THE UNIVERSITY
of EDINBURGH

THE UNIVERSITY *of* EDINBURGH
Edinburgh Clinical Academic Track

Functional genomics in the regulation of the immune response

Nicholas J Parkinson

Doctorate of Philosophy in Genetics and Genomics

The University of Edinburgh - 2021

Contents

Declaration

I declare that this thesis has been composed solely by myself and that the work presented, including all illustrations, is my own except where explicitly acknowledged. Where experiments or analyses were performed in collaboration with colleagues, their specific contributions have been indicated in the 'Statement of Contribution' or in the text. This thesis has not been submitted, in whole or in part, for any other degree or professional qualification.

Nicholas J. Parkinson

July 2021

Abstract

Genetics contribute substantially to the ability of the immune system to respond appropriately to a challenge. Consequently, many infectious and inflammatory diseases have a heritable component. As genome-wide association studies provide increasing data linking genetic variants to disease, we can leverage this information to gain insights into disease biology if we can elucidate the mechanisms underlying an observed association. Data-driven bioinformatic approaches, *in vitro* mechanistic studies and whole-organism approaches to study integrated pathophysiological systems provide complementary information to help establish causal links between specific variants and their effects on a target gene, and between that gene and disease pathogenesis.

Cells of the mononuclear phagocyte system play key roles in regulation of innate and adaptive immune responses, either via secreted mediators or through direct cell-to-cell contacts. To advance understanding of how these regulatory processes, and genetic variations therein, shape the course of disease, I have combined experimental and bioinformatic approaches to explore novel genetic associations between genes for macrophage surface receptors and human disease.

Specifically, functional follow-up of an association between adhesion G-protein-coupled receptor CD97 and severe influenza showed, using a mouse model, that deficiency of this receptor reduces the efficiency of the $CD8^+$ T-lymphocyte response, a process critical to clearance of infected cells. Secondly, I addressed the question of how to identify causal variants in a disease-associated linkage disequilibrium block, for an association between macrophage regulatory receptor SIRPα and schizophrenia. To achieve this, I developed novel methodology for targeted locus screening using CRISPR/Cas9 mutagenesis, and identified a number of plausible causal regulatory variants that could affect expression of

this receptor.

Combining variant-level information with gene-level studies of disease pathophysiology can provide valuable insights into genetic causation of immune dysregulation leading to disease, which may be harnessed for improved personalised disease risk prediction, or to identify therapeutically targetable pathways.

Lay Summary

Small variations in our genetic code can affect how well our immune system responds to a threat. Consequently, our risk of becoming seriously ill from infectious disease such as influenza, or of developing disease caused by inappropriate 'friendly fire' inflammation, is partly inherited. Modern genetic studies are providing us with increasing information linking specific genetic changes to disease, but it is often unclear exactly how these variations produce their effect. Better understanding of the mechanisms underlying these observed associations will give us new insights into how a disease develops.

Macrophages are a diverse group of cells with multiple important functions in the immune system, including engulfing microorganisms and debris, and initiating and regulating both specific and non-specific immune responses. Abnormal function of these cells is thus of paramount interest in infectious and inflammatory disease. In this thesis, I have used a range of complementary experimental and computational approaches to explore the biological mechanisms underlying two novel associations between human disease and genes involved in communication between macrophages and other cell types. I first used an animal model to show that one of the genes in question is required for efficient stimulation of the type of immune response needed to destroy infected cells after viral infection - this could be a possible mechanism for its association with severe influenza. For an important macrophage regulatory gene linked to schizophrenia, most likely via effects on brain development, I have developed and applied a gene editing-based screening technique to help determine more specifically which of the many small genetic variations found in the gene could be responsible for the effects observed.

These varied approaches help us to understand, at different levels, exactly how genetic variations affect the regulation of the immune system and lead to dis-

ease. This information can help with understanding the biology of the disease, prediction of individual risk, and possibly with design of therapeutic strategies.

Acknowledgements

I would first and foremost like to thank my supervisors, Dr Kenny Baillie, Professor Adriano Rossi and Professor Paul Digard, for all their advice and support which has made this PhD possible. Their input has been invaluable throughout this process, and I also appreciate how much I have been welcomed into their respective research groups. In particular I would like to thank Kenny for creating such a wealth of exciting opportunities in an ever broadening field of functional genomics, and for opening so many doors for new collaborations both before and during the pandemic. I am also grateful to him for opening my eyes to the endless possibilities of bioinformatics, and to the beauty of using code to 'automate the boring stuff'.

I would like to thank all the past and present members of the Baillie lab for their daily support, and for teaching me most of what I know about wet and dry lab work - particularly Sara for teaching me everything I needed to know about cell culture, mice and many other wet lab techniques, Tim for teaching me the dark arts of cloning and CRISPR, and James for keeping everything running smoothly and helping to keep everybody sane. Beyond the Baillie group, I must also thank Anna for teaching me everything I know about flow cytometry, Andy and Steve for trying (admittedly sometimes in vain) to keep my coding style on the straight and narrow, and the various members of the Digard, Dutia and Gaunt groups (especially Ellie, Rute and Marlynne) for sharing their practical influenza expertise.

I am grateful to the ECAT scheme directors for all their advice and feedback, especially when choosing PhD projects, and to the Wellcome Trust who funded this PhD research fellowship. A special mention must go to Jo Ness, for her tireless work keeping the ECAT wheels turning, and without whom it is doubtful whether I would ever have got the right paperwork to the right people at the right

time.

Finally, I would like to thank Luisa for her constant support and her tireless (but largely unsuccessful) attempts to convince me of the superiority of Bayesian statistics, Samuel for his critique of the aesthetics of my work and assistance with analysis of immunofluorescent images ("the green ones have got the flu"), and Ms T Parkinson for her promotion of a paperless workspace.

