1) to better understand of the ISG signature present in the CC patients [373]. We cross referenced our significantly differentially expressed genes obtained in both the CC/TT (447) and additive (374) models with genes known to be IRGs in PBMCs. Under the CC/TT model, 358/447, or 80% of all differentially expressed genes were considered IRGs (fig. 6.10a). Under the additive model, 313/374, or 84% of all differentially expressed genes were considered IRGs. All IRGs flagged under the additive model were included in the CC/TT model, showing a high level of agreement between both models (fig. 6.10b).

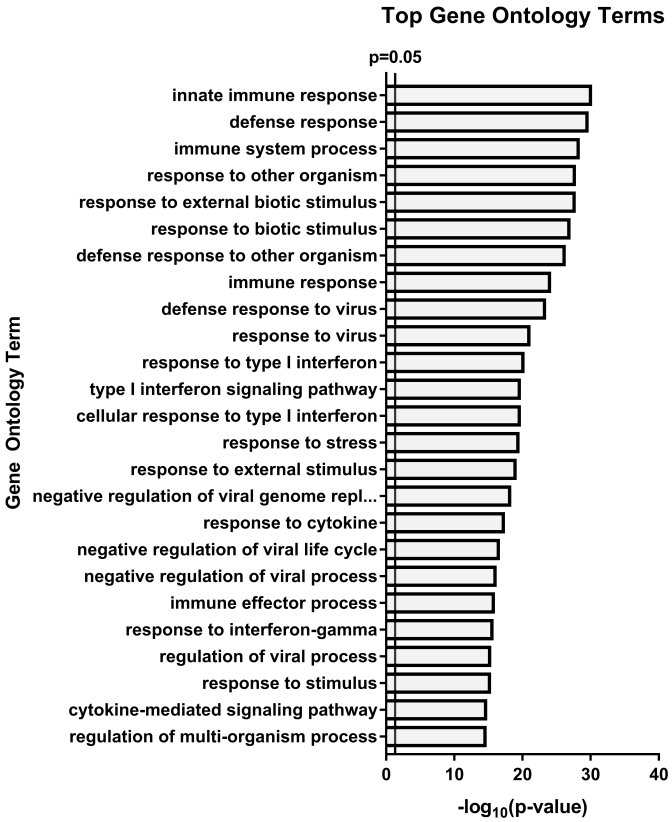


Figure 6.8: The top 20 most significantly enriched gene ontology terms from genes significantly differentially expressed between TT and CC patients. The enriched ontologies are largely related to the innate immune response to infection and its associated cytokines, especially including IFNs.

Overall, these findings indicate that IFN signalling underlies the bulk of differential gene expression between CC and TT patients.

The Interferome database allows for backtracing of a subset of IRGs, to identify which IFN type drives their expression. We used this for our differentially expressed expressed IRGs. In both models, this analysis found signalling attributable to both type I and type II IFN, suggesting higher levels of both type I and type II IFN signalling (fig. 6.11). There were no genes found that were regulated exclusively by type III IFN.

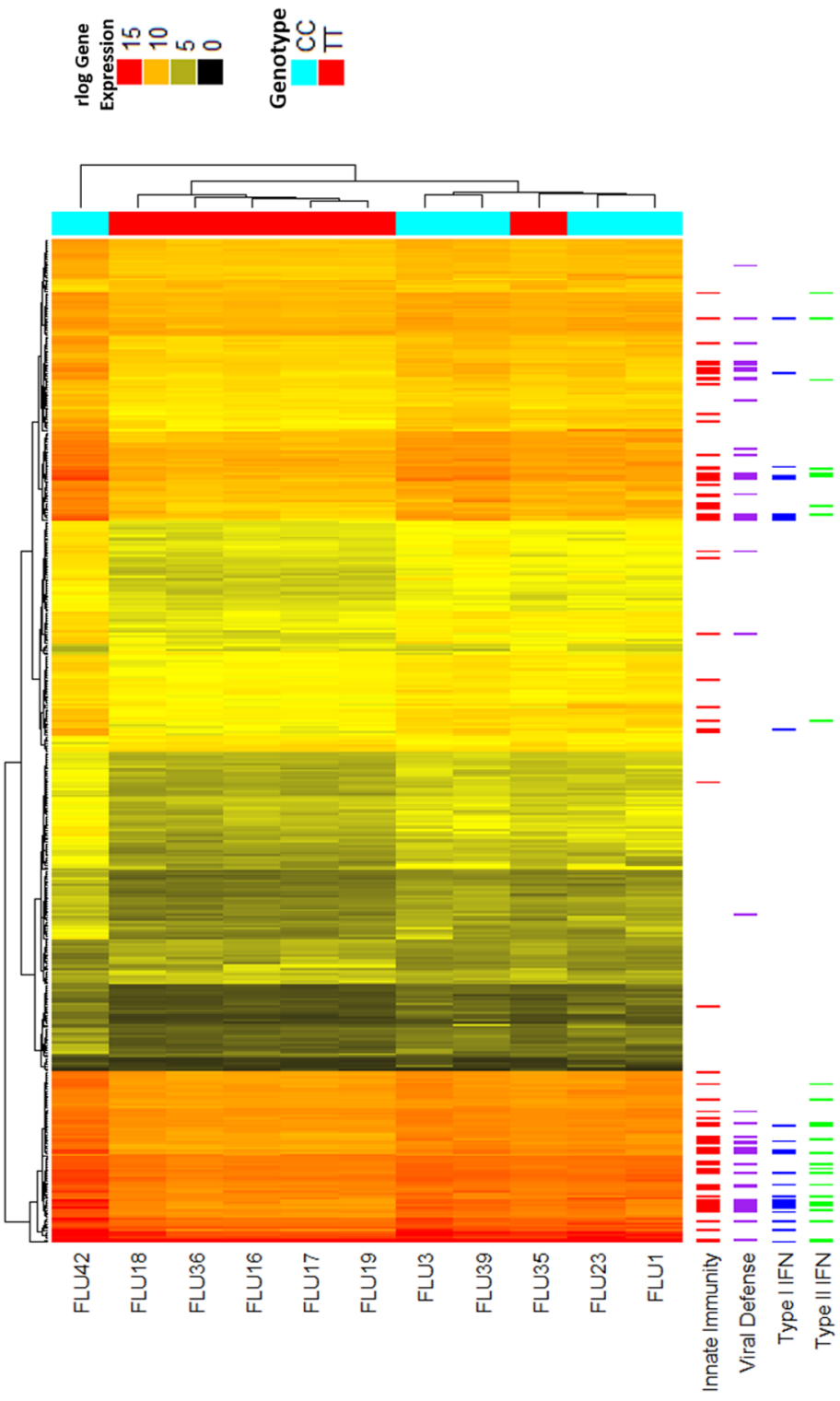
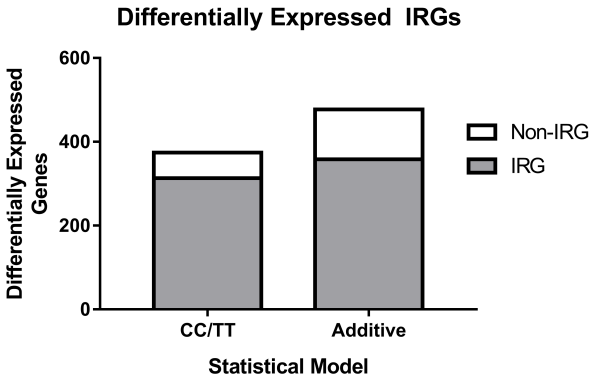
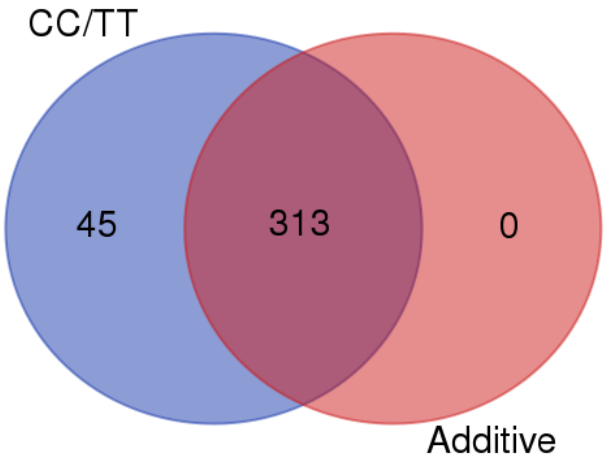


Figure 6.9: Highly significant GO terms of interest superimposed on heatmap of significantly differentially expressed genes between CC and TT patients. GO terms generally overlap with areas of high expression in CC patients compared to TT patients

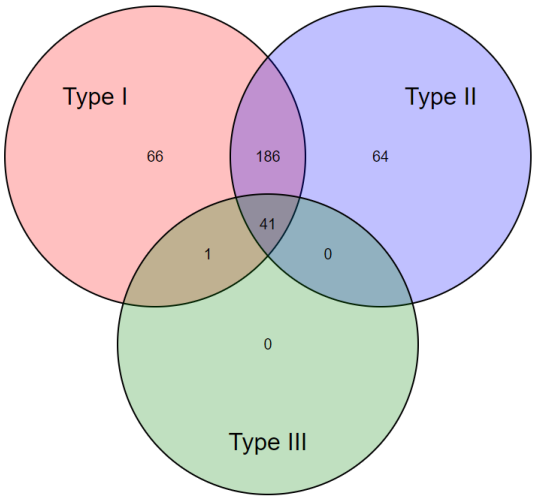


(a) DE Genes Considered IRGs

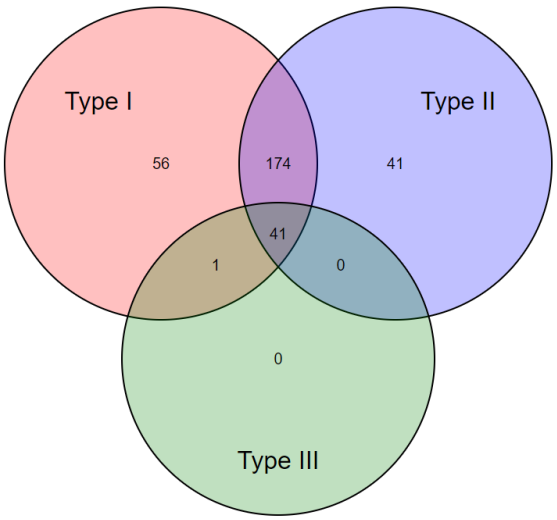


(b) Agreement Between Models

Figure 6.10: ***a.*** *80-85% of all differentially expressed genes were considered IRGs by the interferome database, depending on the statistical model used.* ***b.*** *Statistical models show a high degree of agreement in differentially expressed IRGs*



(a) CC/TT Model



(b) Additive Model

Figure 6.11: Under both statistical models, significantly differentially expressed IRGs were classified based on the IFN type associated with their transcription control. Identified genes were regulated by both types I and II IFNs, with no clear dominance of either IFN type. It was unclear if type-III IFN played a role in this scenario.

Analysis of Serum Cytokine Levels

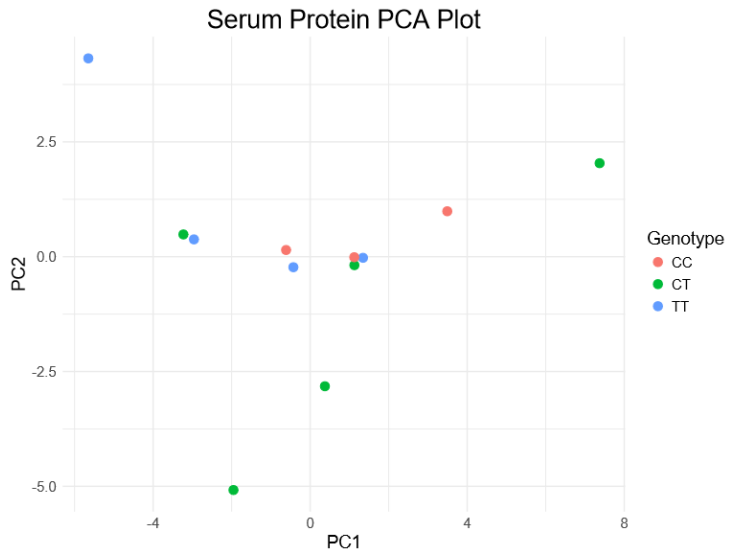
As part of the 2013 Zhang *et al* study involving these participants, serum levels of a number of cytokines were collected. In getting serum levels of cytokines for the same patients analysed above, we hoped to identify similar, and potentially novel cytokine patterns in light of our findings at the mRNA level. It is worth noting that not all patients used above had viable cytokine data; we had data from four TT patients, five CT patients, and three CC patients. Data was collected for 25 different cytokines via a Luminex bead-based multiplex assay.

To take an unbiased approach, we clustered these patients, based on cytokine profile, via PCA (fig. 6.12a). Here, we observed only slight clustering via genotype, although not to the extent that was observed at the mRNA level. Generally speaking, CC patients showed more positive PC1 values, as compared to TT patients.

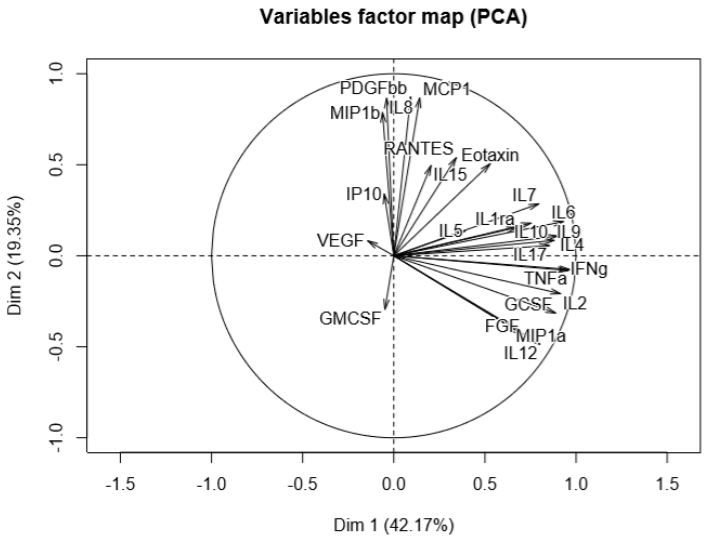
We generated a factor map of the PCA, in order to understand better the contribution of individual cytokines to the clustering of patients (fig. 6.12b). In general, this categorised cytokines into two groups - one of these was heavily related to PC1, and the other to PC2. Using these categorisations as a guide, we analysed the levels of individual cytokines. We used only the additive model in this case, based on the fact that, with the loss of several patients, not using the CT patients would result in a lack of statistical power. We did this for cytokines deemed influential for both PC1 (fig. 6.13) and PC2. In both cases, no statistically significant associations between serum cytokine levels were observed, although cytokines associated with PC1 did show a slight positive correlation between concentration and PC1 cytokines.

6.2.2 Investigation of the Cytokine Profile of Infected DCs

As mentioned in section 6.1, previous experiments conducted within our group have shown, when using monocyte-derived DCs as a model cell system, the CC genotype is associated with increased expression of type-I IFN transcripts in infected cells. We believed that these results could underlie the findings observed with in our 2009 (H1N1)pdm2009 patients, providing a link between the CC genotype and increased production of IFN.



(a) PCA of Serum Cytokine Levels



(b) Factor Map

Figure 6.12: ***a.*** *Serum levels of 25 different cytokines were used to generate a PCA of patient serum profiles. Patients are coloured according to genotype. Some clustering is observable via genotype, although this association is not strong.* ***b.*** *Factor map of the PCA plot of serum cytokine levels. The direction of each cytokine arrow indicates the direction of its influence. The magnitude of the arrow is proportional to its contribution overall. Generally speaking, cytokines associated strongly with either PC1 (Dim 1) or PC2 (Dim 2).*

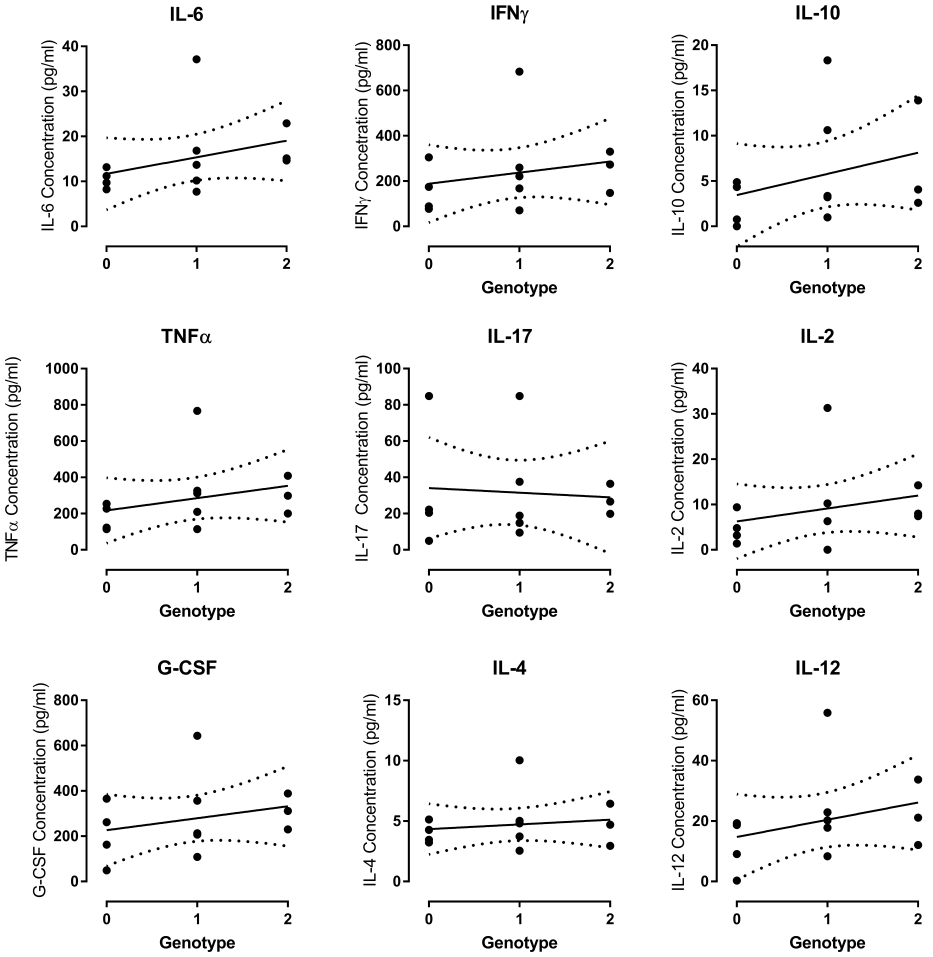


Figure 6.13: Expression of serum cytokines associated with PC1, as shown under an additive model of genotype. In this model, TT patients are shown as 0, CT patients and 1, and CC patients as 2. For these cytokines, there was a general trend of increased expression associated with rs12252-C, although this was not statistically significant.

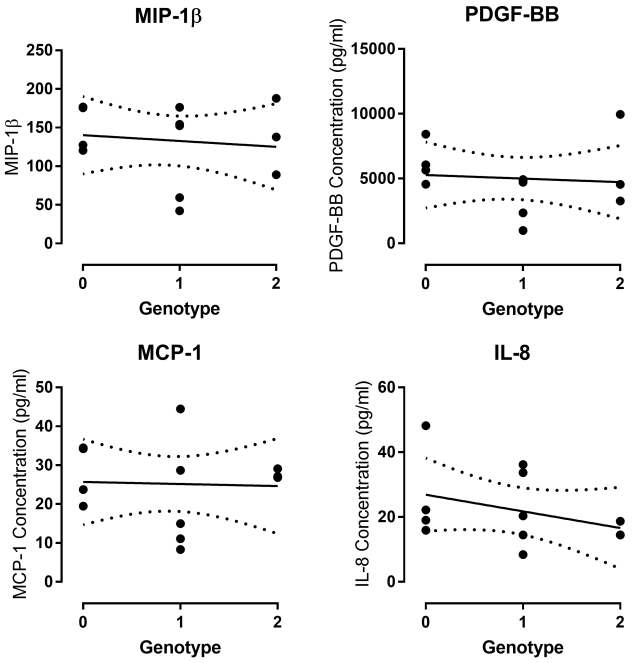
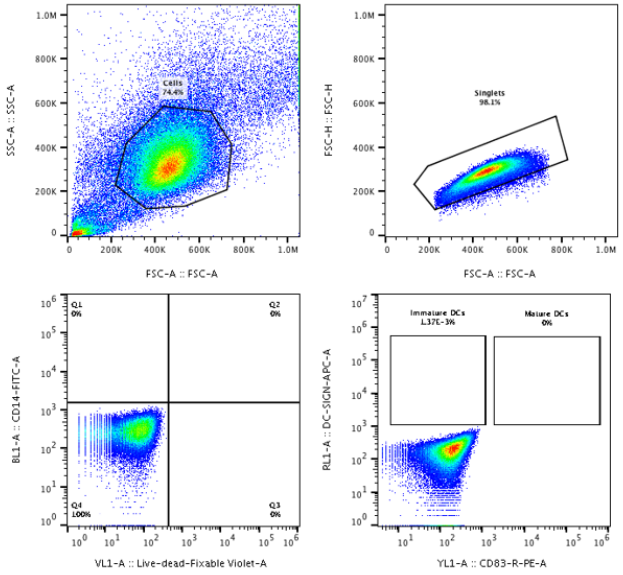


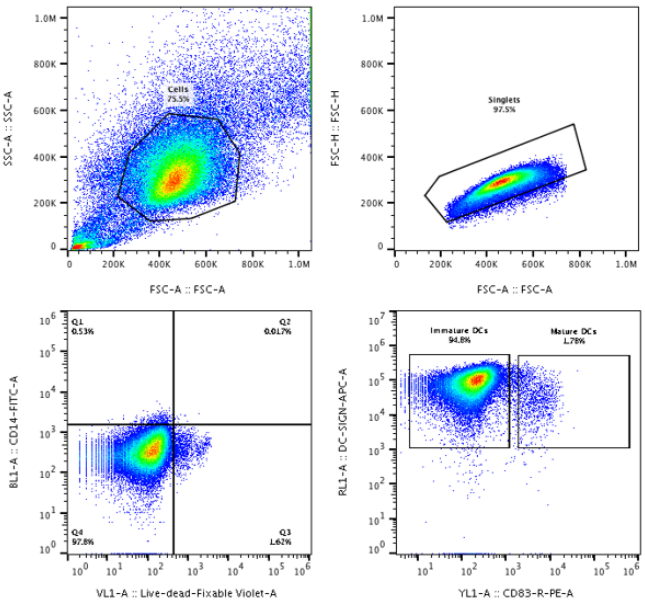
Figure 6.14: Expression of serum cytokines associated with PC2, as shown under an additive model of genotype. In this model, TT patients are shown as 0, CT patients and 1, and CC patients as 2. Here, there was no observable trend between rs12252 genotype and cytokine levels

When performing this experiment, we could not access the same donors as those used in the experiment outlined in section 6.1. We used three CC donors and 3 TT donors in this experiment. For each patient, we isolated monocytes from PBMCs, and generated monocyte-derived DCs.