List of publications

Publications related to, but not part of this thesis:

1. Parkinson N, Rodgers N, Fourman MH, Wang B, Zechner M, Swets MC, Millar J, Law A, Russell CD, Baillie JK et al. 2020. Dynamic data-driven meta-analysis for prioritisation of host genes implicated in COVID-19. *Scientific reports* **10**.

2. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick A, Rawlik K, Parkinson N, Pasko D, Walker S, Richmond A, Fourman MH, ... Baillie JK. 2020. Genetic mechanisms of critical illness in Covid-19. *Nature* doi: 10.1038/s41586-020-03065-y.

3. Clohisey S, Parkinson N, Wang B, Bertin N, Wise H, Tomoiu A, The Fantom Consortium, Summers KM, Hendry RW, Carninci P ... Baillie JK. 2020. Comprehensive characterisation of transcriptional activity during influenza A virus infection reveals biases in cap-snatching of host RNA sequences. *Journal of virology* doi: 10.1128/JVI.01720-19.

4. Li B, Clohisey SM, Chia BS, Wang B, Cui A, Eisenhaure T, Schweitzer LD, Hoover P, Parkinson NJ, Nachshon ... Baillie JK. 2020. Genome-wide CRISPR screen identifies host dependency factors for influenza A virus infection. *Nature communications* **11**.

Statement of Contribution

I would like to acknowledge the following assistance and specific contributions of colleagues to the experiments and analyses described:

Chapter 3:

- Establishment and back-crossing of the Cd97-/- mouse strain:

  Kristin Sauter and Sara Clohisey

- Assistance with mouse experiments, and tissue collection and processing:

  Sara Clohisey, Jun Hu, Marlynne Quigg-Nicol and James Furniss

- Provision of viral strains and plasmids (including rescue of Eng195 strain):

  Bernadette Dutia, Sara Clohisey, Eleanor Gaunt and Yvonne Ligertwood

- Assistance with flow cytometry experimental design:

  Anna Raper and Marlynne Quigg-Nicol

- Design of qPCR primers for influenza A virus segments, and for a subset of host genes analysed:

  Eleanor Gaunt and Jun Hu

- Advice on statistical analysis:

  Helen Brown

Chapter 4:

- Study concept:

  Kenneth Baillie, Peter Hohenstein, Derya Ozdemir and David Hume

- Assistance with sgRNA design:

  Derya Ozdemir

- Original code for genome-wide CRISPR knockout guide design and screen analysis, from which the pipelines developed in this study were adapted:

  Kenneth Baillie

- Refactoring and optimisation of the permutation-based analysis script (prior to adaptation for this study design), and assistance with high performance computing:

  Andy Law

- Cloning of the LentiCRISPRv2-xCas9 plasmid:

  Tim Regan

- Assistance with lentiviral packaging and titration, and optimisation of transduction conditions and antibiotic selection protocols for THP-1 cells:

  Tim Regan

- Assistance with single guide cloning, plasmid production and library preparation:

  James Furniss

- Fluorescence-activated cell sorting:

  Bob Fleming and Graeme Robertson

List of Figures

List of Tables

Abbreviations and Acronyms

ADGRE5	Adhesion G-protein-couple receptor E5, an alias for CD97.
APC	Allophycocyanin.
ARDS	Acute respiratory distress syndrome.
ATP	Adenosine triphsophate.
B2M	Beta-2-microglobulin.
BALF	Bronchoalveolar lavage fluid.
BCA	Bicinchoninic acid.
BMDM	Bone marrow-derived macrophage.
bp	Base pairs.
BSA	Bovine serum albumin.
cCRE	Candidate cis-acting regulatory element.
ChIP	Chromatin immunoprecipitation.
C.I.	Confidence interval.
CRISPR	Clustered regularly interspaced short palindromic repeats.
cRNA	Complementary RNA.
CT	Cycle Threshold.

CTCF	CCCTC-Binding Factor.
DMEM	Dulbecco's Modified Eagle's Medium.
dpi	Days post infection.
EDTA	Ethylenediaminetetraacetic acid.
EGF	Epidermal growth factor.
ELISA	Enzyme-linked immunosorbent assay.
EMR2	EGF-like module-containing mucin-like hormone receptor-like 2.
eQTL	Expression quantitative trait locus.
FBS	Fetal bovine serum.
FDR	False discovery rate.
GAIN	G-protein-coupled receptor autoproteolysis-inducing domain.
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase.
G-CSF	Granulocyte colony-stimulating factor.
GDP	Guanosine diphosphate.
GWAS	Genome-Wide Association Study.
H3K27ac	Histone H3 acetylation at lysine residue 27, a marker of enhancer activity.
HA	Haemagglutinin.
HLA	Human leukocyte antigen.
HPRT1	Hypoxanthine phosphoribosyltransferase 1.
IAV	Influenza A virus.
IF	Immunofluorescence.

IFITM3	Interferon-inducible transmembrane protein 3.
IFNα	Interferon α.
IFNγ	Interferon γ.
IL	Interleukin.
IQR	Interquartile range.
IRF	Interferon regulatory factor.
ISG	Interferon-stimulated gene.
KIR	Killer cell immunoglobulin-like receptor.
LD	Linkage disequilibrium.
LPS	Lipopolysaccharide.
MAF	Minor allele frequency.
MAPK	Mitogen-activated protein kinase.
MDCK	Madin-Darby Canine Kidney cells.
MHC	Major histocompatibility complex.
MOI	Multiplicity of infection, i.e. how many viral units per target cell.
MR	Mendelian Randomisation.
mRNA	Messenger RNA.
NA	Neuraminidase.
NF-κB	Nuclear factor kappa B.
NK	Natural killer.
NP	Viral Nucleoprotein.
PAM	Protospacer-adjacent motif.