We used FACS to profile the differentiation of our monocytes into DCs. We first ensured cells were living via Live/DEAD stain. We then used a combination of CD14 and DC-SIGN antibodies to differentiate between monocytes and DCs (Monocytes: CD14+, DC-SIGN-; DCs: CD14-, DC-SIGN+). After identifying our DC population, we differentiated between immature and mature DCs via a CD83 antibody (Immature: CD83-; Mature: CD83+). As shown in fig. 6.15, we obtained roughly 95% differentiation into DCs after 6 days of incubation. When using S-Flu to infect these cells, we found that an infection rate of roughly 75% was achievable with a dilution of 1/8; this was invariant from a 1/4 dilution fig. 6.16.



(a) Unstained DCs



(b) Stained DCs

Figure 6.15: Gating strategy showing full differentiation of monocyte into monocyte-derived DCs. Cell population was gated by FSC-A/SSC-A. Singlets were identified via gating cells on FSC-A/FSC-H. Live, CD14 negative cells were identified via gating singlets on Live/DEAD- and CD14- population. Immature and mature DCs were identified by gating live, CD14- singlets on DC-SIGN+ and CD83- (immature) and CD83+ populations (mature)

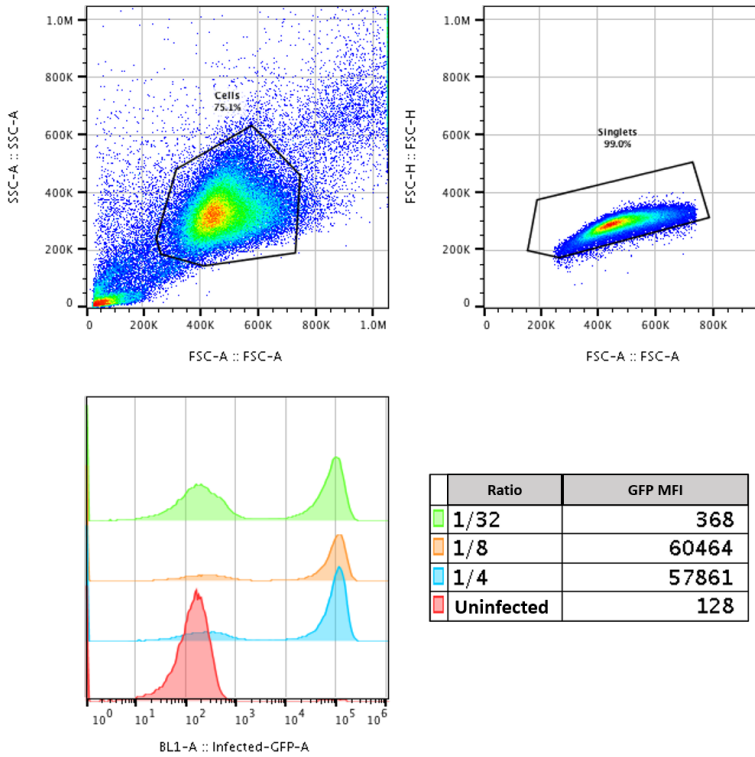


Figure 6.16: Gating strategy for infection of DCs, infected with a 1/8 dilution of S-Flu. Cell population was gated by FSC-A/SSC-A. Singlets were identified via gating cells on FSC-A/FSC-H. Cells were determined to be infected based on a stronger eGFP signal than autofluorescence of uninfected cells. MFI is shown with respect to each sample. There was no difference in infection rate between DCs infected with a 1/4 and 1/8 dilution ratio

When infecting cells from all donors, we observed an mean infection rate of 73% across all donors after 8 hours (fig. 6.17). We saw a marginally higher infection rate in TT donors when compared to CC donors. This difference was not statistically significant. We used cell supernatant to measure levels of soluble IP10 secreted by our infected DCs in response to infection; this protein is IFN inducible. We observed a marginal increase in IP-10 secretion associated with the CC genotype (fig. 6.18a). This was not statistically significant. To complement this, we performed an IFN bioassay, using a cell reporter system outlined in section 2.3.9. Here, we observed marginally higher rates of type-I IFN secretion associated with the CC genotype (fig. 6.18b), although this was not statistically significant.

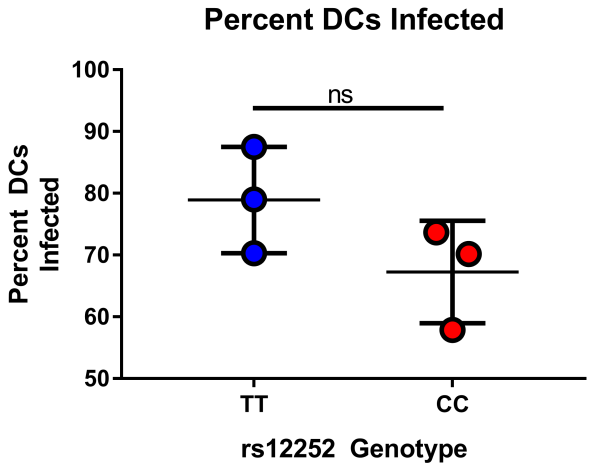


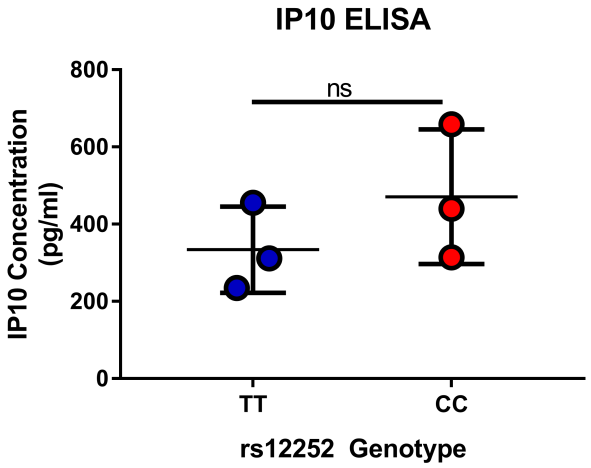
Figure 6.17: Infection of monocyte-derived DCs with S-Flu virus. DCs were homozygous for the T or C allele of rs12252. TT cells showed higher rates of infection, although this difference was not statistically significant.

6.2.3 Infection of CRISPR-Edited HEK293 Cells

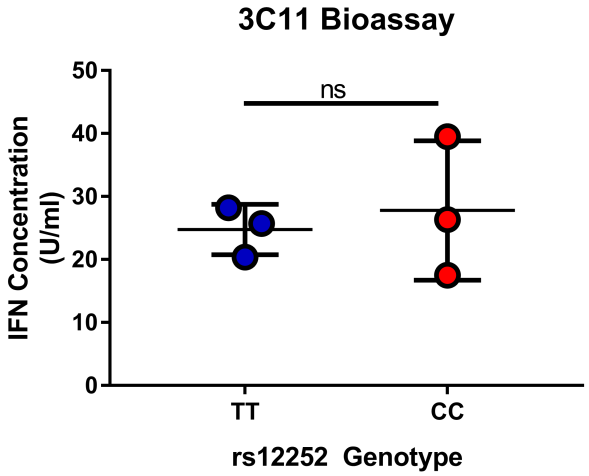
As a second means to investigate differential infection sensitivities with respect to rs12252, we infected our edited HEK293 cell lines with S-Flu. As shown in section 4.2.4, these cells had reduced expression of *IFITM3* associated with rs12252-C. We hypothesised that the CC genotype, with its reduction in *IFITM3* expression would result in an increased sensitivity to IAV infection, and a greater IFN response. When performing these infections, we infected KO cell lines in parallel.

HEK293 Infection Sensitivity With Respect to rs12252 Genotype

We infected cells for both 24 and 16 hours using a range of viral dilutions. Infected cells were determined via FACS, as shown in fig. 6.19. Here, the *IFITM3*-KO cell lines showed an extremely large increase in susceptibility to IAV infection when compared to the WT and KI cells. When using 1/8 and 1/32 dilutions, *IFITM3*-KO cell lines effectively achieved maximum infection, while the other cell lines did not. KI cells showed a modest increase in susceptibility when compared to WT cells. After infection for 16 hours, CC HEK293 cells were infected at 1.14 to 1.53 fold higher rates than WT cells, depending on the concentration of virus (fig. 6.20a). After



(a) Supernatant IP-10 Concentration



(b) Supernatant Type-I IFN Concentration

Figure 6.18: ***a.*** *Supernatant concentration of IP-10, as measured by ELISA. There was no statistically significant difference in production of IP-10 between TT and CC monocyted-derived DCs.* ***b.*** *Measurement of supernatant concentration of type-I IFN, as measured by 3C11 bioassay. There was no statistically significant difference in production of type-I IFN between TT and CC monocyte-derived DCs*

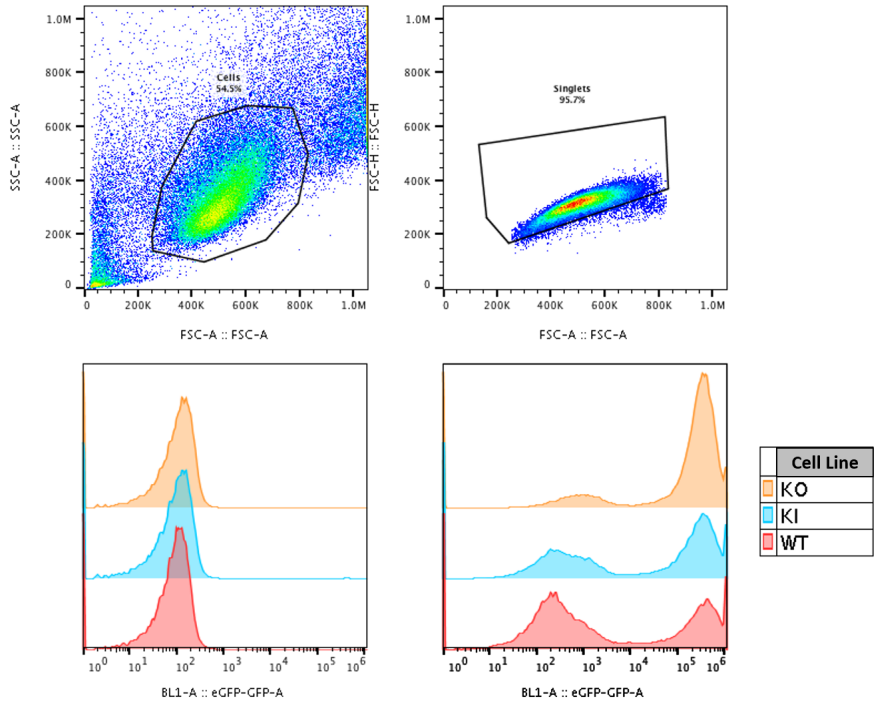


Figure 6.19: Gating strategy for infection of HEK293 cells with S-Flu virus. Cell population was gated by FSC-A/SSC-A. Singlets were identified via gating cells on FSC-A/FSC-H. Cells were determined to be infected based on a stronger eGFP signal than autofluorescence of uninfected cells

24 hours, CC HEK293 cells were infected 1.23 to 1.37 fold higher rates than WT cells (fig. 6.20b). These results indicate higher susceptibility to infection associated with the C allele of rs12252, likely as a result of the reduced *IFITM3* expression.

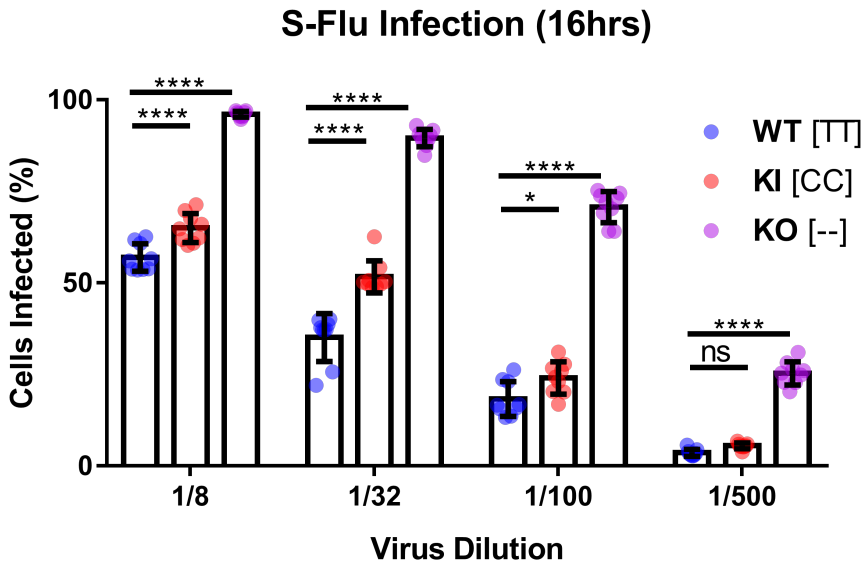
6.3 Discussion

6.3.1 RNA-Sequencing of Patients Reveals rs12252-C Associated IFN Signature

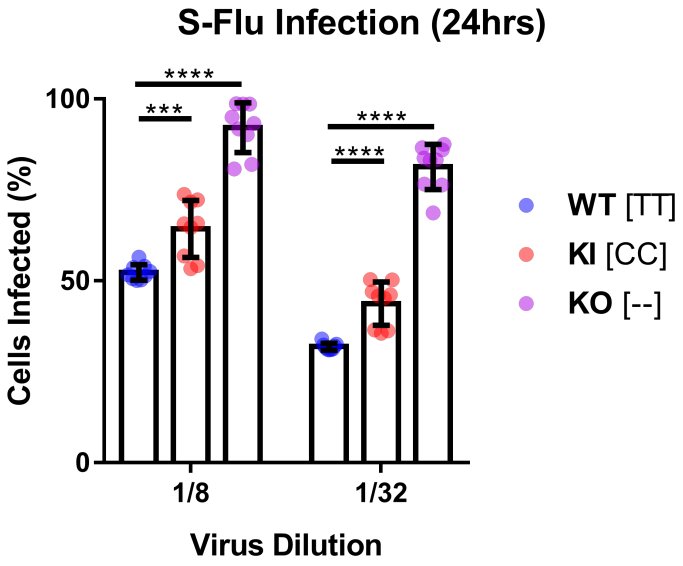
Before this study, there was little understanding of how the rs12252 affected overall IAV infection. Using a series of exploratory experiments, we have uncovered some evidence that the rs12252 genotype may affect the overall molecular signature of the immune response to influenza infection. This study however, is still limited in terms of scope, and in general would require more evidence to support its findings. After observing a heightened innate immune signature in H1N1pdm2009 patients with the CC genotype, when compared to other genotypes, we have found some evidence for differential infection rates in CRISPR-Cas9 edited HEK293 cells. We could not reproduce these findings in DCs.

In our pateint dataset, we uncovered a stark innate immune signature associated with rs12252 genotype associated with interferon stimulation. However, it is worth noting the limitations of this study, and evaluating its conclusions for future perspective. Firstly, the small patient numbers should be acknowledged - seventeen patients total were used in this case due to the difficulty in obtaining influenza samples that are well controlled, especially those with mild disease. This leaves a large amount of room for mitigating factors such as stage of infection, and age to influence the transcriptome of the patients. In addition, our findings from chapter 4 suggest that the expression of *IFITM3* in patients would underlie the observed differential molecular signature. It is questionable whether or not these expression level effects would be appreciable in a cohort of this size. Any relationship between *IFITM3* basal expression and signature within infection is unknowable without healthy samples from the same patients.

The tissue we used in this study was PBMCs - these cells were not directly in contact with the site of infection, the respiratory tract. Despite this, we could



(a) 16hrs Infection



(b) 24hrs Infection

*Figure 6.20: Infection of HEK293 cell lines with S-Flu for 16 (**a**) and 24 (**b**) hours. S-Flu dilutions were from stock solution, diluted with DPBS. KI HEK293 cells were found to be more sensitive to S-Flu infection by a modest amount when compared to WT cells. In contrast, KO cell lines were highly sensitive to S-Flu infection when compared to WT cell lines. Data was analysed with a one-way ANOVA for each dilution factor, with a post-hoc Dunnett-test, with correction for multiple comparisons.*

identify clear molecular signatures associated with this infection. Previous studies have reported clear whole blood molecular signatures associated with influenza infection when compared to molecular controls, supporting our findings [374].

With respect to this study, it is unclear what the significance of the increased ISG signature associated with rs12252-C suggests about *IFITM3* mediated restriction. The best frame of reference for this finding is the results of the MOSAIC study, where the whole blood molecular signatures of 131 patients hospitalised with IAV were stratified by severity [375]. This study used healthy controls as a comparison. In this study, it was observed that the ISG signature, and GO terms specifically associated with viral response were upregulated in a time dependent manner, and were enriched in patients with less severe outcomes to IAV. In contrast, GO terms related to bacterial infection were associated with severe outcomes. While this study used whole blood instead of PBMCs, there will be significant overlap in these transcriptome signatures.

Under the MOSAIC study model, our CC patients would be clearly associated with a mild influenza outcome. Given that we did not report any bacterial terms in our TT patients, the question worth considering is what the significance of stronger viral signatures within the same patient group is. Given the fact that antiviral and inflammatory responses are linked via PRR signalling, it is possible that the CC patients had higher PRR signalling than TT patients on average, but not enough to tip the balance towards severe infection. While it is not entirely reliable to use reductionist studies in this context, this hypothesis would be supported by a 2014 study by Levin *et al*, where antiviral IFNs were seen to be be expressed to saturated levels by low concentrations of type-I IFN, while inflammatory/non-proliferative genes required higher concentrations of type-II IFN for expression [376].

When looking at this data, it is important to recognise that the genes encoding all IFN subtypes were not expressed within the cells profiled. Results from this study and the MOSAIC study suggest that the ISG signature is driven by the release of IFN at the site of infection [375]. Given that in humans, IAV titres generally correlate with local IFN, if the SNP was not affecting IFN levels independently, it would seem likely that there was a greater degree of virus replication associated with rs12252-C [377].

In animal models, there is strong support for excessive IFN signalling incurring severe influenza [148]. It is thought that type-I IFN eventually leads

to excessive proinflammatory cytokine signalling, which ultimately mediates the progression from mild to severe infection. While this may be a plausible link between the differential IFN signatures observed in relation to rs12252 genotype, this would require further investigation to elucidate.

Using the interferome database, we attempted to deconvolve the ISG signature observed in patients, but could not associate it with type-I or type-II IFN specifically. Cross expression of both type-I and type-II IFN have been shown to peak at roughly day 7 post infection - this is consistent with the fact that the patients had been experiencing symptoms of infection for several days at the time of sample collection [153]. It is unsurprising that no type-III IFN specific genes were detected; these generally are less well characterised, and do not affect PBMCs [378].

It is interesting to consider what the implications of higher IFN signalling may have on adaptive immunity. As mentioned briefly in section 6.1, in a 2015 study, Infusini *et al* described an association between high IFITM3 expression in respiratory DCs and improved trafficking to lymph nodes for antigen presentation [368]. It has also been shown that IFNs positively enhance B and T cell function in the context of viral infection, driving the production of neutralising antibodies in the former [379]. Given this, it may be possible that the increased IFN signature associated with rs12252-C may help to induce adaptive immunity in some cases.