PAMP	Pathogen-associated molecular pattern.
PBS	Phosphate buffered saline.
PCR	Polymerase chain reaction.
PE	Phycoerythrin.
pfu	Plaque-forming units.
PMA	Phorbol-12-myristate 13-acetate.
qRT-PCR	Quantitative reverse transcriptase polymerase chain reaction.
rhCSF1	Recombinant human colony-stimulating factor 1.
RIG-I	Retinoic acid-inducible gene I.
RNA	Ribonucleic acid.
RPMI	Roswell Park Memorial Institute 1640 Medium.
SIRPα	Signal regulatory protein alpha.
SNP	Single-nucleotide polymorphism.
sQTL	Splicing quantitative trait locus.
Th1	T helper type 1.
TLR	Toll-like receptor.
TMPRSS2	Transmembrane serine protease 2.
TNFα	Tumour necrosis factor alpha.
vRNA	Viral RNA.
vRNP	Viral nucleoprotein.

Chapter 1

General Introduction

1.1 The immune response is shaped by host genetics.

The immune system must be finely regulated to ensure efficient elimination of pathogens while minimising damage to host tissues, and immune dysregulation is a common cause of disease. Inefficiency of specific aspects of the immune response can allow unchecked spread of viral, bacterial, fungal, protozoal or parasitic infections, while an excessive or inappropriate response can equally lead to severe disease via immunopathology. For example, the systemic inflammatory response to infection is a key feature of sepsis, and specification of a dysregulated host response has recently been included in the internationally accepted definition of sepsis.[1] Similarly, an inappropriate immune response can cause inflammatory or auto-immune disease in the absence of infection, in conditions such as inflammatory bowel disease, asthma and rheumatoid arthritis. Many factors can influence the ability to mount an appropriate immune response, including prior pathogen exposure, age, comorbidities and microbiota, but there is increasing evidence that both susceptibility to infectious disease and risk of non-infectious inflammatory disorders are in part heritable.

1.1.1 Mortality from infectious disease is heritable.

It has been readily apparent in recent viral pandemics, including the SARS-CoV-2 pandemic and the 2009 H1N1 influenza A virus (IAV) pandemic, that infection with a single pathogen can have vastly different effects in different individuals. Infection with SARS-CoV-2 (the causative agent of COVID-19) in most people causes mild to moderate symptoms, often consisting of cough, fever and 'flu-like' illness; many people indeed remain asymptomatic. However a small proportion of those infected develop severe respiratory disease requiring hospital treatment and in some cases invasive ventilation. Similarly, during the 2009 H1N1 ('swine flu') pandemic, most people developed mild disease, while again a small proportion required advanced respiratory support. In this latter case, however, younger patients seemed to be over-represented compared to the seasonal form of the disease, with those under 65 years of age accounting for over 75% of deaths, in contrast to the clear increased risk in older age groups for COVID-19.[2–4]

Some of the variation between people can be explained by known host factors such as age or comorbidities, including chronic lung disorders such as asthma, obesity, chronic renal, hepatic or cardiac disease, diabetes or immunosuppression, the majority of which have epidemiological evidence to indicate they are risk factors for mortality in both influenza and COVID-19.[2,4] However, over 20% of hospitalised COVID-19 patients[4] and an estimated 25-50% of hospitalised 2009 H1N1 patients[3] have had no reported underlying disease. In many cases we cannot explain why, for example, a young healthy adult infected with IAV develops severe viral pneumonia while other members of the same household are barely affected, or why one 55-year-old male with hypertension develops minimal symptoms after infection with SARS-CoV-2 while others in the same risk category die despite treatment. Some of the unexplained variability could relate to viral factors, such as dose or route of infection. Part could furthermore be explained by unidentified host factors, such as previous exposure to similar pathogens (a phenomenon hypothesised to underlie the increased infection rate for 2009 H1N1 IAV in people below 65 years of age).[2] However, there is strong evidence that individual risk for infectious disease, whether defined as risk of infection or risk of developing severe or fatal disease following infection, is determined at least in part by genetics.

Establishing heritability for complex traits such as infection susceptibility re-

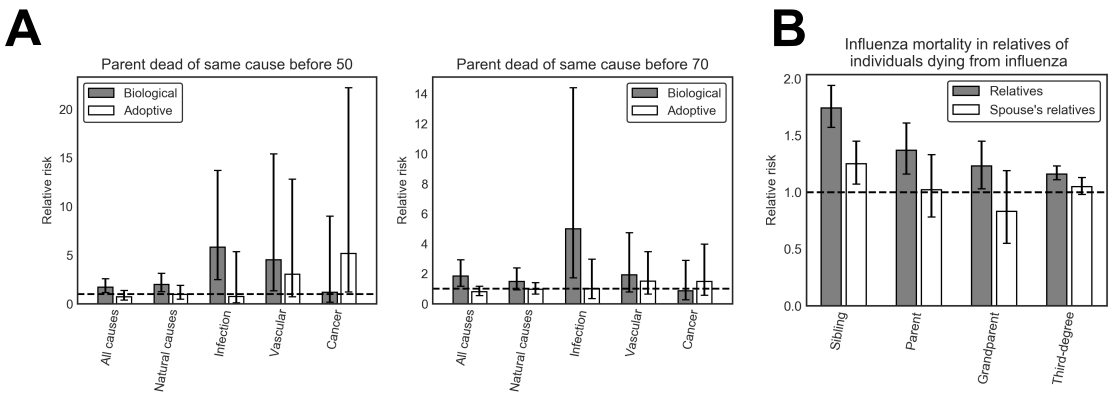

Figure 1.1 – Genetic contribution to susceptibility to infectious disease. A: Relative risk of early death (before 50 or 70 years) in adoptees when either a biological or adoptive had died early of the same cause. There is a 5-fold increase in risk of death from infection with either age cutoff. Original data are from Sørensen *et al.* (1988).[5] B: Relative risk of death from influenza in relatives of individuals dying of influenza compared to relatives of the spouses of influenza victims. The increased risk in biological relatives extends to third degree relatives. Original data are from Albright *et al.* (2008)[6]. Error bars indicate the 95% confidence interval for the relative risk; the horizontal dashed line indicates a relative risk of 1 (i.e. no effect).

quires separation of effects of shared genetics from those of shared environment. Classic approaches to this include twin studies and studies of adopted children. A classic example of this is the 1988 study by Sørensen *et al.* of 1003 Danish adoptees[5], which compared the relative risk of early mortality (defined as death before either 50 or 70 years of age) in adoptees when either a biological parent (i.e. with shared genetics but no shared environment) or an adoptive parent (i.e. shared environment without shared genetics) had died early of the same cause. This study showed a five-fold increase in risk of death from infectious disease overall when a biological parent had succumbed to infection, while no significant increase in risk was apparent if an adoptive parent had died from infectious disease (Figure 1.1). This was similar to the increase in risk for death from cardiovascular disease before the age of 50, and unlike vascular disease or cancer was also significant for death before 70 years of age. While the disease categories are broad (an inevitable consequence of using historic data of this type) and the study has some limitations, such as small numbers of deaths in individual categories and dependence on assumptions such as lack of early-

life environmental effects and random socio-economic placement of adoptees, this provides convincing evidence that genetic background modifies risk independently of environment.

Twin studies take a complementary approach, comparing monozygotic (genetically identical) twins to dizygotic twins. These groups are assumed to have a similar degree of shared environment, and thus a difference in concordance (the probability that, if one twin is affected by a condition, the other will also be affected) provides evidence of a heritable component to the condition. A number of twin studies have been reported for infectious diseases, with substantially greater concordance in monozygotic twins for a range of bacterial and viral diseases, particularly those that result in chronic conditions, including tuberculosis, leprosy, poliomyelitis and hepatitis B.[7] A twin study of COVID-19 has modelled the influence of additive genetic effects versus shared and unique environmental factors on self-reported symptoms in 962 pairs of twins (674 monozygotic and 288 dizygotic, 87% female). This found evidence of heritability for specific self-reported symptoms such as delirium (heritability 49%, 95% confidence interval 32-64%), and for overall predicted COVID-19, determined from symptoms using a model previously trained on over 5000 tested individuals (heritability 31%, 95% confidence interval 11-48%).[8] Estimation of heritability of severe COVID-19 on the basis of the contribution of common polymorphisms to the phenotype indicated a substantially lower heritability of 6.5%, although this is a conservative estimate as it does not include the effect of rare variants with large effect sizes which may occur in at least 3.5% of severe cases.[9,10]

No twin studies have been reported for influenza severity or susceptibility, but a number of studies have used genealogy or case clustering to assess a possible heritable component. The strongest supporting data come from examination of genealogical records of 4855 individuals from Utah for whom influenza (with or without positive viral identification) was recorded as the cause of death over a 100-year period.[6] To distinguish between shared environment and genetic effects, the authors evaluated the relative risk of death from influenza in either biological relatives or relatives of spouses of individuals dying from influenza. The relationship between genetic distance from the influenza victim (or their spouse), and degree of shared environment with the victim, is assumed to be comparable between these two groups. While there was a clear effect of environment (most likely related at least in part to exposure), as an increased risk

of death was seen in spouses and spouses' siblings, there was also evidence of a heritable component. Increased risk was seen in more distant biological relatives up to third-degree relatives (Figure 1.1B), with the magnitude of effect diminishing for more distant relatives, while no such increase was seen in more distant relatives of spouses (although it should be noted that the two groups had overlapping confidence intervals for the risk ratios, and were not directly compared statistically). Further support that the increased risk was genetic rather than purely exposure-based comes from the wide temporal separation of deaths in third-degree relative pairs. A similar study of influenza victims from the 1918 pandemic in Iceland found a strong familial effect, but although the difference in relative risk between relatives and spouses' relatives was nominally significant ($p<0.05$) for aunts, uncles and parents, this did not extend to other relationship levels, and so did not provide strong evidence of heritability.[11] However, as well as including only individuals dying within a six-week period in 1918, this study had a sample size ten-fold lower than the Utah study, with a correspondingly lower power to detect an effect at any relationship level. Familial clustering of H5N1 cases provides some further potential supportive evidence of heritability (in this case, of risk of infection rather than risk of mortality), but the interpretation of these data is the subject of some debate. Given the likelihood of shared exposure, a high proportion of case clusters would be expected to consist purely of blood relatives by chance alone, but conversely the observed spatiotemporal separation of related cases is not completely consistent with shared exposure and could suggest underlying genetic effects.[12]

There are a small number of known examples where a single genetic variant has a large impact on susceptibility to a specific infectious disease. The best characterised of these include the effects of inherited haemoglobin disorders, such as sickle cell disease and possibly β thalassaemia on mortality from malaria. In this case the causative mutations are relatively common in populations living in areas where the disease is endemic, presumably reflecting selection pressure. Most other known monogenic effects on infectious disease risk, in contrast, involve rare mutations, such as loss-of-function mutations in interferon gamma receptor genes which increase susceptibility to opportunistic mycobacterial infections, or the CCR5Δ32 deletion mutation that confers resistance to HIV infection.[7] In most cases, the heritability of infectious disease risk is more likely to arise from complex polygenic effects, involving the interaction of small

effects of many variants. These variants could occur either in host genes required for pathogen entry and replication (particularly in the case of viral infections, which rely on the host machinery to reproduce), or in host genes involved in the immune response.