Despite the strong ISG signature observed when stratifying patients by rs12252 genotype, serum cytokine levels did not associate strongly with genotype. This is partially explainable by the fact that a reduced number of patients had serum cytokine data available. The serum obtained would also be influenced by all cells within the whole blood and tissues, instead of just the PMBCs which we analysed. The cytokine panel did not include type-I IFN, only type-II, which did show a slight correlation with rs12252 genotype. It is possible that stronger IFN signalling would not be as visible with these serum cytokines. In the MOSAIC study, serum cytokines did not vary nearly as much as mRNA levels according to disease severity, suggesting that mRNA may be the most effective way to observe differential IFN responses to infection [375]. In a 2013 Davey *et al* study, twelve different serum biomarkers were found to be associated with progression to severe influenza - including IL-6, IL-10, Interleukin-2 (IL-2), Monocyte Chemoattractant Protein 1 (MCP1), IP10, and TNFα [380]. The serum levels of four of these were somewhat positively correlated with rs12252-C, supporting previous hypotheses.

6.3.2 *In Vitro* Models of Infection Somewhat Recapitulate rs122252-C Associated Phenotypes

When infecting HEK293 cell lines with S-Flu, we saw high levels of infection in KO cells, and low levels in WT cells. This was consistent with our hypothesis that rs12252-C was associated with higher levels of infection, and that this susceptibility would be not be as strong as that in the KO cells. This was likely due to the lower expression of IFITM3 associated with the KI cells. The KO cells were significantly more susceptible to infection than KI cells this difference in susceptibility was only appreciable at lower dilutions. This suggests that low levels of IFITM3 expression are vastly more protective than a lack of the protein.

Overall, our *in vitro* infections of HEK293 cells lends a good deal of evidence our hypothesis that rs12252-C is associated with higher levels of viral replication and an increased IFN response. As higher infection rates would generally be expected to incur higher levels of viral replication, which would ultimately trigger higher levels of PRR signalling, even independent of RIG-I.

Our findings with regards to DC infection did not support the our earlier experiments. It should be noted that both experiments suffer from a lack of statistical power, as both were limited by donor numbers. In our experiment, IFN induction was slightly higher in CC cells despite the fact that they incurred reduced expression compared to the TT cells, which somewhat supports the original experiment. However, it seems clear that this model is not particularly effective without an expanded donor pool. Moving forward, the use of CRISPR-Cas9 to generate different cell types would likely be the best method to investigate the effects of rs12252-C on different cell types.

While this is not true for monocyte-derived DCs, in pDCs IFN signalling occurs primarily via TLR7, not RIG-I. It seems likely that with a single-pass virus, IFITM3 would have no effect on PRR signalling occurring before the entry stage of infection. However, a 2017 report by Stacey *et al* did observe, in the context of Murine Cytomegalovirus infection, TLR3 and TLR9 were more robustly stimulated in *ifitm3* - KO mouse DCs. While this was not observed in the context of TLR7, this may hint at an alternative mechanism by which excessive PRR signalling could be incurred with the reduction IFITM3 expression. While DCs have fully functional RIG-I signalling, it is unclear whether or not this contributed at all to this IFN response due to the actions of NS1 [381].

6.3.3 Concluding Remarks

In this chapter, we have, using a range of *in vivo* and *in vitro* systems, investigated the molecular effects of rs12252-C on natural influenza infection. While the results from this study were mostly exploratory, these findings give insight into the effects of rs12252-C in infection. Overall, it appears likely that this SNP is associated with an increase in IFN signalling during the course of infection, likely due to differential levels of viral replication and PRR signalling. However, given our findings in chapter 4, it is unclear whether or not *IFITM3* expression is reduced in some cells, still leaving some clear gaps in our knowledge.

Moving forward, it is clear that the full effects of rs12252-C on individual cell populations should be elucidated, as expressed in chapter 4. In the meantime, a greater understanding of the effects of rs12252-C on IAV infection could be elucidated via the use of larger patient cohorts, with greater symptomatic characterisation. In addition, genome editing should continue to be used to further generate *in vitro* models for IAV infection that would be more relevant and flexible than HEK293 cells, like induced pleuripotent stem cells, DCs, and A549 cells.

Chapter 7

Final Discussion and Future Directions

Over the course of this project, we have made significant contributions to understanding of rs12252 in and outside the context of IAV infection. Previously, there was relatively little understood about this SNP and its association with viral infection. Since then, significant advances have been made by our group and others to understand the role of rs12252 and the genetics of *IFITM3* in the context of viral infection.

Prior to this study, rs12252-C was thought to affect the splicing profile of *IFITM3* mRNA. However, we have now shown, alongside others, that this SNP does not affect the *IFITM3* splicing profile. This was a significant contribution to the field; since the Everitt *et al* paper, many studies were built off the idea that this splicing occurred. In disproving this hypothesis, we hope that other researchers will be influenced to focus on the real effects of this SNP. Given this, there is not a great deal of further investigation that is needed in the context of this project.

We have established for the first time, that rs12252 likely modulates expression of *IFITM3*. We have accumulated significant evidence for this using *in vitro* models and human donors. This conclusion is most strongly supported by our data from healthy human donors, which showed a robust association between rs12252-C and high *IFITM3* expression. We have also obtained more preliminary data from primary skin fibroblasts, and a CRISPR-Cas9 edited cell line which suggests that in some cases, rs12252-c is associated with reduced expression of *IFITM3*. The idea that rs12252 is a

bidirectional, and potentially tissue-specific eQTL fits well within the context of the literature. To understand this SNP's bidirectional nature, establishing which cells are associated with which directionality would also be a valuable future endeavour.

We also found broader patterns of transcriptional activity associated with this SNP in our healthy donor cohort relating to ISG and ncRNA expression. A relationship between rs12252 and ISG signature could have profound implications for the role of this SNP in viral disease. Differential expression of ncRNAs close to *IFITM3* could suggest a relationship between rs12252 and transcriptional patterns within the region. *In vitro* systems, like our HEK293 cells, could be used for epigenomic studies to further investigate these ncRNAs. Moving forward, we feel as if these observations merit further investigation in new cohorts, and could be of great importance in the context of rs12252.

To our knowledge, we have also established the first instance where the *IFITM3* locus has been edited using CRISPR-Cas9 to introduce novel mutations. We hope to use this in new cell systems, which would be most effective in editing of induced pluripotent stem cells - something that has already been established for several years [382]. It is highly possible that we could use established editing systems to introduce rs12252-C, or other SNPs of interest into the genome of these cells lines and using this system, we could generate a variety of cell types to examine effects of this or other SNPs. This would serve as the best possible system that for investigating the molecular effects and potentially bidirectional nature of this SNP.

As mentioned previously, there has been a great deal of consensus in recent literature that SNPs which modulate *IFITM3* can strongly associate with severity of disease. Particularly, these polymorphisms have all been associated with gene expression, with lower expression correlating with severe disease. This is largely in agreement with our findings regarding the effects of rs12252-C, although effects are more nuanced than other SNPs given its potentially bidirectional nature. If basal *IFITM3* expression is truly a indicator for IAV susceptibility, it may be best to take more holistic approaches to this in the future, where combinatorial risk scores could be calculated based on the presence of SNPs that have been shown to be eQTLs for *IFITM3*. Using our EAS specific eQTL database, we took an approach similar to that outlined above. We identified two novel eQTLs for *IFITM3*, showing that they could work in combination to influence *IFITM3* expression. Genotyping for these SNPs in readily available viral infection

cohorts will hopefully allow us to identify new potential risk variants using this new approach. In the event that we do discover novel risk variants, we could carry out new mechanistic studies to study the effects of the rs28475163 SNP; it is possible this SNP affects TF binding at a defined regulatory site.

It is also worth mentioning the value of having an eQTL database to the scientific community and important to recognise that majority of populations are vastly under-represented in eQTL databases. It is our hope this can be used in an open ended manner moving forward to serve the scientific community at large.

Finally, we have uncovered strong associations between rs12252-C and IFN signatures in mild influenza patients, supporting a working model where low IFITM3 expression at baseline incurs high levels of viral replication and PRR signalling. While this model requires further probing, the stark difference between these patient's ISG signatures is a strong starting point in ascertaining the effects of rs12252 in infection. Moving forward, we would like to follow these studies up, matching patients to healthy samples with more robust data, such as viral titres. At present however, we are moving forward to publish these findings, alongside the findings we obtained with our healthy donor cohort.

To conclude, this thesis has contributed significantly to our understanding of the rs12252 SNP, and host genetics of *IFITM3* in general. Understanding this SNP, and the genetics of susceptibility to viral infections in general, will be an effective future method of identifying at-risk individuals, and better understanding epidemiology of influenza and all viral infections. Understanding of IAV susceptibility could be particularly effective when considering vaccination strategy and high pathology avian influenza strains. It is our hope that in the future, this understanding of the interplay between host genetics and viral infection will help to curtail severe disease.

Approximate Word Count: 50,000

References

1. Shi, M. *et al.* The evolutionary history of vertebrate RNA viruses. *Nature* **556,** 197–202. ISSN: 0028-0836 (2018).
2. Li, C.-X. *et al.* Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife* **4.** ISSN: 2050-084X. doi:10.7554/eLife.05378 (2015).
3. Fiore, A. E. *et al.* Prevention and control of influenza: recommendations of the Advisory Committee on Immunization Practices (ACIP), 2008. *MMWR Recommendations & Reports* **57,** 1–60. ISSN: 1057-5987 (2008).
4. Thompson, W. W. *et al.* Estimates of US influenza-associated deaths made using four different methods. *Influenza and Other Respiratory Viruses* **3,** 37–49. ISSN: 17502640 (2009).
5. Ma, H.-Y. *et al.* Risk factors associated with severe influenza virus infections in hospitalized children during the 2013 to 2014 season. *Journal of Microbiology, Immunology and Infection* **49,** 387–393. ISSN: 1684-1182 (2016).
6. Wong, S.-S. *et al.* Severe Influenza Is Characterized by Prolonged Immune Activation: Results From the SHIVERS Cohort Study. *The Journal of Infectious Diseases* **217,** 245–256. ISSN: 0022-1899 (2018).
7. Fabian, P. *et al.* Influenza Virus in Human Exhaled Breath: An Observational Study. *PLoS ONE* **3,** e2691. ISSN: 1932-6203 (2008).
8. Mubareka, S. *et al.* Transmission of influenza virus via aerosols and fomites in the guinea pig model. *The Journal of infectious diseases* **199,** 858–65. ISSN: 0022-1899 (2009).
9. Taubenberger, J. K. & Morens, D. M. The pathology of influenza virus infections. *Annual review of pathology* **3,** 499–522. ISSN: 1553-4006 (2008).

10. Dadonaite, B., Vijayakrishnan, S., Fodor, E., Bhella, D. & Hutchinson, E. C. Filamentous influenza viruses. *Journal of General Virology* **97,** 1755–1764. ISSN: 0022-1317 (2016).

11. Bouvier, N. M. & Palese, P. The biology of influenza viruses. *Vaccine* **26,** D49–53. ISSN: 0264-410X (2008).

12. Mahy, B. W. J. & Van Regenmortel, M. H. V. *Encyclopedia of Virology Encyclopedia of Virology* **2.** ISBN: 9780123739377 (Academic Press, 2008).

13. Sha, B. & Luo, M. Structure of a bifunctional membrane-RNA binding protein, influenza virus matrix protein M1. *Nature Structural Biology* **4,** 239–244. ISSN: 1072-8368 (1997).

14. Portela, A. & Digard, P. The influenza virus nucleoprotein: a multifunctional RNA-binding protein pivotal to virus replication. *Journal of General Virology* **83,** 723–734. ISSN: 0022-1317 (2002).

15. Goto, H., Muramoto, Y., Noda, T. & Kawaoka, Y. The genome-packaging signal of the influenza A virus genome comprises a genome incorporation signal and a genome-bundling signal. *Journal of virology* **87,** 11316–22. ISSN: 1098-5514 (2013).

16. Trifonov, V., Racaniello, V. & Rabadan, R. The Contribution of the PB1-F2 Protein to the Fitness of Influenza A Viruses and its Recent Evolution in the 2009 Influenza A (H1N1) Pandemic Virus. *PLoS currents* **1,** RRN1006. ISSN: 2157-3999 (2009).

17. Kido, H. Influenza virus pathogenicity regulated by host cellular proteases, cytokines and metabolites, and its therapeutic options. *Proceedings of the Japan Academy. Series B, Physical and biological sciences* **91,** 351–68. ISSN: 1349-2896 (2015).

18. Matrosovich, M. N., Matrosovich, T. Y., Gray, T., Roberts, N. A. & Klenk, H.-D. Human and avian influenza viruses target different cell types in cultures of human airway epithelium. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 4620–4. ISSN: 0027-8424 (2004).

19. Kumlin, U., Olofsson, S., Dimock, K. & Arnberg, N. Sialic acid tissue distribution and influenza virus tropism. *Influenza and Other Respiratory Viruses* **2,** 147–154. ISSN: 17502640 (2008).

20. Rust, M. J., Lakadamyali, M., Zhang, F. & Zhuang, X. Assembly of endocytic machinery around individual influenza viruses during viral entry. *Nature Structural & Molecular Biology* **11,** 567–573. ISSN: 1545-9993 (2004).

21. De Vries, E. *et al.* Dissection of the Influenza A Virus Endocytic Routes Reveals Macropinocytosis as an Alternative Entry Pathway. *PLoS Pathogens* **7,** e1001329. ISSN: 1553-7374 (2011).

22. Edinger, T. O., Pohl, M. O. & Stertz, S. Entry of influenza A virus: host factors and antiviral targets. *Journal of General Virology* **95,** 263–277. ISSN: 0022-1317 (2014).

23. Gillespie, L. *et al.* Endocytic function is critical for influenza A virus infection via DC-SIGN and L-SIGN. *Scientific Reports* **6,** 19428. ISSN: 2045-2322 (2016).

24. Lakadamyali, M., Rust, M. J., Babcock, H. P. & Zhuang, X. Visualizing infection of individual influenza viruses. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 9280–5. ISSN: 0027-8424 (2003).

25. Melikyan, G. B., Niles, W. D. & Cohen, F. S. Influenza virus hemagglutinin-induced cell-planar bilayer fusion: quantitative dissection of fusion pore kinetics into stages. *The Journal of general physiology* **102,** 1151–70. ISSN: 0022-1295 (1993).

26. Safo, M. K. *et al.* Crystal Structures of Influenza A Virus Matrix Protein M1: Variations on a Theme. *PLoS ONE* **9** (ed Menéndez-Arias, L.) e109510. ISSN: 1932-6203 (2014).

27. Cros, J. F. & Palese, P. Trafficking of viral genomic RNA into and out of the nucleus: influenza, Thogoto and Borna disease viruses. *Virus research* **95,** 3–12. ISSN: 0168-1702 (2003).

28. O'Neill, R. E., Jaskunas, R, Blobel, G, Palese, P & Moroianu, J. Nuclear import of influenza virus RNA can be mediated by viral nucleoprotein and transport factors required for protein import. *The Journal of biological chemistry* **270,** 22701–4. ISSN: 0021-9258 (1995).

29. Vreede, F. T., Jung, T. E. & Brownlee, G. G. Model Suggesting that Replication of Influenza Virus Is Regulated by Stabilization of Replicative Intermediates. *Journal of Virology* **78,** 9568–9572. ISSN: 0022-538X (2004).

30. Krug, R. M. Priming of influenza viral RNA transcription by capped heterologous RNAs. *Current topics in microbiology and immunology* **93,** 125–49. ISSN: 0070-217X (1981).

31. O'Neill, R. E., Talon, J & Palese, P. The influenza virus NEP (NS2 protein) mediates the nuclear export of viral ribonucleoproteins. *The EMBO journal* **17,** 288–96. ISSN: 0261-4189 (1998).

32. Shimizu, T., Takizawa, N., Watanabe, K., Nagata, K. & Kobayashi, N. Crucial role of the influenza virus NS2 (NEP) C-terminal domain in M1 binding and nuclear export of vRNP. *FEBS Letters* **585,** 41–46. ISSN: 00145793 (2011).

33. Leser, G. P. & Lamb, R. A. Influenza virus assembly and budding in raft-derived microdomains: A quantitative analysis of the surface distribution of HA, NA and M2 proteins. *Virology* **342,** 215–227. ISSN: 00426822 (2005).

34. Rossman, J. S., Jing, X., Leser, G. P. & Lamb, R. A. Influenza virus M2 protein mediates ESCRT-independent membrane scission. *Cell* **142,** 902–13. ISSN: 1097-4172 (2010).

35. Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. *Nature Reviews Microbiology* **16,** 47–60. ISSN: 1740-1526 (2017).

36. Doud, M. B., Hensley, S. E. & Bloom, J. D. Complete mapping of viral escape from neutralizing antibodies. *PLOS Pathogens* **13** (ed Lauring, A. S.) e1006271. ISSN: 1553-7374 (2017).

37. Koel, B. F. *et al.* Substitutions Near the Receptor Binding Site Determine Major Antigenic Change During Influenza Virus Evolution. *Science* **342,** 976–979. ISSN: 0036-8075 (2013).

38. Mathew, B., Daniel, R. & Campbell, I. Swine-Origin Influenza A (H1N1) Pandemic Revisited. *The Libyan journal of medicine* **4,** 123–5. ISSN: 1993-2820 (2009).

39. Stöhr, K. The Global Agenda on Influenza Surveillance and Control. *Vaccine* **21,** 1744–1748. ISSN: 0264-410X (2003).

40. Biggerstaff, M., Cauchemez, S., Reed, C., Gambhir, M. & Finelli, L. Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *BMC Infectious Diseases* **14,** 480. ISSN: 1471-2334 (2014).

41. Bedford, T. *et al.* Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523,** 217–220. ISSN: 0028-0836 (2015).

42. WHO. *WHO Factsheet - Seasonal Influenza* 2018. `http://www.who.int/mediacentre/factsheets/fs211/en/` (2018).

43. Iuliano, A. D. *et al.* Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet (London, England)* **391,** 1285–1300. ISSN: 1474-547X (2018).

44. Cassini, A. *et al.* Impact of infectious diseases on population health using incidence-based disability-adjusted life years (DALYs): results from the Burden of Communicable Diseases in Europe study, European Union and European Economic Area countries, 2009 to 2013. *Eurosurveillance* **23,** 17–00454. ISSN: 1560-7917 (2018).

45. Malosh, R. E. *et al.* Efficacy and Safety of Oseltamivir in Children: Systematic Review and Individual Patient Data Meta-analysis of Randomized Controlled Trials. *Clinical Infectious Diseases* **66,** 1492–1500. ISSN: 1058-4838 (2018).

46. Taubenberger, J. K. & Morens, D. M. Pandemic influenza–including a risk assessment of H5N1. *Revue scientifique et technique (International Office of Epizootics)* **28,** 187–202. ISSN: 0253-1933 (2009).

47. Taubenberger, J. K. & Morens, D. M. 1918 Influenza: the mother of all pandemics. *Emerging infectious diseases* **12,** 15–22. ISSN: 1080-6040 (2006).

48. Johnson, N. P.A. S. & Mueller, J. Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bulletin of the history of medicine* **76,** 105–15. ISSN: 0007-5140 (2002).

49. Institute for Disease Modelling. *Pandemic Simulation used in Gates Shattuck Lecture — Institute for Disease Modeling* 2018. `http://www.idmod.org/news/node/296` (2018).

50. Taubenberger, J. K. *et al.* Characterization of the 1918 influenza virus polymerase genes. *Nature* **437,** 889–893. ISSN: 0028-0836 (2005).

51. Morens, D. M., Taubenberger, J. K. & Fauci, A. S. The persistent legacy of the 1918 influenza virus. *The New England journal of medicine* **361,** 225–9. ISSN: 1533-4406 (2009).

52. Dawood, F. S. *et al.* Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet. Infectious diseases* **12,** 687–95. ISSN: 1474-4457 (2012).

53. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459,** 1122–1125. ISSN: 0028-0836 (2009).

54. Reed, C. *et al.* Complications Among Adults Hospitalized With Influenza: A Comparison of Seasonal Influenza and the 2009 H1N1 Pandemic. *Clinical Infectious Diseases* **59,** 166–174. ISSN: 1058-4838 (2014).

55. Wan, X.-F. *et al.* Indications that live poultry markets are a major source of human H5N1 influenza virus infection in China. *Journal of virology* **85,** 13432–8. ISSN: 1098-5514 (2011).

56. Bowman, A. S. *et al.* Influenza A(H3N2) Virus in Swine at Agricultural Fairs and Transmission to Humans, Michigan and Ohio, USA, 2016. *Emerging Infectious Diseases* **23,** 1551–1555. ISSN: 1080-6040 (2017).

57. WHO. *Influenza (Avian and other zoonotic) - WHO Fact Sheet* 2018. `http://www.who.int/news-room/fact-sheets/detail/influenza-(avian-and-other-zoonotic)` (2018).

58. Mertz, D. *et al.* Populations at risk for severe or complicated influenza illness: systematic review and meta-analysis. *BMJ (Clinical research ed.)* **347,** f5061. ISSN: 1756-1833 (2013).

59. Tempia, S. *et al.* Risk Factors for Influenza-Associated Severe Acute Respiratory Illness Hospitalization in South Africa, 2012-2015. *Open forum infectious diseases* **4,** ofw262. ISSN: 2328-8957 (2017).