1.1.2 Heritability of the immune response

As for infectious disease, there is evidence from twin studies and other sources that autoimmune and other immune-mediated diseases are partially heritable. For example, conditions such as systemic lupus erythematosus, type I diabetes mellitus, coeliac disease and Crohn's disease all have substantially greater concordance in monozygotic compared to dizygotic twins, indicating the importance of genetic factors in disease development. There is, however, substantial variability in concordance between studies and between diseases, and many have a concordance below 50% even for monozygotic twins - concordance is particularly low for some diseases such as rheumatoid arthritis, suggesting that environmental risk factors may play a greater role, while in contrast concordance may be over 70% for coeliac disease, indicating that this condition is predominantly genetic.[13,14]

Heritability has been demonstrated for many elements of the host immune response, which could contribute to infection susceptibility or to immune-mediated disease. Deep immune phenotyping of peripheral blood mononuclear cells by flow cytometry in a cohort of twins showed that 76% of the 23,394 traits analysed (which included populations of specific immune cell subsets and expression levels of specific markers within these subsets) had evidence of a substantial genetic component, with a mean estimated genetic influence of 45%.[15] Heritability estimates for a wider range of immune phenotypes in the same population ranged from 0 to 96%, with the most highly heritable traits including counts of FcγR2$^+$ myeloid dendritic cells and some subsets of natural killer (NK) and T cells.[16] Although heritable traits could be identified for all broad groups of leukocytes examined (B cells, T cells, NK cells, monocytes and dendritic cells), differences in the likelihood of heritability were apparent across different innate and adaptive immune processes: for example, traits related to CD4$^+$ T-cell polarisation, innate-like T-cell populations and the majority of monocyte traits were more frequently influenced predominantly by environmental factors, while traits

concerning NK-cell, conventional T-cell, regulatory T-cell and dendritic cell populations were more frequently under predominantly genetic control.[15] A genome wide association study (GWAS) of the most highly heritable traits within the same population and in a similar replication cohort, using a model accounting for relatedness, identified replicable associations meeting a stringent threshold for genome-wide significance for a number of traits. These included associations between variants at the locus containing the *FCGR2A* gene and a number of different traits on T cells, B cells and dendritic cells - notably, the associated variants, which seem to exert a particularly strong influence on gene expression in monocytes, have also been associated with a number of immune-mediated diseases including systemic lupus erythematosus and inflammatory bowel disease.[16]

Another twin-based study has, however, suggested that for many components of the immune system, heritable factors are less important than environmental influences. A study of 105 twin pairs in the SRI International Twin Research Registry[17] found highly variable heritability for serum proteins related to immune function, circulating immune cell subpopulations, and *in vitro* responses to cytokines. Of the cell populations studied, 61% had no detectable heritability in circulating numbers, although heritability was greater than 50% for neutrophil counts and higher for specific T-lymphocyte subpopulations such as those expressing CD27 (an activation marker), but not for overall $CD4^+$ or $CD8^+$ T-cell counts. Heritability was high (>75%) for serum concentrations of some cytokines such as interleukin-6 (IL6), interleukin-12p40 and interferon alpha (IFNα), but lower for many others and undetectable for some such as IFNβ and IL-1α. Phosphorylation of STAT transcription factors in response to cytokine stimulation in T lymphocytes and monocytes was highly heritable for IL-2 and IL-7, but a heritable component was undetectable for IL-6 and IL-10, and cell type-dependent for interferons. There was no observable heritability (defined as <20%) in the antibody response to a multivalent IAV vaccine, although this contrasted with high heritabilities previously reported for other vaccines such as those for measles, polio and hepatitis B. Apparent heritability of some cell populations such as regulatory T cells tended to decline with age, presumably due to increased influence of accumulated environmental exposures. Differences in heritability estimates between studies could thus be related to subject age distribution, as well as to technical factors such as surface markers used to define

each cell subset. These twin studies have obvious limitations in that the small sample sizes can only represent a fraction of naturally occurring genetic variation across a population and will limit statistical power to detect small effects, and in that measured phenotypes were for the most part limited to peripheral blood without stimulation, but they do serve to illustrate that while the human immune response may be influenced by a complex mix of environmental and genetic factors, naturally-occurring genetic variation has an impact on specific immune components which could modify the efficiency of the response to viral or other infectious disease, or could modulate the potential for immunopathology.

1.1.3 Value of identifying host genetic factors involved in infection and immunity

There is now a strong body of evidence as to the heritability of susceptibility to infectious and inflammatory disease. This is highly likely to extend to influenza, which is the focus of one of the genetic associations discussed in this thesis. The genetic basis of susceptibility is more than just of academic interest: if we can establish which host genes are most important in determining disease outcomes, we may be able to harness this for disease prevention or therapy. If a gene is strongly implicated in disease, we do not have to intervene at the genetic level for therapeutic benefit, but such an observation would justify investigation into the therapeutic benefits of either targeting the gene product itself, or other mediators in the biological pathways associated with that gene. This is a particularly attractive prospect in diseases caused by viruses such as IAV with a high mutation rate: while a virus can acquire resistance to antivirals, it will be harder to develop resistance to therapies targeting host viral dependency or restriction factors, and resistance will not be a substantial problem for therapies targeting an excessive host immune response.[18] More precise data on variants (as opposed to whole genes) associated with risk may also facilitate improved risk prediction at the individual patient level, conceptually allowing a personalised approach to prophylaxis and intervention. In non-human diseases the benefits could extend even further, enabling implementation of genetically-informed breeding strategies to minimise disease at a population level.