60. The Writing Committee of the World Health Organization (WHO) Consultation on Human Influenza. *Avian Influenza A (H5N1) Infection in Humans* tech. rep. 13 (2005), 1374–1385. doi:10.1056/NEJMra052211. `http://www.nejm.org/doi/abs/10.1056/NEJMra052211`.

61. Ciancanelli, M. J. *et al.* Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency. *Science* **348,** 448–453. ISSN: 0036-8075 (2015).

62. Esposito, S. *et al.* Toll-like receptor 3 gene polymorphisms and severity of pandemic A/H1N1/2009 influenza in otherwise healthy children. *Virology journal* **9,** 270. ISSN: 1743-422X (2012).

63. Martinez-Ocaña, J. *et al.* Plasma cytokine levels and cytokine gene polymorphisms in Mexican patients during the influenza pandemic A(H1N1)pdm09. *Journal of Clinical Virology* **58,** 108–113. ISSN: 13866532 (2013).

64. Liu, Y. *et al.* Genetic variants in IL1A and IL1B contribute to the susceptibility to 2009 pandemic H1N1 influenza A virus. *BMC immunology* **14,** 37. ISSN: 1471-2172 (2013).

65. Herrera-Ramos, E. *et al.* Surfactant protein A genetic variants associate with severe respiratory insufficiency in pandemic influenza A virus infection. *Critical Care* **18,** R127. ISSN: 1364-8535 (2014).

66. Chen, Y. *et al.* Functional variants regulating LGALS1 (Galectin 1) expression affect human susceptibility to influenza A(H7N9). *Scientific Reports* **5,** 8517. ISSN: 2045-2322 (2015).

67. Cheng, Z. *et al.* Identification of TMPRSS2 as a Susceptibility Gene for Severe 2009 Pandemic A(H1N1) Influenza and A(H7N9) Influenza. *Journal of Infectious Diseases* **212,** 1214–1221. ISSN: 0022-1899 (2015).

68. Garcia-Etxebarria, K. *et al.* No Major Host Genetic Risk Factor Contributed to A(H1N1)2009 Influenza Severity. *PloS one* **10,** e0135983. ISSN: 1932-6203 (2015).

69. Pommerenke, C. *et al.* Global Transcriptome Analysis in Influenza-Infected Mouse Lungs Reveals the Kinetics of Innate and Adaptive Host Immune Responses. *PLoS ONE* **7** (ed Kassiotis, G.) e41169. ISSN: 1932-6203 (2012).

70. Cohen, M. *et al.* Influenza A penetrates host mucus by cleaving sialic acids with neuraminidase. *Virology Journal* **10,** 321. ISSN: 1743-422X (2013).

71. Bahadoran, A. *et al.* Immune Responses to Influenza Virus and Its Correlation to Age and Inherited Factors. *Frontiers in microbiology* **7,** 1841. ISSN: 1664-302X (2016).

72. Sarvestani, S. T. & McAuley, J. L. The role of the NLRP3 inflammasome in regulation of antiviral responses to influenza A virus infection. *Antiviral Research* **148,** 32–42. ISSN: 01663542 (2017).

73. Brandes, M., Klauschen, F., Kuchen, S. & Germain, R. N. A systems analysis identifies a feedforward inflammatory circuit leading to lethal influenza infection. *Cell* **154,** 197–212. ISSN: 1097-4172 (2013).

74. Pulendran, B. & Maddur, M. S. Innate immune sensing and response to influenza. *Current topics in microbiology and immunology* **386,** 23–71. ISSN: 0070-217X (2015).

75. Helft, J. *et al.* Cross-presenting CD103+ dendritic cells are protected from influenza virus infection. *Journal of Clinical Investigation* **122,** 4037–4047. ISSN: 0021-9738 (2012).

76. Cao, W. *et al.* Rapid Differentiation of Monocytes into Type I IFN-Producing Myeloid Dendritic Cells as an Antiviral Strategy against Influenza Virus Infection. *The Journal of Immunology* **189,** 2257–2265. ISSN: 0022-1767 (2012).

77. Lin, K. L., Sweeney, S., Kang, B. D., Ramsburg, E. & Gunn, M. D. CCR2-antagonist prophylaxis reduces pulmonary immune pathology and markedly improves survival during influenza infection. *Journal of immunology (Baltimore, Md. : 1950)* **186,** 508–15. ISSN: 1550-6606 (2011).

78. Steinman, R. M. & Banchereau, J. Taking dendritic cells into medicine. *Nature* **449,** 419–426. ISSN: 0028-0836 (2007).

79. Tversky, J. R. *et al.* Human blood dendritic cells from allergic subjects have impaired capacity to produce interferon-alpha via Toll-like receptor 9. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **38,** 781–8. ISSN: 1365-2222 (2008).

80. Sandbulte, M. R., Boon, A. C. M., Webby, R. J. & Riberdy, J. M. Analysis of cytokine secretion from human plasmacytoid dendritic cells infected with H5N1 or low-pathogenicity influenza viruses. *Virology* **381,** 22–8. ISSN: 1096-0341 (2008).

81. Van de Sandt, C. E., Kreijtz, J. H.C. M. & Rimmelzwaan, G. F. Evasion of Influenza A Viruses from Innate and Adaptive Immune Responses. *Viruses* **4,** 1438–1476. ISSN: 1999-4915 (2012).

82. Jost, S. & Altfeld, M. Control of Human Viral Infections by Natural Killer Cells. *Annual Review of Immunology* **31,** 163–194. ISSN: 0732-0582 (2013).

83. Mandelboim, O. *et al.* Recognition of haemagglutinins on virus-infected cells by NKp46 activates lysis by human NK cells. *Nature* **409,** 1055–1060. ISSN: 00280836 (2001).

84. Ho, A. W. S. *et al.* Lung CD103+ Dendritic Cells Efficiently Transport Influenza Virus to the Lymph Node and Load Viral Antigen onto MHC Class I for Presentation to CD8 T Cells. *The Journal of Immunology* **187,** 6011–6021. ISSN: 0022-1767 (2011).

85. Chen, X. *et al.* Host Immune Response to Influenza A Virus Infection. *Frontiers in immunology* **9,** 320. ISSN: 1664-3224 (2018).

86. McKinstry, K. K. *et al.* Memory CD4+ T cells protect against influenza through multiple synergizing mechanisms. *Journal of Clinical Investigation* **122,** 2847–2856. ISSN: 0021-9738 (2012).

87. Seibert, C. W. *et al.* Recombinant IgA Is Sufficient To Prevent Influenza Virus Transmission in Guinea Pigs. *Journal of Virology* **87,** 7793–7804. ISSN: 0022-538X (2013).

88. Bedford, T. *et al.* Integrating influenza antigenic dynamics with molecular evolution. *eLife* **3,** e01914. ISSN: 2050-084X (2014).

89. Zost, S. J. *et al.* Contemporary H3N2 influenza viruses have a glycosylation site that alters binding of antibodies elicited by egg-adapted vaccine strains. *Proceedings of the National Academy of Sciences of the United States of America* **114,** 12578–12583. ISSN: 1091-6490 (2017).

90. Milián, E. & Kamen, A. A. Current and emerging cell culture manufacturing technologies for influenza vaccines. *BioMed research international* **2015,** 504831. ISSN: 2314-6141 (2015).

91. CDC. *Seasonal Influenza Vaccine Effectiveness, 2005-2018* 2018. https://www.cdc.gov/flu/professionals/vaccination/effectiveness-studies.htm (2018).

92. Waffarn, E. E. & Baumgarth, N. Protective B cell responses to flu–no fluke! *Journal of immunology (Baltimore, Md. : 1950)* **186,** 3823–9. ISSN: 1550-6606 (2011).

93. Lee, L. Y.-H. *et al.* Memory T cells established by seasonal human influenza A infection cross-react with avian influenza A (H5N1) in healthy individuals. *Journal of Clinical Investigation* **118,** 3478–90. ISSN: 0021-9738 (2008).

94. Spitaels, J., Roose, K. & Saelens, X. Influenza and Memory T Cells: How to Awake the Force. *Vaccines* **4.** ISSN: 2076-393X. doi:10.3390/vaccines4040033. http://www.ncbi.nlm.nih.gov/pubmed/27754364http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5192353 (2016).

95. Belshe, R. B. *et al.* The Efficacy of Live Attenuated, Cold-Adapted, Trivalent, Intranasal Influenzavirus Vaccine in Children. *New England Journal of Medicine* **338,** 1405–1412. ISSN: 0028-4793 (1998).

96. Mohn, K. G. I., Zhou, F., Brokstad, K. A., Sridhar, S. & Cox, R. J. Boosting of Cross-Reactive and Protection-Associated T Cells in Children After Live Attenuated Influenza Vaccination. *The Journal of Infectious Diseases* **215,** 1527–1535. ISSN: 0022-1899 (2017).

97. Pebody, R., McMenamin, J. & Nohynek, H. Live attenuated influenza vaccine (LAIV): recent effectiveness results from the USA and implications for LAIV programmes elsewhere. *Archives of disease in childhood* **103,** 101–105. ISSN: 1468-2044 (2018).

98. Berthoud, T. K. *et al.* Potent CD8+ T-Cell Immunogenicity in Humans of a Novel Heterosubtypic Influenza A Vaccine, MVA-NP+M1. *Clinical Infectious Diseases* **52,** 1–7. ISSN: 1058-4838 (2011).

99. Lillie, P. J. *et al.* Preliminary Assessment of the Efficacy of a T-Cell–Based Influenza Vaccine, MVA-NP+M1, in Humans. *Clinical Infectious Diseases* **55,** 19–25. ISSN: 1537-6591 (2012).

100. Powell, T. J. *et al.* Examination of Influenza Specific T Cell Responses after Influenza Virus Challenge in Individuals Vaccinated with MVA-NP+M1 Vaccine. *PLoS ONE* **8** (ed Alsharifi, M.) e62778. ISSN: 1932-6203 (2013).

101. Iwasaki, A. & Pillai, P. S. Innate immunity to influenza virus infection. *Nature Reviews Immunology* **14,** 315–328. ISSN: 1474-1733 (2014).

102. Pawson, T & Scott, J. D. Signaling through scaffold, anchoring, and adaptor proteins. *Science (New York, N.Y.)* **278,** 2075–80. ISSN: 0036-8075 (1997).

103. Thompson, M. R., Kaminski, J. J., Kurt-Jones, E. A. & Fitzgerald, K. A. Pattern recognition receptors and the innate immune response to viral infection. *Viruses* **3,** 920–40. ISSN: 1999-4915 (2011).

104. Lund, J. M. *et al.* Recognition of single-stranded RNA viruses by Toll-like receptor 7. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 5598–603. ISSN: 0027-8424 (2004).

105. Zhang, Z. *et al.* Structural Analysis Reveals that Toll-like Receptor 7 Is a Dual Receptor for Guanosine and Single-Stranded RNA. *Immunity* **45,** 737–748. ISSN: 10747613 (2016).

106. Seo, S.-U. *et al.* MyD88 Signaling Is Indispensable for Primary Influenza A Virus Infection but Dispensable for Secondary Infection. *Journal of Virology* **84,** 12713–12722. ISSN: 0022-538X (2010).

107. Jeisy-Scott, V. *et al.* TLR7 Recognition Is Dispensable for Influenza Virus A Infection but Important for the Induction of Hemagglutinin-Specific Antibodies in Response to the 2009 Pandemic Split Vaccine in Mice. *Journal of Virology* **86,** 10988–10998. ISSN: 0022-538X (2012).

108. Lester, S. N. & Li, K. Toll-Like Receptors in Antiviral Innate Immunity. *Journal of Molecular Biology* **426,** 1246–1264. ISSN: 0022-2836 (2014).

109. Guillot, L. *et al.* Involvement of Toll-like Receptor 3 in the Immune Response of Lung Epithelial Cells to Double-stranded RNA and Influenza A Virus. *Journal of Biological Chemistry* **280,** 5571–5580. ISSN: 0021-9258 (2005).

110. Wu, W. *et al.* RIG-I and TLR3 are both required for maximum interferon induction by influenza virus in human lung alveolar epithelial cells. *Virology* **482,** 181–8. ISSN: 1096-0341 (2015).

111. Wisskirchen, C., Ludersdorfer, T. H., Muller, D. A., Moritz, E. & Pavlovic, J. The Cellular RNA Helicase UAP56 Is Required for Prevention of Double-Stranded RNA Formation during Influenza A Virus Infection. *Journal of Virology* **85,** 8646–8655. ISSN: 0022-538X (2011).

112. Schulz, O. *et al.* Toll-like receptor 3 promotes cross-priming to virus-infected cells. *Nature* **433,** 887–892. ISSN: 0028-0836 (2005).

113. Kawasaki, T. & Kawai, T. Toll-like receptor signaling pathways. *Frontiers in immunology* **5,** 461. ISSN: 1664-3224 (2014).

114. Kato, H. *et al.* Cell Type-Specific Involvement of RIG-I in Antiviral Response. *Immunity* **23,** 19–28. ISSN: 10747613 (2005).

115. Rehwinkel, J. *et al.* RIG-I Detects Viral Genomic RNA during Negative-Strand RNA Virus Infection. *Cell* **140,** 397–408. ISSN: 00928674 (2010).

116. Kolakofsky, D., Kowalinski, E. & Cusack, S. A structure-based model of RIG-I activation. *RNA (New York, N.Y.)* **18,** 2118–27. ISSN: 1469-9001 (2012).

117. Onomoto, K. *et al.* Critical Role of an Antiviral Stress Granule Containing RIG-I and PKR in Viral Detection and Innate Immunity. *PLoS ONE* **7** (ed Kanai, A.) e43031. ISSN: 1932-6203 (2012).

118. Koyama, S. *et al.* Differential role of TLR- and RLR-signaling in the immune responses to influenza A virus infection and vaccination. *Journal of immunology (Baltimore, Md. : 1950)* **179,** 4711–20. ISSN: 0022-1767 (2007).

119. Kowalinski, E. *et al.* Structural Basis for the Activation of Innate Immune Pattern-Recognition Receptor RIG-I by Viral RNA. *Cell* **147,** 423–435. ISSN: 00928674 (2011).

120. Bergsbaken, T., Fink, S. L. & Cookson, B. T. Pyroptosis: host cell death and inflammation. *Nature Reviews Microbiology* **7,** 99–109. ISSN: 1740-1526 (2009).

121. Tate, M. D. *et al.* Reassessing the role of the NLRP3 inflammasome during pathogenic influenza A virus infection via temporal inhibition. *Scientific Reports* **6,** 27912. ISSN: 2045-2322 (2016).

122. McAuley, J. L. *et al.* Activation of the NLRP3 Inflammasome by IAV Virulence Protein PB1-F2 Contributes to Severe Pathophysiology and Disease. *PLoS Pathogens* **9** (ed Gack, M. U.) e1003392. ISSN: 1553-7374 (2013).

123. Hoffmann, H.-H., Schneider, W. M. & Rice, C. M. Interferons and viruses: an evolutionary arms race of molecular interactions. *Trends in immunology* **36,** 124–38. ISSN: 1471-4981 (2015).

124. Genin, P., Lin, R., Hiscott, J. & Civas, A. Differential Regulation of Human Interferon A Gene Expression by Interferon Regulatory Factors 3 and 7. *Molecular and Cellular Biology* **29,** 3435–3450. ISSN: 0270-7306 (2009).

125. Ho, H. H. & Ivashkiv, L. B. Role of STAT3 in Type I Interferon Responses. *Journal of Biological Chemistry* **281,** 14111–14118. ISSN: 0021-9258 (2006).

126. Piehler, J., Thomas, C., Garcia, K. C. & Schreiber, G. Structural and dynamic determinants of type I interferon receptor assembly and their functional interpretation. *Immunological reviews* **250,** 317–34. ISSN: 1600-065X (2012).

127. Schroder, K., Hertzog, P. J., Ravasi, T. & Hume, D. A. Interferon-γ: an overview of signals, mechanisms and functions. *Journal of Leukocyte Biology* **75,** 163–189. ISSN: 07415400 (2004).

128. Gołąb, J. Interleukin 18-interferon gamma inducing actor - a novel player in tumour immunotherapy? *Cytokine* **12,** 332–338. ISSN: 10434666 (2000).

129. Kraaij, M. D. *et al.* Human monocytes produce interferon-gamma upon stimulation with LPS. *Cytokine* **67,** 7–12. ISSN: 1043-4666 (2014).

130. Fultz, M. J., Barber, S. A., Dieffenbach, C. W. & Vogel, S. N. Induction of IFN-γ in macrophages by lipopolysaccharide. *International Immunology* **5,** 1383–1392. ISSN: 0953-8178 (1993).

131. Briscoe, J. *et al.* JAKs, STATs and Signal Transduction in Response to the Interferons and other Cytokines [and Discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences* **351,** 167–171. ISSN: 0962-8436 (1996).

132. Sakamoto, E. *et al.* Type I and type II interferons delay human neutrophil apoptosis via activation of STAT3 and up-regulation of cellular inhibitor of apoptosis 2. *Journal of Leukocyte Biology* **78,** 301–309. ISSN: 0741-5400 (2005).

133. Hermant, P. & Michiels, T. Interferon-λ in the Context of Viral Infections: Production, Response and Therapeutic Implications. *Journal of Innate Immunity* **6,** 563–574. ISSN: 1662-8128 (2014).

134. Iversen, M. B., Ank, N., Melchjorsen, J. & Paludan, S. R. Expression of Type III Interferon (IFN) in the Vaginal Mucosa Is Mediated Primarily by Dendritic Cells and Displays Stronger Dependence on NF- B than Type I IFNs. *Journal of Virology* **84,** 4579–4586. ISSN: 0022-538X (2010).

135. Lazear, H. M., Nice, T. J. & Diamond, M. S. Interferon-λ: Immune Functions at Barrier Surfaces and Beyond. *Immunity* **43,** 15–28. ISSN: 1097-4180 (2015).

136. Marcello, T. *et al.* Interferons α and λ Inhibit Hepatitis C Virus Replication With Distinct Signal Transduction and Gene Regulation Kinetics. *Gastroenterology* **131,** 1887–1898. ISSN: 00165085 (2006).

137. Gough, D. J., Messina, N. L., Clarke, C. J. P., Johnstone, R. W. & Levy, D. E. Constitutive Type I Interferon Modulates Homeostatic Balance through Tonic Signaling. *Immunity* **36,** 166–174. ISSN: 10747613 (2012).

138. Mostafavi, S. *et al.* Parsing the Interferon Transcriptional Network and Its Disease Associations. *Cell* **164,** 564–78. ISSN: 1097-4172 (2016).

139. Gough, D. J. *et al.* Functional Crosstalk between Type I and II Interferon through the Regulated Expression of STAT1. *PLoS Biology* **8** (ed Virgin, S. W.) e1000361. ISSN: 1545-7885 (2010).

140. Zurney, J., Howard, K. E. & Sherry, B. Basal Expression Levels of IFNAR and Jak-STAT Components Are Determinants of Cell-Type-Specific Differences in Cardiac Antiviral Responses. *Journal of Virology* **81,** 13668–13680. ISSN: 0022-538X (2007).

141. Crow, M. K. Type I interferon in organ-targeted autoimmune and inflammatory diseases. *Arthritis research & therapy* **12 Suppl 1,** S5. ISSN: 1478-6362 (2010).

142. Hunter, C. A., Gabriel, K. E., Radzanowski, T, Neyer, L. E. & Remington, J. S. Type I interferons enhance production of IFN-gamma by NK cells. *Immunology letters* **59,** 1–5. ISSN: 0165-2478 (1997).

143. Mack, E. A., Kallal, L. E., Demers, D. A. & Biron, C. A. Type 1 interferon induction of natural killer cell gamma interferon production for defense during lymphocytic choriomeningitis virus infection. *mBio* **2,** e00169–11. ISSN: 2150-7511 (2011).

144. Rayamajhi, M., Humann, J., Kearney, S., Hill, K. K. & Lenz, L. L. Antagonistic crosstalk between type I and II interferons and increased host susceptibility to bacterial infections. *Virulence* **1,** 418–22. ISSN: 2150-5608 (2010).

145. Eshleman, E. M., Delgado, C., Kearney, S. J., Friedman, R. S. & Lenz, L. L. Down regulation of macrophage IFNGR1 exacerbates systemic L. monocytogenes infection. *PLOS Pathogens* **13** (ed Coers, J.) e1006388. ISSN: 1553-7374 (2017).

146. Jaitin, D. A. *et al.* Inquiring into the differential action of interferons (IFNs): an IFN-alpha2 mutant with enhanced affinity to IFNAR1 is functionally similar to IFN-beta. *Molecular and cellular biology* **26,** 1888–97. ISSN: 0270-7306 (2006).

147. Qian, J. *et al.* Pathogen recognition receptor signaling accelerates phosphorylation-dependent degradation of IFNAR1. *PLoS pathogens* **7,** e1002065. ISSN: 1553-7374 (2011).

148. Davidson, S., Crotta, S., McCabe, T. M. & Wack, A. Pathogenic potential of interferon $\alpha\beta$ in acute influenza infection. *Nature Communications* **5,** 3864. ISSN: 2041-1723 (2014).

149. Li, W., Moltedo, B. & Moran, T. M. Type I interferon induction during influenza virus infection increases susceptibility to secondary Streptococcus pneumoniae infection by negative regulation of $\gamma\delta$ T cells. *Journal of virology* **86,** 12304–12. ISSN: 1098-5514 (2012).

150. Galani, I. E. *et al.* Interferon-λ Mediates Non-redundant Front-Line Antiviral Protection against Influenza Virus Infection without Compromising Host Fitness. *Immunity* **46,** 875–890.e6. ISSN: 1074-7613 (2017).

151. Davidson, S. *et al.* IFNλ is a potent anti-influenza therapeutic without the inflammatory side effects of IFNα treatment. *EMBO molecular medicine* **8,** 1099–112. ISSN: 1757-4684 (2016).

152. Ivashkiv, L. B. & Donlin, L. T. Regulation of type I interferon responses. *Nature Reviews Immunology* **14,** 36–49. ISSN: 1474-1733 (2014).

153. Stifter, S. A. *et al.* Functional Interplay between Type I and II Interferons Is Essential to Limit Influenza A Virus-Induced Tissue Inflammation. *PLOS Pathogens* **12,** e1005378. ISSN: 1553-7374 (2016).

154. Califano, D *et al.* IFN-γ increases susceptibility to influenza A infection through suppression of group II innate lymphoid cells. *Mucosal Immunology* **11,** 209–219. ISSN: 1933-0219 (2018).

155. Liu, Q., Zhou, Y.-h. & Yang, Z.-q. The cytokine storm of severe influenza and development of immunomodulatory therapy. *Cellular & molecular immunology* **13,** 3–10. ISSN: 2042-0226 (2016).

156. Rusinova, I. *et al.* Interferome v2.0: an updated database of annotated interferon-regulated genes. *Nucleic acids research* **41,** D1040–6. ISSN: 1362-4962 (2013).

157. Schreiber, G. The molecular basis for differential type I interferon signaling. *The Journal of biological chemistry* **292,** 7285–7294. ISSN: 1083-351X (2017).

158. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498,** 236–240. ISSN: 0028-0836 (2013).

159. Schoggins, J. W. & Rice, C. M. Interferon-stimulated genes and their antiviral effector functions. *Current opinion in virology* **1,** 519–25. ISSN: 1879-6265 (2011).

160. Villalón-Letelier, F., Brooks, A., Saunders, P., Londrigan, S. & Reading, P. Host Cell Restriction Factors that Limit Influenza A Infection. *Viruses* **9,** 376. ISSN: 1999-4915 (2017).

161. Deeg, C. M. *et al.* In vivo evasion of MxA by avian influenza viruses requires human signature in the viral nucleoprotein. *The Journal of Experimental Medicine* **214,** 1239–1248. ISSN: 0022-1007 (2017).

162. Matzinger, S. R., Carroll, T. D., Dutra, J. C., Ma, Z.-M. & Miller, C. J. Myxovirus Resistance Gene A (MxA) Expression Suppresses Influenza A Virus Replication in Alpha Interferon-Treated Primate Cells. *Journal of Virology* **87,** 1150–1158. ISSN: 0022-538X (2013).

163. Pavlovic, J, Haller, O & Staeheli, P. Human and mouse Mx proteins inhibit different steps of the influenza virus multiplication cycle. *Journal of virology* **66,** 2564–9. ISSN: 0022-538X (1992).

164. Gusho, E., Baskar, D. & Banerjee, S. New advances in our understanding of the "unique" RNase L in host pathogen interaction and immune signaling. *Cytokine.* ISSN: 10434666. doi:10.1016/j.cyto.2016.08.009 (2016).

165. Dauber, B. & Wolff, T. Activation of the Antiviral Kinase PKR and Viral Countermeasures. *Viruses* **1,** 523–544. ISSN: 1999-4915 (2009).

166. Wang, X., Hinson, E. R. & Cresswell, P. The Interferon-Inducible Protein Viperin Inhibits Influenza Virus Release by Perturbing Lipid Rafts. *Cell Host & Microbe* **2,** 96–105. ISSN: 19313128 (2007).

167. Blanc, M. *et al.* The Transcription Factor STAT-1 Couples Macrophage Synthesis of 25-Hydroxycholesterol to the Interferon Antiviral Response. *Immunity* **38,** 106–118. ISSN: 1074-7613 (2013).

168. Di Pietro, A. *et al.* TRIM22 Inhibits Influenza A Virus Infection by Targeting the Viral Nucleoprotein for Degradation. *Journal of Virology* **87,** 4523–4533. ISSN: 0022-538X (2013).

169. Zhao, C., Hsiang, T.-Y., Kuo, R.-L. & Krug, R. M. ISG15 conjugation system targets the viral NS1 protein in influenza A virus-infected cells. *Proceedings of the National Academy of Sciences* **107,** 2253–2258. ISSN: 0027-8424 (2010).

170. Liu, C.-H., Zhou, L., Chen, G. & Krug, R. M. Battle between influenza A virus and a newly identified antiviral activity of the PARP-containing ZAPL protein. *Proceedings of the National Academy of Sciences* **112,** 14048–14053. ISSN: 0027-8424 (2015).

171. Tang, Q., Wang, X. & Gao, G. The Short Form of the Zinc Finger Antiviral Protein Inhibits Influenza A Virus Protein Expression and Is Antagonized by the Virus-Encoded NS1. *Journal of Virology* **91** (ed Lyles, D. S.) ISSN: 0022-538X. doi:10.1128/JVI.01909-16 (2017).

172. Dittmann, M. *et al.* A Serpin Shapes the Extracellular Environment to Prevent Influenza A Virus Maturation. *Cell* **160,** 631–643. ISSN: 00928674 (2015).

173. Qu, H. *et al.* Influenza A Virus-induced expression of ISG20 inhibits viral replication by interacting with nucleoprotein. *Virus Genes* **52,** 759–767. ISSN: 0920-8569 (2016).

174. Gack, M. U. *et al.* Influenza A Virus NS1 Targets the Ubiquitin Ligase TRIM25 to Evade Recognition by the Host Viral RNA Sensor RIG-I. *Cell Host & Microbe* **5,** 439–449. ISSN: 19313128 (2009).

175. Rajsbaum, R. *et al.* Species-Specific Inhibition of RIG-I Ubiquitination and IFN Induction by the Influenza A Virus NS1 Protein. *PLoS Pathogens* **8** (ed Pekosz, A.) e1003059. ISSN: 1553-7374 (2012).

176. Jureka, A., Kleinpeter, A., Cornilescu, G., Cornilescu, C. & Petit, C. Structural Basis for a Novel Interaction between the NS1 Protein Derived from the 1918 Influenza Virus and RIG-I. *Structure* **23,** 2001–2010. ISSN: 09692126 (2015).

177. Talon, J *et al.* Activation of interferon regulatory factor 3 is inhibited by the influenza A virus NS1 protein. *Journal of virology* **74,** 7989–96. ISSN: 0022-538X (2000).

178. Wang, X *et al.* Influenza A virus NS1 protein prevents activation of NF-kappaB and induction of alpha/beta interferon. *Journal of virology* **74,** 11566–73. ISSN: 0022-538X (2000).

179. Anastasina, M. *et al.* Influenza virus NS1 protein binds cellular DNA to block transcription of antiviral genes. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1859,** 1440–1448. ISSN: 1874-9399 (2016).

180. Nemeroff, M. E., Barabino, S. M., Li, Y, Keller, W & Krug, R. M. Influenza virus NS1 protein interacts with the cellular 30 kDa subunit of CPSF and inhibits 3'end formation of cellular pre-mRNAs. *Molecular cell* **1,** 991–1000. ISSN: 1097-2765 (1998).

181. Lu, Y, Wambach, M, Katze, M. G. & Krug, R. M. Binding of the influenza virus NS1 protein to double-stranded RNA inhibits the activation of the protein kinase that phosphorylates the elF-2 translation initiation factor. *Virology* **214,** 222–8. ISSN: 0042-6822 (1995).

182. Silverman, R. H. Viral encounters with 2',5'-oligoadenylate synthetase and RNase L during the interferon antiviral response. *Journal of virology* **81,** 12720–9. ISSN: 1098-5514 (2007).

183. Hanagata, N. IFITM5 mutations and osteogenesis imperfecta. *Journal of bone and mineral metabolism* **34,** 123–31. ISSN: 1435-5604 (2016).

184. Sällman Almén, M., Bringeland, N., Fredriksson, R. & Schiöth, H. B. The Dispanins: A Novel Gene Family of Ancient Origin That Contains 14 Human Members. *PLoS ONE* **7** (ed Uversky, V. N.) e31961. ISSN: 1932-6203 (2012).

185. Zhang, Z., Liu, J., Li, M., Yang, H. & Zhang, C. Evolutionary Dynamics of the Interferon-Induced Transmembrane Gene Family in Vertebrates. *PLoS ONE* **7,** e49265. ISSN: 1932-6203 (2012).

186. Andreu, P. *et al.* Identification of the IFITM Family as a New Molecular Marker in Human Colorectal Tumors. *Cancer Research* **66,** 1949–1955. ISSN: 0008-5472 (2006).

187. Zhang, D., Wang, H., He, H., Niu, H. & Li, Y. Interferon induced transmembrane protein 3 regulates the growth and invasion of human lung adenocarcinoma. *Thoracic cancer* **8,** 337–343. ISSN: 1759-7714 (2017).

188. Tanaka, S. S., Yamaguchi, Y. L., Tsoi, B., Lickert, H. & Tam, P. P. IFITM/Mil/Fragilis Family Proteins IFITM1 and IFITM3 Play Distinct Roles in Mouse Primordial Germ Cell Homing and Repulsion. *Developmental Cell* **9,** 745–756. ISSN: 1534-5807 (2005).

189. Friedman, R. L., Manly, S. P., McMahon, M, Kerr, I. M. & Stark, G. R. Transcriptional and posttranscriptional regulation of interferon-induced gene expression in human cells. *Cell* **38,** 745–55. ISSN: 0092-8674 (1984).

190. Matsumoto, A. K. *et al.* Functional dissection of the CD21/CD19/TAPA-1/Leu-13 complex of B lymphocytes. *The Journal of experimental medicine* **178,** 1407–17. ISSN: 0022-1007 (1993).

191. Brass, A. L. *et al.* The IFITM Proteins Mediate Cellular Resistance to Influenza A H1N1 Virus, West Nile Virus, and Dengue Virus. *Cell* **139,** 1243–1254. ISSN: 00928674 (2009).

192. Narayana, S. K. *et al.* The interferon-induced transmembrane proteins, IFITM1, IFITM2, and IFITM3 inhibit hepatitis C virus entry. *Journal of Biological Chemistry* **290,** 25946–25959. ISSN: 1083351X (2015).

193. Savidis, G. *et al.* The IFITMs Inhibit Zika Virus Replication. *Cell Reports* **15,** 2323–2330. ISSN: 22111247 (2016).

194. Zhang, L.-K., Chai, F., Li, H.-Y., Xiao, G. & Guo, L. Identification of Host Proteins Involved in Japanese Encephalitis Virus Infection by Quantitative Proteomics Analysis. *Journal of Proteome Research* **12,** 2666–2678. ISSN: 1535-3893 (2013).

195. Huang, I.-C. *et al.* Distinct Patterns of IFITM-Mediated Restriction of Filoviruses, SARS Coronavirus, and Influenza A Virus. *PLoS Pathogens* **7** (ed Baric, R. S.) e1001258. ISSN: 1553-7374 (2011).

196. Wrensch, F., Winkler, M. & Pöhlmann, S. IFITM Proteins Inhibit Entry Driven by the MERS-Coronavirus Spike Protein: Evidence for Cholesterol-Independent Mechanisms. *Viruses* **6,** 3683–3698. ISSN: 1999-4915 (2014).

197. Lu, J., Pan, Q., Rong, L., Liu, S.-L. & Liang, C. The IFITM Proteins Inhibit HIV-1 Infection. *Journal of Virology* **85,** 2126–2137. ISSN: 0022-538X (2011).

198. Qian, J. *et al.* Primate lentiviruses are differentially inhibited by interferon-induced transmembrane proteins. *Virology* **474.** ISSN: 10960341. doi:10.1016/j.virol.2014.10.015 (2015).

199. Mudhasani, R. *et al.* IFITM-2 and IFITM-3 but Not IFITM-1 Restrict Rift Valley Fever Virus. *Journal of Virology* **87,** 8451–8464. ISSN: 0022-538X (2013).

200. Everitt, A. R. *et al.* IFITM3 restricts the morbidity and mortality associated with influenza. *Nature* **484,** 519–523. ISSN: 0028-0836 (2012).

201. Weidner, J. M. *et al.* Interferon-induced cell membrane proteins, IFITM3 and tetherin, inhibit vesicular stomatitis virus infection via distinct mechanisms. *Journal of virology* **84,** 12646–57. ISSN: 1098-5514 (2010).

202. Smith, S. E. *et al.* Chicken interferon-inducible transmembrane protein 3 restricts influenza viruses and lyssaviruses in vitro. *Journal of virology* **87,** 12957–66. ISSN: 1098-5514 (2013).

203. Anafu, A. A., Bowen, C. H., Chin, C. R., Brass, A. L. & Holm, G. H. Interferon-inducible transmembrane protein 3 (IFITM3) restricts reovirus cell entry. *The Journal of biological chemistry* **288,** 17261–71. ISSN: 1083-351X (2013).

204. Zhang, W. *et al.* Human respiratory syncytial virus infection is inhibited by IFN-induced transmembrane proteins. *Journal of General Virology* **96,** 170–182. ISSN: 0022-1317 (2015).

205. Weston, S. *et al.* Alphavirus Restriction by IFITM Proteins. *Traffic* **17.** ISSN: 16000854. doi:10.1111/tra.12416 (2016).

206. Muñoz-Moreno, R. *et al.* Antiviral Role of IFITM Proteins in African Swine Fever Virus Infection. *PLOS ONE* **11** (ed Beer, M.) e0154366. ISSN: 1932-6203 (2016).

207. Stacey, M. A. *et al.* The antiviral restriction factor IFN-induced transmembrane protein 3 prevents cytokine-driven CMV pathogenesis. *Journal of Clinical Investigation* **127.** ISSN: 15588238. doi:10.1172/JCI84889 (2017).

208. Gorman, M. J., Poddar, S., Farzan, M. & Diamond, M. S. The Interferon-Stimulated Gene Ifitm3 Restricts West Nile Virus Infection and Pathogenesis. *Journal of virology* **90,** 8212–25. ISSN: 1098-5514 (2016).

209. Everitt, A. R. *et al.* Defining the Range of Pathogens Susceptible to Ifitm3 Restriction Using a Knockout Mouse Model. *PLoS ONE* **8** (ed Ikeda, Y.) e80723. ISSN: 1932-6203 (2013).

210. Ranjbar, S., Haridas, V., Jasenosky, L. D., Falvo, J. V. & Goldfeld, A. E. A Role for IFITM Proteins in Restriction of Mycobacterium tuberculosis Infection. *Cell Reports* **13,** 874–883. ISSN: 22111247 (2015).

211. Shen, C. *et al.* A Functional Promoter Polymorphism of IFITM3 Is Associated with Susceptibility to Pediatric Tuberculosis in Han Chinese Population. *PLoS ONE* **8,** 1–7. ISSN: 19326203 (2013).

212. Naderi, M. *et al.* Evaluation of interferon-induced transmembrane protein-3 (IFITM3) rs7478728 and rs3888188 polymorphisms and the risk of pulmonary tuberculosis. *Biomedical Reports* **5,** 634–638. ISSN: 2049-9434 (2016).

213. Bailey, C. C., Zhong, G., Huang, I.-C. & Farzan, M. IFITM-Family Proteins: The Cell's First Line of Antiviral Defense. *Annual Review of Virology* **1.** ISSN: 2327-056X. arXiv: NIHMS150003 (2014).

214. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research* **22,** 1775–89. ISSN: 1549-5469 (2012).

215. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking* **13,** 311–319. ISSN: 1947-5535 (2015).

216. LEWIN, A. R., REID, L. E., McMAHON, M., STARK, G. R. & KERR, I. M. Molecular analysis of a human interferon-inducible gene family. *European Journal of Biochemistry* **199,** 417–423. ISSN: 0014-2956 (1991).

217. Cuddapah, S., Cui, K. & Zhao, K. Transcriptional enhancer factor 1 (TEF-1/TEAD1) mediates activation of IFITM3 gene by BRGl. *FEBS Letters* **582,** 391–397. ISSN: 0014-5793 (2008).

218. Scott, R., Siegrist, F., Foser, S. & Certa, U. Interferon-Alpha Induces Reversible DNA Demethylation of the Interferon-Induced Transmembrane Protein-3 Core Promoter in Human Melanoma Cells. *Journal of Interferon & Cytokine Research* **31,** 601–608. ISSN: 1079-9907 (2011).

219. Allen, E. K. *et al.* SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans. *Nature Medicine* **23,** 975. ISSN: 1078-8956 (2017).

220. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410. ISSN: 00222836 (1990).

221. Amini-Bavil-Olyaee, S. *et al.* The antiviral effector IFITM3 disrupts intracellular cholesterol homeostasis to block viral entry. *Cell Host and Microbe* **13.** ISSN: 19346069. doi:10.1016/j.chom.2013.03.006. arXiv: NIHMS150003 (2013).

222. Smith, R. A., Young, J, Weis, J. J. & Weis, J. H. Expression of the mouse fragilis gene products in immune cells and association with receptor signaling complexes. *Genes & Immunity* **7,** 113–121. ISSN: 1466-4879 (2006).

223. Yount, J. S., Karssemeijer, R. A. & Hang, H. C. S-palmitoylation and ubiquitination differentially regulate interferon-induced transmembrane protein 3 (IFITM3)-mediated resistance to influenza virus. *The Journal of biological chemistry* **287,** 19631–41. ISSN: 1083-351X (2012).

224. Jia, R. *et al.* The N-Terminal Region of IFITM3 Modulates Its Antiviral Activity by Regulating IFITM3 Cellular Localization. *Journal of Virology* **86,** 13697–13707. ISSN: 0022-538X (2012).

225. Hach, J. C., McMichael, T., Chesarino, N. M. & Yount, J. S. Palmitoylation on Conserved and Nonconserved Cysteines of Murine IFITM1 Regulates Its Stability and Anti-Influenza A Virus Activity. *Journal of Virology* **87,** 9923–9927. ISSN: 0022-538X (2013).

226. Ling, S. *et al.* Combined approaches of EPR and NMR illustrate only one transmembrane helix in the human IFITM3. *Scientific Reports* **6,** 24029. ISSN: 2045-2322 (2016).

227. Chesarino, N. M. *et al.* IFITM3 requires an amphipathic helix for antiviral activity. *EMBO reports* **18,** 1740–1751. ISSN: 1469-3178 (2017).

228. John, S. P. *et al.* The CD225 Domain of IFITM3 Is Required for both IFITM Protein Association and Inhibition of Influenza A Virus and Dengue Virus Replication. *Journal of Virology* **87,** 7837–7852. ISSN: 0022-538X (2013).

229. Yount, J. S. *et al.* Palmitoylome profiling reveals S-palmitoylation–dependent antiviral activity of IFITM3. *Nature Chemical Biology* **6,** 610–614. ISSN: 1552-4450 (2010).

230. Chesarino, N. M., McMichael, T. M., Hach, J. C. & Yount, J. S. Phosphorylation of the Antiviral Protein Interferon-inducible Transmembrane Protein 3 (IFITM3) Dually Regulates Its Endocytosis and Ubiquitination. *Journal of Biological Chemistry* **289,** 11986–11992. ISSN: 1083351X (2014).

231. Chesarino, N. M., McMichael, T. M. & Yount, J. S. E3 Ubiquitin Ligase NEDD4 Promotes Influenza Virus Infection by Decreasing Levels of the Antiviral Protein IFITM3. *PLoS Pathogens* **11.** ISSN: 15537374 (2015).

232. Jia, R. *et al.* The C-Terminal Sequence of IFITM1 Regulates Its Anti-HIV-1 Activity. *PLOS ONE* **10** (ed Pöhlmann, S.) e0118794. ISSN: 1932-6203 (2015).

233. Yu, J. *et al.* IFITM Proteins Restrict HIV-1 Infection by Antagonizing the Envelope Glycoprotein. *Cell Reports* **13.** ISSN: 22111247 (2015).

234. Lin, T. Y. *et al.* Amphotericin B Increases Influenza A Virus Infection by Preventing IFITM3-Mediated Restriction. *Cell Reports* **5.** ISSN: 22111247. doi:10.1016/j.celrep.2013.10.033 (2013).

235. Li, K. *et al.* IFITM Proteins Restrict Viral Membrane Hemifusion. *PLoS Pathogens* **9.** ISSN: 15537366. doi:10.1371/journal.ppat.1003124 (2013).

236. Desai, T. M. *et al.* IFITM3 Restricts Influenza A Virus Entry by Blocking the Formation of Fusion Pores following Virus-Endosome Hemifusion. *PLoS Pathogens* **10.** ISSN: 15537374 (2014).

237. Wee, Y. S., Roundy, K. M., Weis, J. J. & Weis, J. H. Interferon-inducible transmembrane proteins of the innate immune response act as membrane organizers by influencing clathrin and v-ATPase localization and function. *Innate Immunity* **18,** 834–845. ISSN: 1753-4259 (2012).