1.2 Mononuclear phagocytes as regulators and effectors of the immune response

While recognised heritable components of the immune system involve a wide range of different immune cell types[15,17], mononuclear phagocytes are of particular interest due to the extent of their involvement in both innate and adaptive immune function, and thus their potential to modulate diseases of many different aetiologies. The mononuclear phagocyte system consists of a broad range of cells, including monocytes, macrophages and dendritic cells, derived from common myeloid progenitor cells. These are a highly diverse group of cells, and specialised forms are adapted to tissue-specific physiological niches, such as alveolar macrophages in the lung, Kupffer cells in the liver and microglia in the central nervous system. They exist in a broad spectrum of differentiation and activation states, and attempts to categorise them into discreet subtypes risk over-simplifying the true population complexity - for example, even the existence of dendritic cells as an entity distinct from macrophages is controversial, as no single marker, function or developmental origin can completely separate them from the continuum that comprises the mononuclear phagocyte population as a whole.[19]

Mononuclear phagocytes are the primary immune sentinels in tissues, detecting and phagocytosing pathogens, foreign material and cellular debris. They express a range of pattern recognition receptors to detect and respond to potential 'danger' signals, including pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs). Pattern recognition receptors include the toll-like receptors (TLRs) and NOD-like receptors, located either on the cell surface or intracellularly in endosome membranes, which respond to a range of microbial, viral and 'self' products. For example, TLR4 responds to bacterial lipopolysaccharide (LPS) but also to products of extracellular matrix breakdown or host cell lysis, such as heparan sulphate and high mobility group box protein-1, while TLR3 responds to double-stranded RNA in endosomes. Receptor activation triggers a range of signalling pathways, via adaptor proteins such as myeloid differentiation primary response (MyD88) and TIR domain-containing adaptor inducing IFN-β (TRIF), which initiate a variety of different phosphorylation cascades, leading finally to activation and nuclear translocation of transcription factors such as nuclear factor kappa B (NF-κB)

and interferon regulatory factors (IRFs), stimulating the production of interferons and other cytokines.[20] This is one of the key triggers of the early innate immune response. The factors secreted by the macrophages (or other mononuclear phagocytes) have numerous functions including immune cell chemotaxis, activation of effector cells (including the macrophages themselves), cytotoxic effects (e.g. tumor necrosis factor α) and anti-inflammatory effects (e.g. interleukin 10). Dysregulation of this aspect of macrophage signalling function thus has the potential for substantial effects on the magnitude and phenotype of the ensuing inflammatory response.

Antigen presentation functions of mononuclear phagocytes are equally critical in initiation of the adaptive immune response. While dendritic cells are considered to be the most highly adapted for efficient antigen presentation to T cells, many macrophage populations can perform this function to some extent, and claims that dendritic cells alone are capable of activating naïve T cells are disputed on the basis of evidence of induction of a primary T-cell response *in vivo* by certain populations of activated macrophages.[19] Antigen presenting cells process and present endogenous antigen (including viral products) via major histocompatibility complex (MHC) class I, a receptor that is present on most host cells, and exogenous phagocytosed antigen via MHC class II, a receptor that is more specific to antigen-presenting cells. If a T cell carries a T-cell receptor with a strong affinity for the specific MHC-mounted antigen, either for MHC I-bound antigen for $CD8^+$ T cells or MHC II-bound antigen for $CD4^+$ T cells, antigen presentation coupled with appropriate co-stimulation via interactions of other surface receptors will result in clonal expansion and activation of the lymphocytes. Antigen presentation by some subsets of macrophages may also be involved in development of antigen-specific peripheral tolerance.[21] The MHC is encoded by a hyper-variable gene cluster, known in humans as the human leukocyte antigen (*HLA*) locus. Genetic variability at this locus determines the range of antigens to which a host can respond, and as such is likely to be a key determinant of susceptibility to infectious disease. *HLA* variability furthermore explains a substantial proportion of the heritability of autoimmune diseases.[14] However, a wide range of other genes, such as those affecting antigen processing, cell migration or co-stimulation, could equally affect the efficiency of antigen presentation and T-cell stimulation. As T-cell help is required for efficient B-cell responses, this function of mononuclear phagocytes is essential for both the cell-mediated and

antibody-mediated aspects of a primary adaptive immune response.

Effector functions of macrophages are augmented by macrophage activation and polarisation. Polarised forms of activated macrophages are commonly designated 'M1', promoted by, among other factors, cytokines such as interferon-γ (IFNγ) and interleukin-2 (IL-2) that are associated with T helper type 1 (Th1) cells, or 'M2', promoted by Fc-gamma receptor activation or by Th2-associated cytokines such as IL-4, IL-10, IL-13. This is an over-simplification, and these classifications are best viewed as extremes on a broad phenotypic spectrum. Moreover, these are not terminally differentiated states, and macrophages retain much plasticity to convert between activation states. A large number of genes are differentially expressed on macrophage activation and polarisation, to determine the resulting cellular phenotypes. M1-like macrophages, which are considered pro-inflammatory and primed for microbial killing, have augmented MHC II expression, phagocytic function, nitric oxide synthase 2 activity and chemotactic activity, and secrete numerous products including prostaglandins, complement components and mediators of coagulation and fibrinolysis. M2-like macrophages promote resolution of inflammation, for example by secretion of IL-10, and tissue repair, for example by promotion of fibroblast migration and secretion of growth factors such as transforming growth factor beta and fibroblast-like growth factor beta. An excessive M1-like response could lead to increased host tissue damage, while an inappropriate M2-like response can be associated with fibrotic disease or with poor outcomes in cancer.[22] Therefore, tight regulation of the activation process and effector functions is required throughout the course of the response to an inflammatory stimulus, to maintain the appropriate balance of pathogen clearance versus inflammation resolution and tissue remodelling.