238. Fu, B., Wang, L., Li, S. & Dorf, M. E. ZMPSTE24 defends against influenza and other pathogenic viruses. *The Journal of Experimental Medicine,* jem.20161270. ISSN: 0022-1007 (2017).

239. Jiang, L.-q. *et al.* IFITM3 inhibits virus-triggered induction of type I interferon by mediating autophagosome-dependent degradation of IRF3, 1–10. ISSN: 1672-7681 (2017).

240. Compton, A. A. *et al.* Natural mutations in IFITM3 modulate post-translational regulation and toggle antiviral specificity. *EMBO reports* **17,** 1657 LP –1671 (2016).

241. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74. ISSN: 0028-0836 (2015).

242. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291. ISSN: 0028-0836 (2016).

243. Woods, L. *et al.* The 100,000 Genomes Project Protocol. *Figshare* (2017).

244. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database : the journal of biological databases and curation* **2016.** ISSN: 1758-0463 (2016).

245. Zhang, Y.-H. *et al.* Interferon-induced transmembrane protein-3 genetic variant rs12252-C is associated with severe influenza in Chinese individuals. *Nature Communications* **4,** 1418. ISSN: 2041-1723 (2013).

246. Mills, T. C. *et al.* IFITM3 and susceptibility to respiratory viral infections in the community. *The Journal of infectious diseases* **209,** 1028–31. ISSN: 1537-6613 (2014).

247. López-Rodríguez, M. *et al.* IFITM3 and severe influenza virus infection. No evidence of genetic association. *European Journal of Clinical Microbiology & Infectious Diseases* **35,** 1811–1817. ISSN: 0934-9723 (2016).

248. David, S. *et al.* Population genetics of IFITM3 in Portugal and Central Africa reveals a potential modifier of influenza severity. *Immunogenetics* **70,** 169–177. ISSN: 0093-7711 (2018).

249. Randolph, A. G. *et al.* Evaluation of IFITM3 rs12252 Association With Severe Pediatric Influenza Infection. *The Journal of Infectious Diseases* **216,** 1–8. ISSN: 0022-1899 (2017).

250. Gaio, V. *et al.* Hospitalization Risk Due to Respiratory Illness Associated with Genetic Variation at IFITM3 in Patients with Influenza A(H1N1)pdm09 Infection: A Case-Control Study. *PLOS ONE* **11** (ed Wang, J.) e0158181. ISSN: 1932-6203 (2016).

251. Carter, T. C. *et al.* Pilot screening study of targeted genetic polymorphisms for association with seasonal influenza hospital admission. *Journal of Medical Virology* **90,** 436–446. ISSN: 01466615 (2018).

252. Mehrbod, P. *et al.* Association of IFITM3 rs12252 polymorphisms, BMI, diabetes, and hypercholesterolemia with mild flu in an Iranian population. *Virology journal* **14,** 218. ISSN: 1743-422X (2017).

253. Wang, Z. *et al.* Early hypercytokinemia is associated with interferon-induced transmembrane protein-3 dysfunction and predictive of fatal H7N9 infection. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 769–74. ISSN: 1091-6490 (2014).

254. Lee, N. *et al.* IFITM3, TLR3, and CD55 Gene SNPs and Cumulative Genetic Risks for Severe Outcomes in Chinese Patients With H7N9/H1N1pdm09 Influenza. *The Journal of Infectious Diseases* **216,** 97–104. ISSN: 0022-1899 (2017).

255. Pan, Y. *et al.* IFITM3 Rs12252-C Variant Increases Potential Risk for Severe Influenza Virus Infection in Chinese Population. *Frontiers in cellular and infection microbiology* **7,** 294. ISSN: 2235-2988 (2017).

256. Zhang, Y. *et al.* Interferon-induced transmembrane protein-3 rs12252-C is associated with rapid progression of acute HIV-1 infection in Chinese MSM cohort. *Aids* **29,** 889–894. ISSN: 0269-9370 (2015).

257. Xu-Yang, Z. *et al.* Interferon-Induced Transmembrane Protein 3 Inhibits Hantaan Virus Infection, and Its Single Nucleotide Polymorphism rs12252 Influences the Severity of Hemorrhagic Fever with Renal Syndrome. *Frontiers in immunology* **7,** 535. ISSN: 1664-3224 (2016).

258. XUAN, Y. *et al.* IFITM3 rs12252 T>C polymorphism is associated with the risk of severe influenza: a meta-analysis. *Epidemiology and Infection* **143,** 2975–2984. ISSN: 0950-2688 (2015).

259. Prabhu, S. S., Chakraborty, T. T., Kumar, N. & Banerjee, I. Association between IFITM3 rs12252 polymorphism and influenza susceptibility and severity: A meta-analysis. *Gene* **674,** 70–79. ISSN: 0378-1119 (2018).

260. Qin, L. *et al.* High Level Antibody Response to Pandemic Influenza H1N1/09 Virus Is Associated With Interferon-Induced Transmembrane Protein-3 rs12252-CC in Young Adults. *Frontiers in cellular and infection microbiology* **8,** 134. ISSN: 2235-2988 (2018).

261. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19. ISSN: 1097-4172 (2016).

262. Williams, D. E. J. *et al.* IFITM3 polymorphism rs12252-C restricts influenza A viruses. *PLoS ONE* **9.** ISSN: 19326203 (2014).

263. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760. ISSN: 1367-4803 (2009).

264. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14,** 178–92. ISSN: 1477-4054 (2013).

265. Maruyama, T. *et al.* Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nature Biotechnology* **33,** 538–542. ISSN: 1087-0156 (2015).

266. Song, J. *et al.* RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. *Nature Communications* **7,** 10548. ISSN: 2041-1723 (2016).

267. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Research* **42,** e168–e168. ISSN: 1362-4962 (2014).

268. Powell, T. J., Silk, J. D., Sharps, J., Fodor, E. & Townsend, A. R. M. Pseudotyped Influenza A Virus as a Vaccine for the Induction of Heterotypic Immunity. *Journal of Virology* **86,** 13397–13406. ISSN: 0022-538X (2012).

269. Martín, J *et al.* Studies of the binding properties of influenza hemagglutinin receptor-site mutants. *Virology* **241,** 101–11. ISSN: 0042-6822 (1998).

270. De Vries, R. P. *et al.* A single mutation in Taiwanese H6N1 influenza hemagglutinin switches binding to human-type receptors. *EMBO molecular medicine* **9,** 1314–1325. ISSN: 1757-4684 (2017).

271. Chen, Z. *et al.* Generation of live attenuated novel influenza virus A/California/7/09 (H1N1) vaccines with high yield in embryonated chicken eggs. *Journal of virology* **84,** 44–51. ISSN: 1098-5514 (2010).

272. R Core Team. *: A language and environment for statistical computing.* 2016. `https://www.r-project.org/` (2017).

273. Andrews S. *FastQC: a quality control tool for high throughput sequence data* `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/` (2017).

274. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10. ISSN: 2226-6089 (2011).

275. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21. ISSN: 1460-2059 (2013).

276. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Research* **44,** D710–D716. ISSN: 0305-1048 (2016).

277. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930. ISSN: 1367-4803 (2014).

278. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079. ISSN: 1367-4803 (2009).

279. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7,** 562–578. ISSN: 1754-2189 (2012).

280. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550. ISSN: 1474-760X (2014).

281. Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011,** bar030–bar030. ISSN: 1758-0463 (2011).

282. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-0-387-98140-6. `http://ggplot2.org` (Springer-Verlag New York, 2009).

283. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25,** 25–29. ISSN: 1061-4036 (2000).

284. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45,** D331–D338. ISSN: 0305-1048 (2017).

285. Alexa, A., Rg, J., Hrer, R. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. **22,** 1600–1607 (2006).

286. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358. ISSN: 1367-4803 (2012).

287. Jia, R. *et al.* Identification of an endocytic signal essential for the antiviral action of IFITM3. *Cellular microbiology* **16,** 1080–93. ISSN: 1462-5822 (2014).

288. Green, E. D., Guyer, M. S. & National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470,** 204–213. ISSN: 0028-0836 (2011).

289. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Research* **37,** e67–e67. ISSN: 1362-4962 (2009).

290. Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7,** 1009–1015. ISSN: 1548-7091 (2010).

291. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics (Oxford, England)* **31,** 3555–7. ISSN: 1367-4811 (2015).

292. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 0028-0836 (2012).

293. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome research* **12,** 996–1006. ISSN: 1088-9051 (2002).

294. Stave, J. W. & Lindpaintner, K. Antibody and antigen contact residues define epitope and paratope size and structure. *Journal of immunology (Baltimore, Md. : 1950)* **191,** 1428–35. ISSN: 1550-6606 (2013).

295. Pang, J., Zeng, X., Xiao, R.-p., Lakatta, E. G. & Lin, L. Design, generation, and testing of mammalian expression modules that tag membrane proteins. *Protein science : a publication of the Protein Society* **18,** 1261–71. ISSN: 1469-896X (2009).

296. Liu, Y. *et al.* Impact of Alternative Splicing on the Human Proteome. *Cell Reports* **20,** 1229–1241 (2017).

297. Piggin, C. L. *et al.* ELF5 isoform expression is tissue-specific and significantly altered in cancer. *Breast cancer research : BCR* **18,** 4. ISSN: 1465-542X (2016).

298. Shi, M. *et al.* Premature termination codons are recognized in the nucleus in a reading-frame-dependent manner. *Cell Discovery* **1,** 15001. ISSN: 2056-5968 (2015).

299. Soergel, D. A., Lareau, L. F. & Brenner, S. E. Regulation of Gene Expression by Coupling of Alternative Splicing and NMD. `https://www.ncbi.nlm.nih.gov/books/NBK6088/` (2013).

300. Anderson, C. L., Zundel, M. A. & Werner, R. Variable promoter usage and alternative splicing in five mouse connexin genes. *Genomics* **85,** 238–244. ISSN: 08887543 (2005).

301. Araujo, P. R. *et al.* Before It Gets Started: Regulating Translation at the 5′ UTR. *Comparative and Functional Genomics* **2012,** 1–8. ISSN: 1531-6912 (2012).

302. Wu, W.-L. *et al.* Δ20 IFITM2 differentially restricts X4 and R5 HIV-1. *Proceedings of the National Academy of Sciences of the United States of America* **114,** 7112–7117. ISSN: 1091-6490 (2017).

303. Foster, T. L. *et al.* Resistance of Transmitted Founder HIV-1 to IFITM-Mediated Restriction. *Cell Host and Microbe* **20,** 429–442. ISSN: 19346069 (2016).

304. Brubaker, S. W., Gauthier, A. E., Mills, E. W., Ingolia, N. T. & Kagan, J. C. A bicistronic MAVS transcript highlights a class of truncated variants in antiviral immunity. *Cell* **156,** 800–11. ISSN: 1097-4172 (2014).

305. Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Human molecular genetics* **24,** R111–9. ISSN: 1460-2083 (2015).

306. Brodie, A., Azaria, J. R. & Ofran, Y. How far from the SNP may the causative genes be? *Nucleic Acids Research* **44,** 6046–6054. ISSN: 0305-1048 (2016).

307. Wood, A. R. *et al.* Imputation of Variants from the 1000 Genomes Project Modestly Improves Known Associations and Can Identify Low-frequency Variant - Phenotype Associations Undetected by HapMap Based Imputation. *PLoS ONE* **8** (ed Arking, D. E.) e64343. ISSN: 1932-6203 (2013).

308. Makvandi-Nejad, S. *et al.* Lack of Truncated IFITM3 Transcripts in Cells Homozygous for the rs12252-C Variant That is Associated With Severe Influenza Infection. *The Journal of Infectious Diseases* **217,** 257–262. ISSN: 0022-1899 (2018).

309. Rebbeck, T. R., Spitz, M. & Wu, X. Assessing the function of genetic variants in candidate gene association studies. *Nature Reviews Genetics* **5,** 589–597. ISSN: 1471-0056 (2004).

310. Breker, M. & Schuldiner, M. The emergence of proteome-wide technologies: systematic analysis of proteins comes of age. *Nature Reviews Molecular Cell Biology* **15,** 453–464. ISSN: 1471-0072 (2014).

311. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368,** 20120362. ISSN: 1471-2970 (2013).

312. Rädle, B. *et al.* Metabolic Labeling of Newly Transcribed RNA for High Resolution Gene Expression Profiling of RNA Synthesis, Processing and Decay in Cell Culture. *Journal of Visualized Experiments,* 1–11. ISSN: 1940-087X (2013).

313. Burger, K. *et al.* 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biology* **10,** 1623–1630. ISSN: 1547-6286 (2013).

314. Kuwano, Y., Rabinovic, A., Srikantan, S., Gorospe, M. & Demple, B. Analysis of Nitric Oxide-Stabilized mRNAs in Human Fibroblasts Reveals HuR-Dependent Heme Oxygenase 1 Upregulation. *Molecular and Cellular Biology* **29,** 2622–2635. ISSN: 0270-7306 (2009).

315. Li, G. *et al.* Small molecules enhance CRISPR/Cas9-mediated homology-directed genome editing in primary cells. *Scientific Reports* **7,** 8943. ISSN: 2045-2322 (2017).

316. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22,** 1760–1774. ISSN: 1088-9051 (2012).

317. Chen, J.-H., Ozanne, S. E. & Hales, C. N. in *Methods in molecular biology (Clifton, N.J.)* 179–189 (2007). doi:10.1007/978-1-59745-361-5_14. http://www.ncbi.nlm.nih.gov/pubmed/17634582http://link.springer.com/10.1007/978-1-59745-361-5{_}14.

318. Yu, Q. *et al.* DNA-Damage-Induced Type I Interferon Promotes Senescence and Inhibits Stem Cell Function. *Cell Reports* **11,** 785–797. ISSN: 2211-1247 (2015).

319. Yang, H., Wang, H., Ren, J., Chen, Q. & Chen, Z. J. cGAS is essential for cellular senescence. *Proceedings of the National Academy of Sciences* **114,** E4612–E4620. ISSN: 0027-8424 (2017).

320. Wagner, J. R. *et al.* Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human. *PLoS Computational Biology* **6** (ed Wasserman, W. W.) e1000849. ISSN: 1553-7358 (2010).

321. Milani, L. *et al.* Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells. *Nucleic acids research* **35,** e34. ISSN: 1362-4962 (2007).

322. Serre, D. *et al.* Differential Allelic Expression in the Human Genome: A Robust Approach To Identify Genetic and Epigenetic Cis-Acting Mechanisms Regulating Gene Expression. *PLoS Genetics* **4** (ed Dermitzakis, E. T.) e1000006. ISSN: 1553-7404 (2008).

323. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology* **33,** 187–197. ISSN: 1087-0156 (2015).

324. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551,** 464–471. ISSN: 0028-0836 (2017).

325. Vink, J. M. *et al.* Differential gene expression patterns between smokers and non-smokers: cause or consequence? *Addiction Biology* **22,** 550–560. ISSN: 13556215 (2017).

326. Wu, C., MacLeod, I. & Su, A. I. BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Research* **41,** D561–D565. ISSN: 0305-1048 (2013).

327. Wakim, L. M., Gupta, N., Mintern, J. D. & Villadangos, J. A. Enhanced survival of lung tissue-resident memory CD8+ T cells during infection with influenza virus due to selective expression of IFITM3. *Nature Immunology* **14,** 238–245. ISSN: 1529-2908 (2013).

328. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660. ISSN: 0036-8075 (2015).

329. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS genetics* **8,** e1002431. ISSN: 1553-7404 (2012).

330. Hauberg, M. E. *et al.* Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression. *The American Journal of Human Genetics* **100,** 885–894. ISSN: 0002-9297 (2017).

331. Tats, A., Tenson, T. & Remm, M. Preferred and avoided codon pairs in three domains of life. *BMC Genomics* **9,** 463. ISSN: 1471-2164 (2008).

332. Speth, C & Oberbäumer, I. Expression of Basement Membrane Proteins: Evidence for Complex Post-transcriptional Control Mechanisms. *Experimental Cell Research* **204,** 302–310. ISSN: 00144827 (1993).

333. Chen, C. Y., Xu, N & Shyu, A. B. mRNA decay mediated by two distinct AU-rich elements from c-fos and granulocyte-macrophage colony-stimulating factor transcripts: different deadenylation kinetics and uncoupling from translation. *Molecular and cellular biology* **15,** 5777–88. ISSN: 0270-7306 (1995).

334. Chen, C.-Y. A., Ezzeddine, N. & Shyu, A.-B. Messenger RNA half-life measurements in mammalian cells. *Methods in enzymology* **448,** 335–57. ISSN: 1557-7988 (2008).

335. Jao, C. Y. & Salic, A. Exploring RNA transcription and turnover in vivo by using click chemistry. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 15779–84. ISSN: 1091-6490 (2008).

336. Brenet, F. *et al.* DNA methylation of the first exon is tightly linked to transcriptional silencing. *PloS one* **6,** e14524. ISSN: 1932-6203 (2011).

337. Edwards, J. R. *et al.* Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome research* **20,** 972–80. ISSN: 1549-5469 (2010).

338. Tuller, T. & Zur, H. Multiple roles of the coding sequence 5′ end in gene expression regulation. *Nucleic Acids Research* **43,** 13–28. ISSN: 1362-4962 (2015).

339. Capon, F. *et al.* A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Human Molecular Genetics* **13,** 2361–2368. ISSN: 09646906 (2004).

340. Birnbaum, R. Y. *et al.* Coding exons function as tissue-specific enhancers of nearby genes. *Genome research* **22,** 1059–68. ISSN: 1549-5469 (2012).

341. Ritter, D. I., Dong, Z., Guo, S. & Chuang, J. H. Transcriptional Enhancers in Protein-Coding Exons of Vertebrate Developmental Genes. *PLoS ONE* **7** (ed Schubert, M.) e35202. ISSN: 1932-6203 (2012).

342. Tumpel, S., Cambronero, F., Sims, C., Krumlauf, R. & Wiedemann, L. M. A regulatory module embedded in the coding region of Hoxa2 controls expression in rhombomere 2. *Proceedings of the National Academy of Sciences* **105,** 20077–20082. ISSN: 0027-8424 (2008).

343. Stergachis, A. B. *et al.* Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution. *Science* **342,** 1367–1372. ISSN: 0036-8075 (2013).

344. Sha, Y., Phan, J. H. & Wang, M. D. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* **2015,** 6461–4. ISSN: 1557-170X (2015).

345. Hsieh, W.-L. *et al.* IFI27, a novel epidermal growth factor-stabilized protein, is functionally involved in proliferation and cell cycling of human epidermal keratinocytes. *Cell Proliferation* **48,** 187–197. ISSN: 09607722 (2015).

346. Purbey, P. K. *et al.* Defined Sensing Mechanisms and Signaling Pathways Contribute to the Global Inflammatory Gene Expression Output Elicited by Ionizing Radiation. *Immunity* **47,** 421–434.e3. ISSN: 1074-7613 (2017).

347. Seng, L.-G., Daly, J., Chang, K.-C. & Kuchipudi, S. V. High Basal Expression of Interferon-Stimulated Genes in Human Bronchial Epithelial (BEAS-2B) Cells Contributes to Influenza A Virus Resistance. *PLoS ONE* **9** (ed Nevels, M.) e109023. ISSN: 1932-6203 (2014).

348. Aamer, H. A., Rajakumar, P., Nyaundi, J. & Murphey-Corb, M. Resistance to simian immunodeficiency virus low dose rectal challenge is associated with higher constitutive TRIM5α expression in PBMC. *Retrovirology* **11,** 39. ISSN: 1742-4690 (2014).

349. Takaoka, A *et al.* Cross talk between interferon-gamma and -alpha/beta signaling components in caveolar membrane domains. *Science (New York, N.Y.)* **288,** 2357–60. ISSN: 0036-8075 (2000).

350. Erlandsson, L *et al.* Interferon-beta is required for interferon-alpha production in mouse fibroblasts. *Current biology : CB* **8,** 223–6. ISSN: 0960-9822 (1998).

351. Jablonska, J., Leschner, S., Westphal, K., Lienenklaus, S. & Weiss, S. Neutrophils responsive to endogenous IFN-β regulate tumor angiogenesis and growth in a mouse tumor model. *Journal of Clinical Investigation* **120,** 1151–1164. ISSN: 0021-9738 (2010).

352. Arakura, F. *et al.* Genetic control directed toward spontaneous IFN-alpha/IFN-beta responses and downstream IFN-gamma expression influences the pathogenesis of a murine psoriasis-like skin disease. *Journal of immunology (Baltimore, Md. : 1950)* **179,** 3249–57. ISSN: 0022-1767 (2007).

353. Riemer, A. B., Keskin, D. B. & Reinherz, E. L. Identification and validation of reference genes for expression studies in human keratinocyte cell lines treated with and without interferon-γ - a method for qRT-PCR reference gene determination. *Experimental dermatology* **21,** 625–9. ISSN: 1600-0625 (2012).

354. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485,** 381–385. ISSN: 0028-0836 (2012).

355. The Ensembl Group. *TSL Definitions (Ensembl)* `https://www.ensembl.org/Help/Glossary?id=492` (2018).

356. Strausberg, R. L. *et al.* Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proceedings of the National Academy of Sciences* **99,** 16899–16903. ISSN: 0027-8424 (2002).

357. Kornienko, A. E. *et al.* Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biology* **17,** 14. ISSN: 1474-760X (2016).

358. Pertea, M. The human transcriptome: an unfinished story. *Genes* **3,** 344–60. ISSN: 2073-4425 (2012).

359. Crampton, N., Bonass, W. A., Kirkham, J., Rivetti, C. & Thomson, N. H. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Research* **34,** 5416–5425. ISSN: 0305-1048 (2006).

360. Hobson, D. J., Wei, W., Steinmetz, L. M. & Svejstrup, J. Q. RNA polymerase II collision interrupts convergent transcription. *Molecular cell* **48,** 365–74. ISSN: 1097-4164 (2012).

361. Yao, C. *et al.* Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *The American Journal of Human Genetics* **100,** 571–580. ISSN: 0002-9297 (2017).

362. Marigorta, U. M. *et al.* Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nature Genetics* **49,** 1517–1521. ISSN: 1061-4036 (2017).

363. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 0028-0836 (2012).

364. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98,** 1487–9. ISSN: 1592-8721 (2013).

365. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330. ISSN: 0028-0836 (2015).

366. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome biology* **16,** 56. ISSN: 1474-760X (2015).

367. Van Nas, A. *et al.* Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics* **185,** 1059–68. ISSN: 1943-2631 (2010).

368. Infusini, G. *et al.* Respiratory DC Use IFITM3 to Avoid Direct Viral Infection and Safeguard Virus-Specific CD8+ T Cell Priming. *PLOS ONE* **10** (ed Tregoning, J. S.) e0143539. ISSN: 1932-6203 (2015).

369. Burgess, S. & Thompson, S. G. Use of allele scores as instrumental variables for Mendelian randomization. *International journal of epidemiology* **42,** 1134–44. ISSN: 1464-3685 (2013).

370. Ciancanelli, M. J., Abel, L., Zhang, S.-Y. & Casanova, J.-L. Host genetics of severe influenza: from mouse Mx1 to human IRF7. *Current opinion in immunology* **38,** 109–20. ISSN: 1879-0372 (2016).

371. Ueno, S *et al.* PU.1 induces apoptosis in myeloma cells through direct transactivation of TRAIL. *Oncogene* **28,** 4116–25. ISSN: 1476-5594 (2009).

372. Bailey, C. C., Huang, I.-C., Kam, C. & Farzan, M. Ifitm3 Limits the Severity of Acute Influenza in Mice. *PLoS Pathogens* **8** (ed Basler, C. F.) e1002909. ISSN: 1553-7374 (2012).

373. Rusinova, I. *et al.* INTERFEROME v2.0: an updated database of annotated interferon-regulated genes. *Nucleic Acids Research* **41,** D1040–D1046. ISSN: 0305-1048 (2012).

374. Parnell, G. P. *et al.* A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia. *Critical Care* **16,** R157. ISSN: 1364-8535 (2012).

375. Dunning, J. *et al.* Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza. *Nature Immunology* **19,** 625–635. ISSN: 1529-2908 (2018).

376. Levin, D. *et al.* Multifaceted activities of type I interferon are revealed by a receptor antagonist. *Science signaling* **7,** ra50. ISSN: 1937-9145 (2014).

377. Baccam, P., Beauchemin, C., Macken, C. A., Hayden, F. G. & Perelson, A. S. Kinetics of influenza A virus infection in humans. *Journal of virology* **80,** 7590–9. ISSN: 0022-538X (2006).

378. Witte, K *et al.* Despite IFN-λ receptor expression, blood immune cells, but not keratinocytes or melanocytes, have an impaired response to type III interferons: implications for therapeutic applications of these cytokines. *Genes & Immunity* **10,** 702–714. ISSN: 1466-4879 (2009).

379. Kiefer, K., Oropallo, M. A., Cancro, M. P. & Marshak-Rothstein, A. Role of type I interferons in the activation of autoreactive B cells. *Immunology and Cell Biology* **90,** 498–504. ISSN: 08189641 (2012).

380. Davey, R. T. *et al.* The association between serum biomarkers and disease outcome in influenza A(H1N1)pdm09 virus infection: results of two international observational cohort studies. *PloS one* **8,** e57121. ISSN: 1932-6203 (2013).

381. Mikkelsen, S. S. *et al.* RIG-I-mediated activation of p38 MAPK is essential for viral induction of interferon and activation of dendritic cells: dependence on TRAF2 and TAK1. *The Journal of biological chemistry* **284,** 10774–82. ISSN: 0021-9258 (2009).

382. Hockemeyer, D. & Jaenisch, R. Induced Pluripotent Stem Cells Meet Genome Editing. *Cell stem cell* **18,** 573–86. ISSN: 1875-9777 (2016).

383. Huang, X. *et al.* Precise Determination of Time to Reach Viral Load Set Point After Acute HIV-1 Infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **61,** 448–454. ISSN: 1525-4135 (2012).

384. GTEx Consortium, T. G. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45,** 580–5. ISSN: 1546-1718 (2013).

385. Smedley, D. *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research* **43,** W589–W598. ISSN: 0305-1048 (2015).

Appendix A

Infection Cohorts used for Analysis

A.1 You'An Hospital Pandemic H1N1 Cohort

Our *in-vivo* studies of IAV infection were facilitated by the use of a patient cohort from You'An Hospital in Beijing, China. This hospital is affiliated with Capital Medical University, and is located in the Fengtai Disctrict in Beijing. You'an hospital specialises in the treatment of infectious diseases. This cohort formed the basis of the Zhang *et al* study investigating the effects of rs12252 on severe influenza virus infection [245].

Patients used in this cohort were admitted to hospital with influenza infection. Over the course of the 2009 H1N1 pandemic, patients with mild as well as severe infection were admitted to hospital during this period. This was an unusual practice relative to other hospitals in China, but allowed for a unique perspective into patients with mild influenza infection. All patients included within this cohort were confirmed for pandemic H1N1 infection via PCR and gave informed consent for inclusion into the study. Patients that had taken immunosuppressants within 1 month of the study were excluded, alongside any patients with confounding illness or pregnancy. This included patients with positive bacterial infections during hospitalisation consistent with secondary infection.

At the time of discharge or death, patients were sorted into mild and severe infection categories via a set of clinical criteria. Mild patients developed symptoms typical of influenza infection, including fever, headache and coughing. Leukocyte profile of these patients remained within the normal range. Severe patients also presented with the above, but additionally deteriorated to develop more severe clinical symptoms. Symptoms of severe patients included:

- Pneumatic consolidation
- Bilateral infiltrates (via chest radiograph)
- Respiratory Failure
- Disseminated intravascular coagulopathy
- Renal Failure

All patients in the mild group survived infection. 19/32, or 59% of severe patients survived. All severe patients were treated with oseltamivir, as opposed to only 38/51 of the mild patients. Blood samples used for analysis were collected 2-3 days after admission to hospital.

Studies using this cohort were approved by the local ethics committee of Beijing You'An Hospital, and the University of Oxford Tropical Ethics Committee, under OXTREC 1001-13.

A.2 You'An Hospital Acute HIV-1 Cohort

We used an acute HIV-1 infection cohort to analyse the splicing of *IFITM3* with respect to rs12252. This cohort was collected and shared by collaborators at Beijing You'An Hospital, as the 2009 H1N1pdm09 cohort explained in Appendix A.1. This cohort was used in the 2015 Zhang *et al* study, and is described in more detail a 2012 Huang *et al* study [256, 383]. The cohort consists of men who have sex with men from Beijing, China who were enrolled in a longitudinal study. These men were healthy at the time of recruitment, but were checked every 2 months for anti-HIV Abs. In cases where active HIV infection was detected, patients were monitored by

collection of whole blood and PBMC samples at regular intervals. None of the subjects were treated by antiretroviral therapy.

Studies using this cohort were approved by the local ethics committee of Beijing You'An Hospital, and the University of Oxford Tropical Ethics Committee, under OXTREC 16-10.

Appendix B

Primer and Guide RNA Sequences

Primers were all ordered as anhydrous oligonucleotides. Upon receipt, primers were dissolved to a concentration of 100µM. All primers used are shown in the table below:

Table B.1 sgRNA Sequences

Primer Name	**Sequence(5' to 3')**
*IFITM2*_E1_F	CTGGCCAGCTCTGCATTTGA
*IFITM2*_E1_R	CAGCATAGCCACTTCCTGCT
*IFITM3*_E1_F	GGCTCACTGAGAACCATCCC
*IFITM3*_E1_R	CACCCTCTGAGCATTCCCTG
*IFITM3*_HDR_F	GGTCTTCGCTGGACACCATG
*IFITM3*_HDR_R	GCAGCAGCACCAGAAACACG
*IFITM3*_E2_F	CAGGAAGATGGTTGGCGACG
*IFITM3*_E2_R	ACCTGCTCAGTCTCAGAACCACG
Meth_Direct_R	CCCTCTAAAACATTCCCTAAAACCATAC
rs12252_MiSeq_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GTCACACTGTCCAAACCTTCTTCT
rs12252_MiSeq_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGTTCACGGAGTAGGCGAATGC

All guide RNA sequences used are shown in the table below:

Table B.2 sgRNA Sequences

Guide RNA Name	**Sequence (5' to 3')**
sg3	CACCCTCTGAGCATTCCCTGGGG
sg23	GTCAAATCGGGATTTGGCCGGGG
sg25	AGGCCTGGAAGATCAGCACTGGG
sg_HDR_v2	GGGGGCTGGCCACTGTTGACAGG

Appendix C

Biological Databases Used for Analysis

Over the course of our *IFITM3* studies, we used a variety of different biological databases to obtain information. These databases are all widely available to the public at large.

C.1 1000 Genomes Database

The 1000 genomes databases is a catalogue of human genetic variants developed by the 1000 genomes project consortium [241]. This project, in various phases, catalogued and identified human genetic variants throughout different world populations. In total, 2504 individual samples underwent gDNA sequencing to varying depths - these samples were subdivided into population backgrounds. These backgrounds were derived from self-reporting. All patients were reported as healthy, although no phenotype information is available for this dataset.

All genotyped individuals were divided into general world populations. These world populations were classified as:

- African
- American (AMR)

- East Asian
- European
- South Asian

Of particular interest to us in our current studies are the EAS and EUR world populations at large. These populations are subdivided into smaller populations, generally sourced from individuals within a single city - this provides a more focused cohort. The subpopulations of EAS are:

- Chinese Dai in Xishuangbanna China
- Han Chinese in Beijing China
- Southern Han Chinese China
- Japanese in Tokyo Japan
- Kinh in Ho Chi Minh City Vietnam

As a reference, the CHB subpopulation was considered the most similar to the populations in cohorts used in our work. The subpopulations within EUR are as follows:

- Utah Residents with Northern and Western European Ancestry
- Finnish in Finland
- British in England and Scotland
- Iberian populations in Spain
- Toscani in Italy

This resource has allowed us to examine population level frequencies of genetic variants, as well as perform LD analysis using this dataset.

C.2 ExAC Database

The Exome Aggregation Consortium Database, or ExAC database, is a large scale aggregation of exome sequencing data derived from a wide range of smaller databases [242]. This database aggregates samples from 60,706 individual samples. Due to differential levels of consent obtained in the smaller databases that make up ExAC, there is no phenotype data available for obtainable samples. In addition, the data is not specific to tissue, meaning the aggregate data is averaged across the tissues available in the database. However, there are no samples derived from tumour tissue - these could potentially cause false reporting of variants derived from tumour specific genetic differentiation.

This database makes available exon coverage data for annotated human transcripts and genes. More importantly, the database offers a wealth of information regarding exome SNPs. Due to the extremely large number of individual samples, this database is powerful enough to detect extremely rare exome genetic variants. In addition, the database is derived from a wide range of world populations. This allows for more accurate reporting of allele frequencies of exome variants than the frequencies reported in 1000 genomes. However, the world populations used are not subdivided beyond their greater world population (e.g. EUR, EAS). Of use to our studies is the fact that this database contains 4,328 samples from individuals of EAS descent, meaning rs12252 is well reported in this database.

In addition, this database shows a degree of raw read alignment to genomes at variant sites - this was useful to us in obtaining data related to allelic expression.

C.3 GTEx Database

The Genotype Tissue Expression project or GTEx is a large scale database of tissue specific gene expression and genotype level data [384]. This database is based on rapid autopsy of postmortem donors. All donors were reviewed to ensure the donors were clear of cancer and infectious disease [215]. These donors were used for DNA and RNA sequencing - the RNA sequencing was performed on a tissue by tissue basis. This database currently contains 714 donors total, with 620 donors having genotype and RNA-Sequencing

data. The combination of RNA and DNA sequencing allowed the authors to perform eQTL analysis - this database provides a powerful resource in investigating potential eQTLs. In its current state, 85.2% of all GTEx donors are reported as Caucasian. Unfortunately, this fact limits our studies and to some extent, its power as an eQTL database in discovering eQTLs that are associated with variants more common in other world populations.

In addition, this database can be used to examine tissue specific expression levels of all genes queried, providing a useful reference for the tissue specific expression of genes of interest. The quantity of samples available in this database is highly variable on a tissue by tissue basis, but is generally well informed with regards to tissues we are interested in. To name a few, there eQTL analysis has been performed on 250-500 samples in muscle, whole blood, lung and transformed fibroblasts.

C.4 ensembl Database

The ensembl Database is an web-based resource aggregating data related to the human genome [276]. Included in this is SNP data, genome sequences, and genome annotations. Gene annotations are generally acquired through the GENCODE and HAVANA sister projects. SNP data is largely derived from 1000 genomes. The ensembl resource provides reference genome sequences and genome annotations. We used reference genomes and annotations from the GRCh38.87 version of the reference genome, as assembled by ensembl.

In addition, we used the ensembl BioMart resource - this resource aggregates a range of data from different research disciplines. Importantly, this database allows for conversion between gene and transcript IDs obtained from different resources [385].

C.4.1 Interferome Database

We used the Interferome database to perform in-depth investigations of observed IFN signatures [373]. The Interferome database is a curated database of genes classified as IRGs. This data is managed by the Paul Herzog group via Monash University. Data used to populate the

database was obtained via corroboration of publicly available microarray and RNA-Sequencing data. For inclusion into the database, an IRG must be differentially expressed after IFN treatment, with a minimum fold change of 2, and a maximum p-value of 0.05. Information regarding IRG regulation by IFN type and tissue is also made available in this database.

Appendix D

Quality Control of RNA-Seq Cohorts

D.1 Quality Control of Primary Human Fibroblasts

Each sample library was prepared with 100bp paired-end reads, with roughly 25 million reads per sample. When samples were returned, we performed analysis of quality using FASTQC [273]. RNA samples showed extremely high average quality base scores in both forward and reverse reads (fig. D.1). As expected, the average read quality declined at the end of the reads.

The sequencing data passed all relevant FastQC quality checks. We found no evidence of contaminating sequences in the returned samples. With these results, we had no concerns about data quality moving forward.

D.1.1 Quality Control of PBMC and DC Datasets

RNA samples from IAV+ DCs were sent to the Wellcome Trust Centre for Human Genetics. cDNA libraries were prepared with paired-end reads of 100bp in length. Libraries were sequenced on a HiSeq 2000 platform. Upon return of samples, quality was checked with FastQC [273]. Resultant samples had consistently high quality scores across read length, with predictable

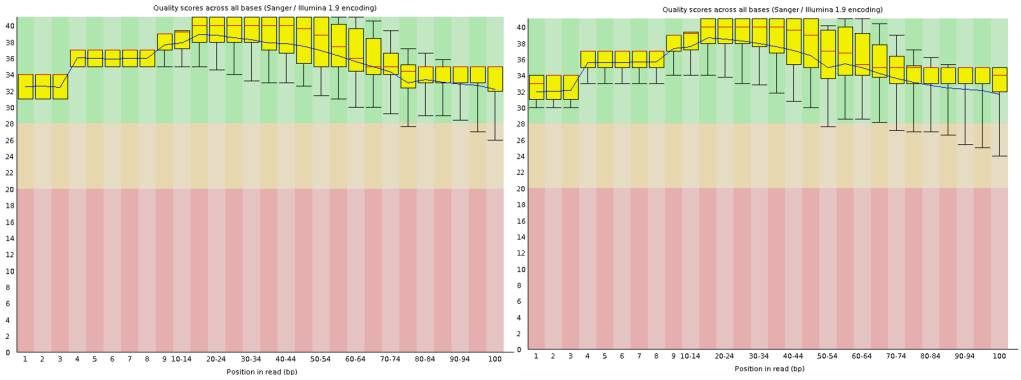
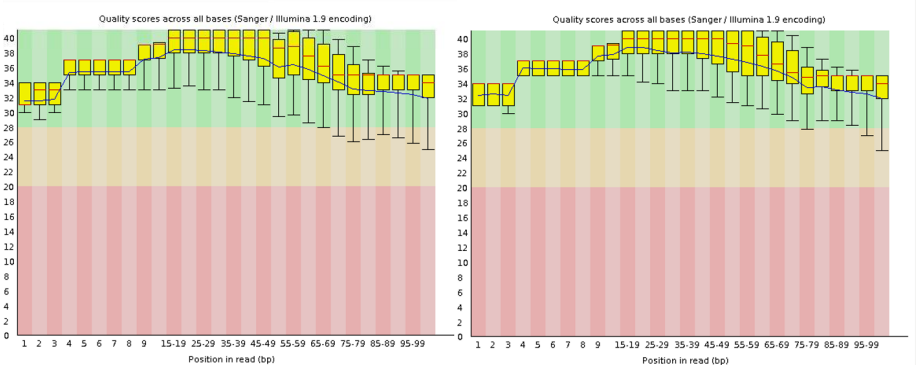


Figure D.1: Quality scores across all bases for an example forward (left) and reverse (read) file. Y-axis displays Illumina Phred quality scores - this can be treated as logarithmic confidence in base accuracy. Average read quality scores were extremely high

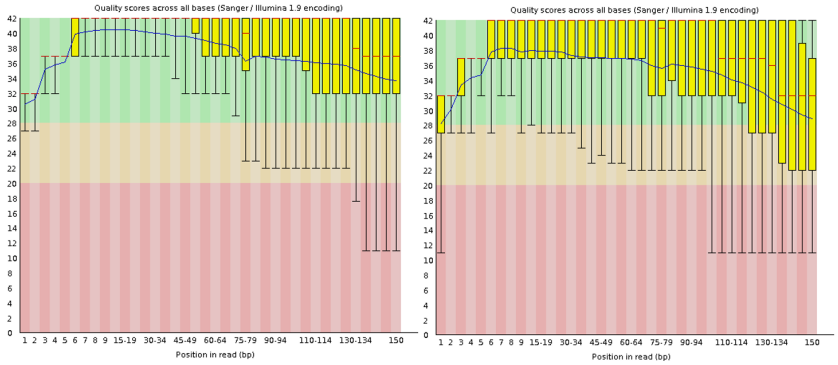
deterioration of sequence quality near the end of the read (fig. D.2a).

PBMCs from 2009 pH1N1 patients and HIV patients were derived from cells frozen in Freezing Medium. Cells were thawed before RNA extraction. RNA samples were submitted to BGI in Beijing, China for sequencing. RNA-Sequencing was performed on a HiSeq 4000 platform. Reads were prepared using standard Illumina protocols, giving 100bp, paired end reads, with 50-60 million reads per sample. Upon obtaining sequencing data from BGI, we performed quality control using FastQC. RNA samples showed high average quality base scores in both forward and reverse reads (fig. D.2b and fig. D.2c). These sequences were not as high in quality as with other RNA-Sequencing cohorts examined - this is likely a result of the fact that RNA-Specific freezing was not used.

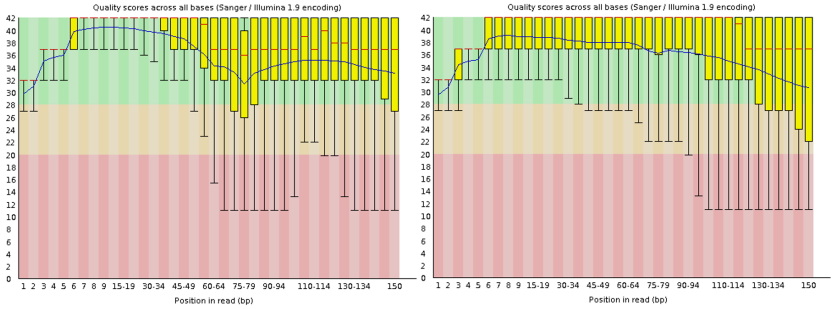
In PBMC samples, we observed contaimination of samples with illumina specific primers. To remove sources of contaminating sequences, we used cutadapt to remove the aforementioned sequences, and remove all resultant empty reads [274].