The range of important roles of mononuclear phagocytes both as effectors and orchestrators of the immune response make dysregulation of macrophage function a prime candidate mechanism in the pathogenesis of many immune-mediated diseases. While multiple intrinsic and extrinsic factors shape the mononuclear phagocyte response, host genetics are likely to explain some of the difference in response between individuals. For example, the link between genetics, monocyte-macrophage function and disease is becoming increasingly apparent in inflammatory bowel diseases including Crohn's disease and ulcerative colitis, immune-mediated diseases with a multifactorial aetiology

for which a number of genome-wide association studies GWAS have been performed. Comparison of GWAS results to transcriptomic data from monocytes has shown that inflammatory bowel disease-associated loci (but not loci associated with rheumatoid arthritis) are highly enriched in promoters that are differentially expressed in monocytes either on differentiation to macrophages or on stimulation with LPS.[23] Similarly, network density analysis using promoter-level co-expression data across a wide range of cell types and conditions has shown that these inflammatory bowel disease-associated loci are preferentially associated with co-expression clusters principally active either in intestinal tissues and epithelial cells (as expected based on disease localisation) or in monocytes and macrophages.[24]

As fewer large scale genome-wide association studies have been performed for infectious disease severity or susceptibility, it is harder to systematically evaluate enrichment of mononuclear phagocyte-associated genes among the results. Nevertheless, given the key roles of these cells in the response to infection, variants in genes with effector or regulatory functions in monocytes, macrophages or dendritic cells would be plausible candidates to modify disease severity. In influenza, for example, while genetic associations with disease include genes with a range of functions in viral replication, innate and adaptive immunity, a number are involved in mononuclear phagocyte responses, as will be reviewed Section 1.4. In this thesis, two novel genetic associations will be investigated between human disease and variants in genes highly expressed in mononuclear phagocytes. The two disease phenotypes, severe influenza and schizophrenia, while very different in clinical presentation and in pathogenesis, are both complex phenotypes in which immune dysregulation, interacting with other factors such as viral replication or neurotransmitter dysfunction, has been implicated as a component of the pathophysiology. In each case, the gene in question encodes a surface receptor involved in intercellular interactions between mononuclear phagocytes and other cell types, with likely roles in immune regulation.

1.3 Causation in genetic associations

1.3.1 Association versus causation

Heritability of a disease implies that a genetic variant or combination of genetic variants causes the disease, or causes an increased risk of that disease. The primary goal of genetic studies is to find and understand these causative factors. Candidate gene studies aim to associate variants in a gene of interest with a disease where a prior hypothesis exists as to the relationship between the gene and the disease. While focusing on a small number of variants increases statistical power by reducing the burden of multiple comparisons, this approach has limited scope to find new information and is prone to bias. Genome-wide association studies (GWAS) in contrast are an unbiased approach that can detect associations in previously unsuspected regions of the genome, and while they require greater sample sizes to achieve adequate statistical power due to the large number of hypotheses simultaneously tested, they have proven substantially more replicable than the candidate gene approach.[25] Genetic associations from either technique, however, provide us only with correlation, not causation.

A range of causal pathways could result in correlation between a genetic variant and a disease phenotype, as illustrated in Figure 1.2.[26] A causal pathway could be non-existent, in the case of spurious correlations related to technical error or sampling bias: for example, if using blood donors as a control group, variants at the *ABO* locus could be spuriously correlated with disease due to bias in selection of donors with blood groups in high demand. The simplest model of a true causative association would be a pathway whereby a single variant induces a measurable change in function of a single target gene, which has effects on a biological pathway (via a mechanism involving an unspecified number of intermediaries), which in turn modulates disease outcome (Figure 1.2A). Mechanistic inference could be complicated by pleiotropy (Figure 1.2B), where a single variant has effects on more than one gene, or where that gene has effects on more than one downstream mediator. This is particularly a concern for analyses using genetic variants as an instrument to determine causality of one of these downstream mediators. Gene-environment correlations can furthermore provide alternative causal pathways. A genetic variant could modify exposure to an environmental variable which is the true cause of modified disease risk (Figure