(a) IAV+ DC Samples



(b) 2009 pH1N1 Patient PBMC Samples



(c) HIV Patient PBMC Samples

Figure D.2: ***a.*** *Quality scores across all bases for an example forward (left) and reverse (read) file. Y-axis displays Illumina Phred quality scores - this can be treated as logarithmic confidence in base accuracy. Average read quality scores were consistently high* ***b & c.*** *PBMC RNA quality scores across all bases for forward (left) and reverse (right) reads. Average read quality scores were relatively high.*

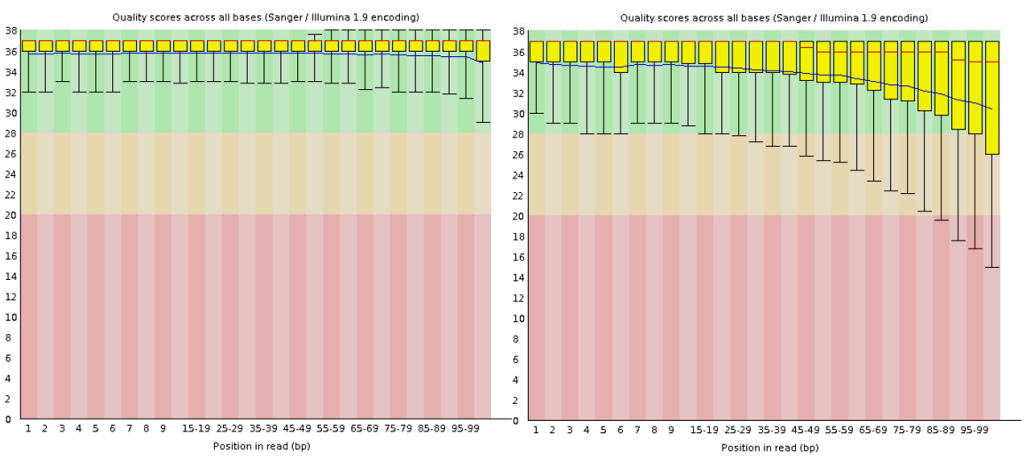


Figure D.3: Quality scores across all bases for an example forward (left) and reverse (right) file. Y-axis displays Illumina Phred quality scores, or logarithmic confidence in base accuracy. Average read quality scores were extremely high

D.2 Quality Control of Healthy Donor Cohort

Library quality was analysed by FastQC[273]. RNA samples showed extremely high average quality base scores in both forward and reverse reads (fig. D.3). As expected, the quality of forward and reverse reads deteriorated near the ends of the reads.

We observed a number of overrepresented sequences. These sequences were not identified by FastQC. We used NCBI nucleotide BLAST to identify these sequences - upon performing alignments, these sequences showed 100% alignment to human alpha globin mRNA. This alignment is not unexpected, given the fact that the RNA used in this experiment is derived from whole blood containing erythrocyte RNA. We did not feel as if the overrepresentation of these sequences would impact our experiment at large.

Given these metrics, we had high confidence in the quality of this dataset moving forward.

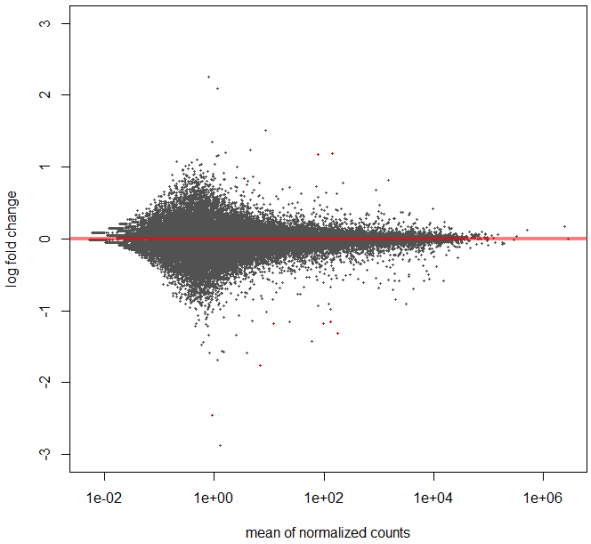
Appendix E

Additional Analysis of Healthy Donor Cohort

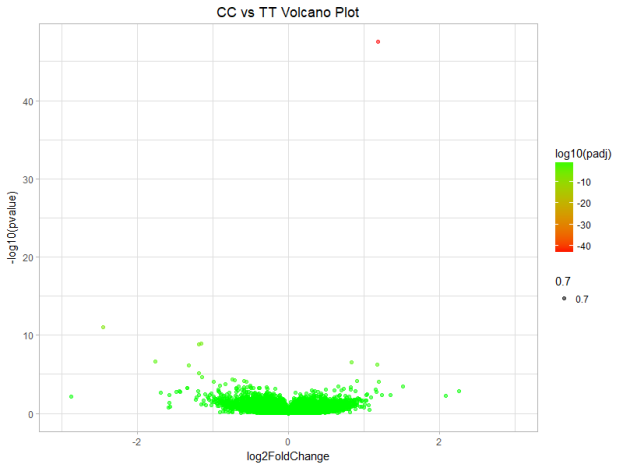
We plotted MA and Volcano plots to visualise the overall expression trends between CC and TT donors (fig. E.1). Since these plots serve to show expression differences between two sample groups, we used homozygous patients only. We were forced to plot the volcano plot using the unadjusted p-value, as the multiple test correction flattened the data significantly, not allowing for visualisation of any clear data trends. In both of these plots, there appeared to be low differential expression genes in general, with a small number of genes upregulated in both directions.

Before analysis of expression data via PCA, we determined the best method offered by DESeq2 for transforming normalised count data. The three methods offered by DESeq2 are a normal $\log_2$ transformation, a method known as regularised log transformation, and a variance stabilising transformation. We performed all three transformations on our results data, and plotted the mean counts for each gene against their standard deviation. We found that the variance stabilising transformation most effectively controlled standard deviation across all levels of expression (fig. E.2).

PCA was performed on CC/TT donors, and all donors. In both cases, PCA plots showed limited dispersion, with low levels of variance between all samples in general. A clear separation of two populations of donors was visible on the PC1 axis (figs. E.3a and E.3b) in both plots. To investigate the underlying gene expression factors governing the separation of donors, we performed differential gene expression analysis based on each donors

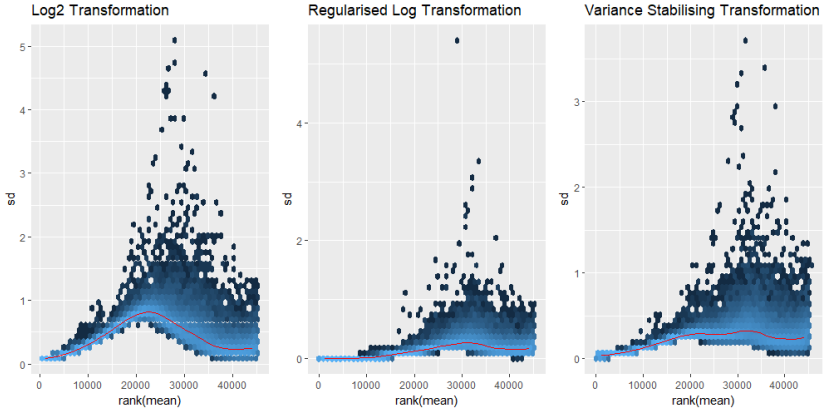


(a) MA Plot

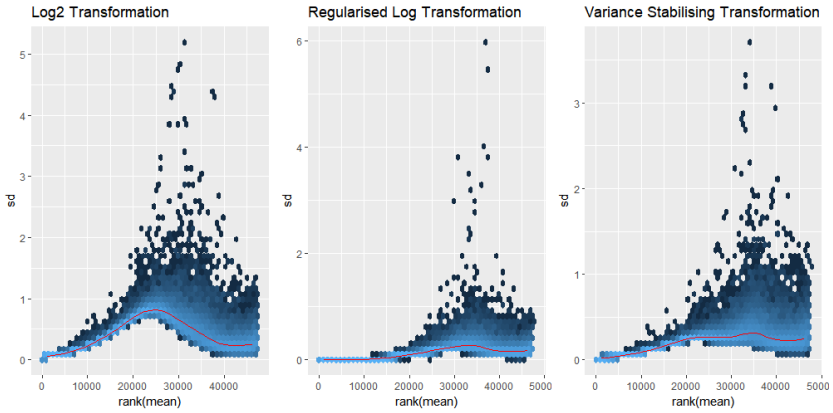


(b) Volcano Plot

Figure E.1: ***a.*** *MA plot showing overall gene expression pattern between CC and TT donors. There is relatively little differential expression shown, yielding an MA plot with few extreme points.* ***b.*** *Volcano plot showing statistical significance of differential expression plotted against the* $log_2(foldchange)$*. Again, there is relatively little observable differential expression in this plot compared to typical volcano plots.*

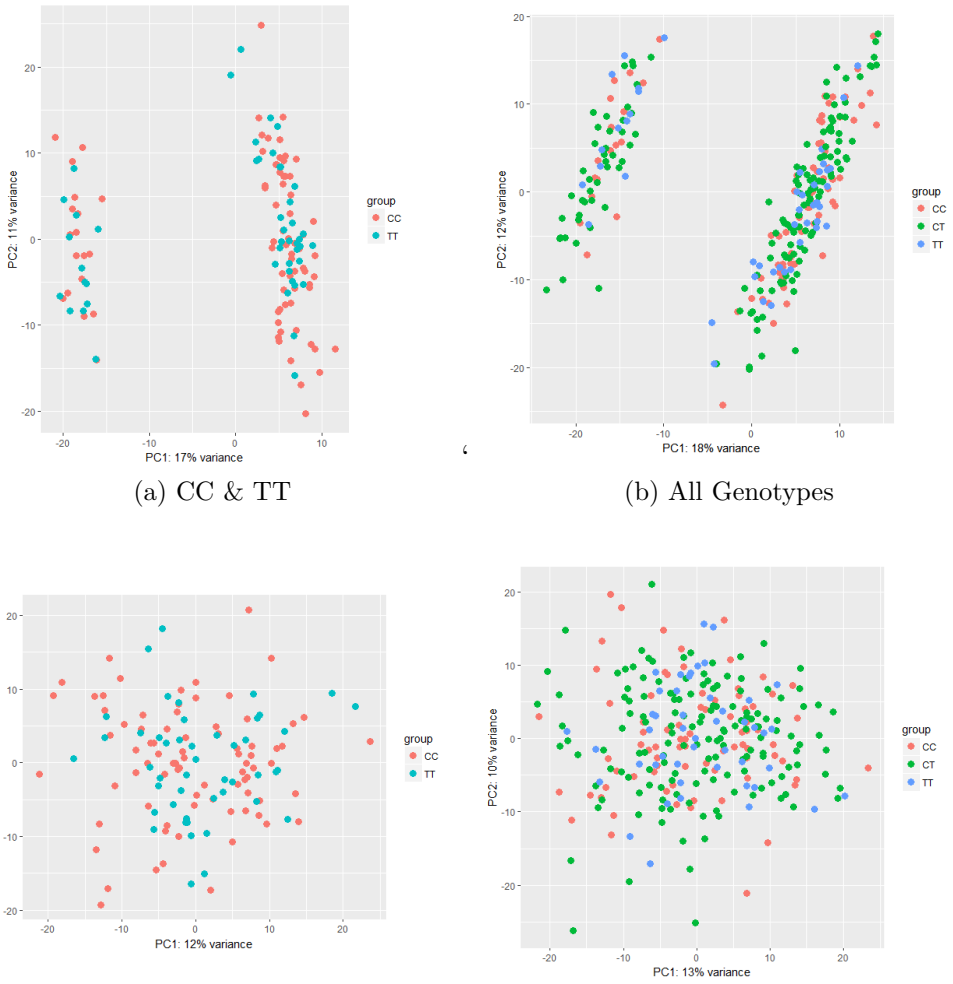


(a) CC vs TT Data Transformations



(b) CC, TT & CT Genotype Data Transformations

Figure E.2: ***a.*** *Graphics showing mean gene expression plotted against the standard deviation for each gene in CC and TT donors. The relative linear nature of the red trendline in the variance stabilisation method indicates that it is the ideal transformation method.* ***b.*** *The same transformation methods used above are applied to CC, TT and CT donors. Again, the variance stabilising method has the most stable trendline.*



(a) CC & TT

(b) All Genotypes

(c) CC & TT, No X/Y Genes

(d) All Genotypes, No X/Y Genes

Figure E.3: PCA plots showing clustering of samples. ***a & b.*** *PCA plots show clear clustering between two distinct donor groups. Clustering was shown to be based on the sex of the donor, with male and female donors clustering separately.* ***c & d.*** *PCA plots made without considering genes on X and Y chromosomes show no clear clustering pattern. Genotype does not have a strong effect on clustering, and sample to sample distances are small*

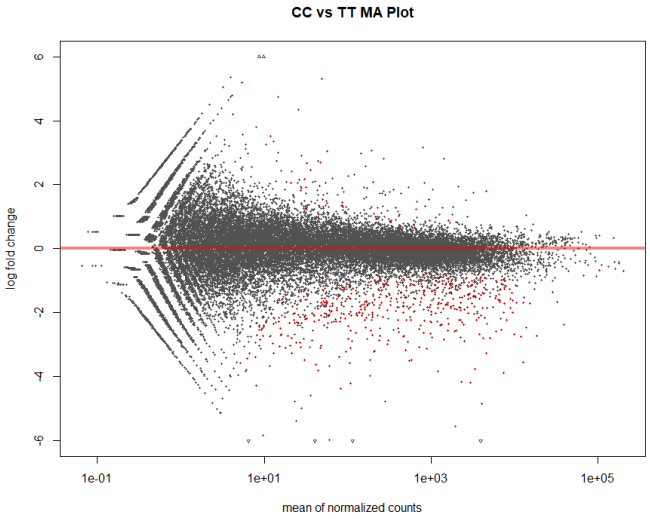
orientation on the PC1 axis. Here, we could see that this was driven by expression of Y chromosome genes, indicating that the separation of donors was related to the donors' sex. Based on this finding, we performed PCA again without the consideration of X or Y chromosome genes. Upon doing so, we saw no clustering in either PCA plot, with very limited dispersion among donors in general (figs. E.3c and E.3d). Finally, we attempted to identify if *IFITM3* expression dictated either the PC1 or PC2 axes. Here, we performed linear regression between *IFITM3* and both principle components. PC1 showed a statistically significant relationship with *IFITM3* expression ($p < 2 \times 10^{-16}$). However, this showed a relatively small adjusted r^2 value of 0.0258.

Appendix F

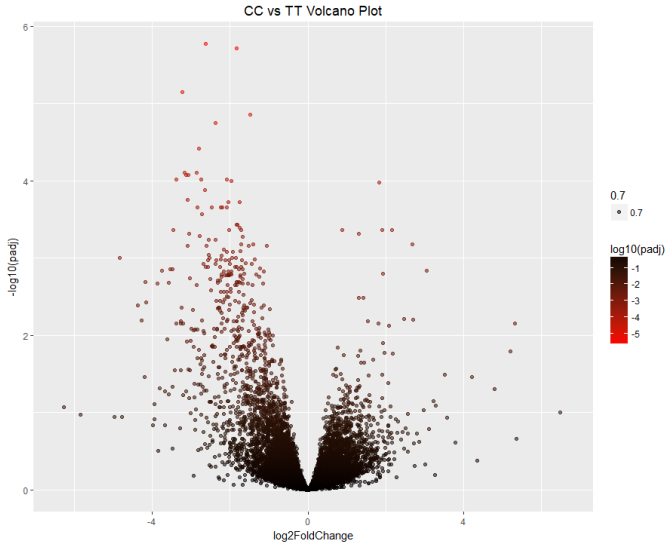
Additional Analysis of 2009 H1N1 Cohort

We plotted both MA and volcano plots to visualise overall results trends between CC and TT patients (fig. F.1). As in section 4.2.5, the MA plot was produced by plotting mean expression of all genes are against the log fold change in expression between the two sample groups. In a volcano plot, the log fold change in expression is plotted against the $-log_{10}$ of the adjusted p-value. In both of these plots, we observed a larger degree of significant upregulation in CC patient samples than TT patient samples.

Before analysis of expression data via PCA, we determined the best method offered by DESeq2 for transforming normalised count data. We performed all possible transformations on our data, and plotted the mean counts for each gene against their respective standard deviations. For both the CC/TT and additive model, it the standard deviation of genes was kept the most consistent across all expression levels by the variance stabilising transformation (fig. F.2). For this reason, we used this transformation going forward. In addition, by plotting the mean expression level for all genes against $-\log 10(p - value)$ for each gene, it could be observed that significantly differentially expressed genes did not favour genes of high or low expression levels.

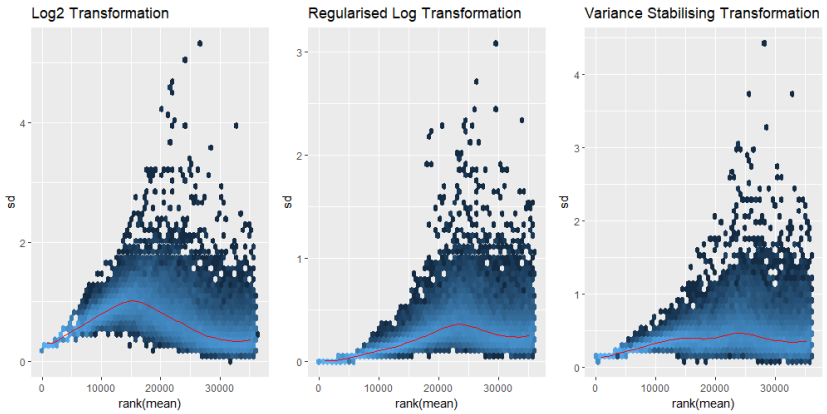


(a) MA Plot

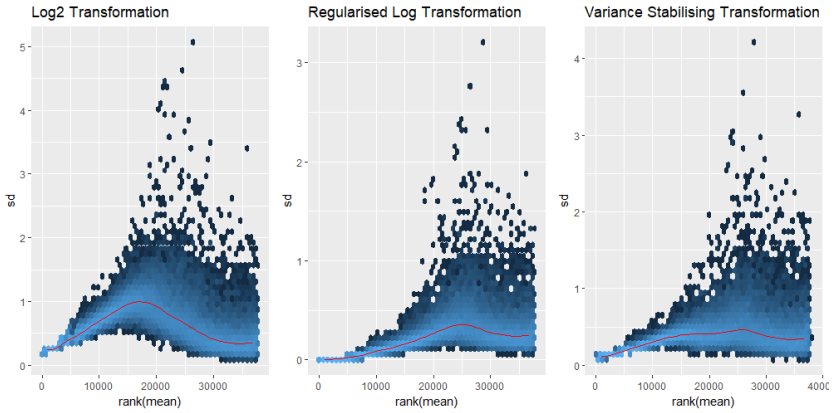


(b) Volcano Plot

Figure F.1: ***a.*** *MA plot showing $log_2(fold-change)$ in gene expression plotted against mean expression between CC and TT patients. Genes with significantly differential expression are shown in red. A pattern of greater upregulation is present on the negative side of the $log2(fold-change)$, representing genes upregulated in CC patients.* ***b.*** *Volcano plot showing statistical significance of differential expression plotted against the $log_2(fold-change)$. Again, greater upregulation of genes is apparent at the negative side, representing CC patients.*



(a) CC vs TT Data Transformations



(b) CC, TT & CT Genotype Data Transformations

Figure F.2: ***a.*** *Mean gene expression plotted against the standard deviation for each gene in CC and TT patients. The relative linear nature of the red trendline in the variance stabilisation method indicates that it is the ideal transformation method.* ***b.*** *The same transformation methods used above were applied to CC, TT and CT patients. Again, the variance stabilising method has the most stable trendline.*

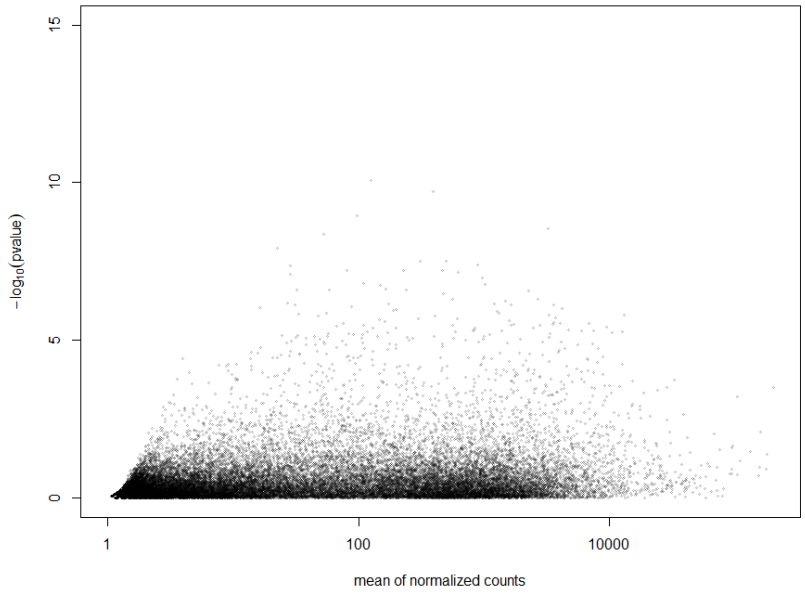


Figure F.3: Mean gene expression levels plotted against their respective p-values. The distribution of p-values is relatively even, with some skew towards higher expression genes. This is to be expected, as low expression genes tend to have high standard deviations, and therefore low p-values