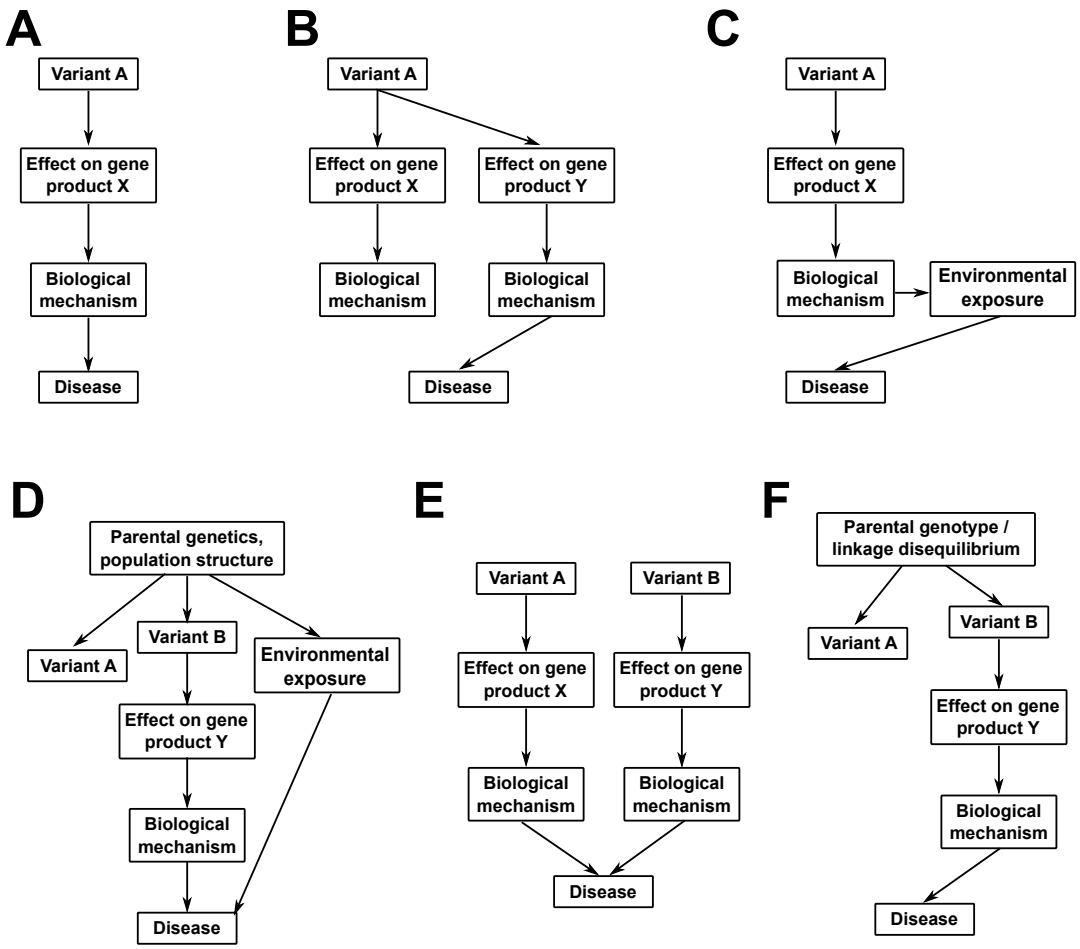

Figure 1.2 – Directed acyclic graphs of causal pathways in genetic associations. Possible pathways are shown where Variants A and B are associated with the disease of interest. An arrow between two nodes indicates that the first causes the second. A: A simple causal pathway involving no extrinsic factors. The biological mechanism linking the effect on the gene product to the disease could still in this case be highly complex, involving additional intermediates. B: Pleiotropy, with a variant affecting more than one gene; either or both genes could be causally linked to the outcome. C: Indirect causality via an extrinsic (environmental) factor. D: Effect of population stratification. E: Independent causal effects of two variants. F: Linkage disequilibrium, with one causal and one proxy variant.

1.2C). For example, variants in the nicotinic receptor gene *CHRNA3* have been associated with smoking behaviour, and this is the most likely explanation for its repeatable association with lung cancer.[27,28] This type of association could still be considered a true causal link, but its indirect nature involving extrinsic factors

would complicate attempts to investigate the underlying biological pathways.

Population structure can provide substantial confounding: if a disease has a greater prevalence in a particular population sub-group, due either to genetic, socio-economic or environmental factors (for example where there is geographic clustering of subpopulations), any variant which has a different allele frequency in that sub-group compared to the rest of the study population may appear to be correlated with the disease phenotype (Figure 1.2D). A variety of methods are available to control for population structure, such as adjustment using genetic principal components in GWAS analysis, but it may be difficult to eliminate such confounding completely.[29] Some types of confounding or alternative causal pathways that pose problems with causal inference in other epidemiologic studies are unlikely to be a concern in genetic association studies, as genotype of an individual is fixed. As such, reverse causation is not possible, at least at an individual level, and similarly it is unlikely that an unmeasured third factor such as an environmental exposure could influence both disease and genotype (although at a population level, genotype biases in case selection or geographic biases in ancestry as discussed above could have similar effects). This substantially simplifies interpretation of direction of causal effects.

Where two variants are both correlated to the phenotype, it is possible that they have independent causal pathways (Figure 1.2E). Interactions between genotypes are also possible, whereby the causal effect of a variant is synergistic with or dependent on genotype at another locus. Such interactions can be difficult to detect, although improved methods for epistasis detection in GWAS analyses are becoming available.[30] In many cases, however, one variant will not be involved in the causal pathway but will be acting as a proxy for a true causative variant with which it is correlated, as a result of linkage disequilibrium (Figure 1.2F).

1.3.2 Linkage disequilibrium complicates causal inference.

Genetic polymorphisms are not inherited individually but rather tend to be inherited together in haplotype blocks, regions of the genome with little meiotic recombination, and thus genotypes for the variants within such a block will be highly correlated. The probability that the alleles of two variants on a single

chromosome in a gamete will have arisen from different parental chromosome copies depends on the probability of recombination occurring in the intervening chromosomal region, which will depend largely (but not exclusively) on the distance between them. Where a pair of alleles at nearby loci are co-segregated more often than would be expected by chance under the null hypothesis of independent occurrence, they are said to be in 'linkage disequilibrium', a property that can be quantified by the following metrics, where p_A and p_B are the individual probabilities of the alleles of interest occurring at simple biallelic loci A and B respectively, and p_{AB} is the probability of the two alleles co-occurring.[31]

1. The coefficient of linkage disequilibrium D, which is the difference between observed and expected allele frequencies.:

$$D = p_{AB} - p_A \times p_B$$

2. D', the coefficient of linkage disequilibrium normalised to the maximum possible difference between observed and expected allele frequencies (taking into account differences in allele frequency between the two loci):

$$D' = \frac{D}{D_{max}}$$

where

$$D_{max} = min\{p_A(1-p_B),\ p_B(1-p_A)\} \text{ if } D > 0$$

$$D_{max} = min\{p_A p_B,\ (1-pA)(1-pB)\} \text{ if } D < 0$$

3. r^2, the squared correlation coefficient between the alleles at the two loci:

$$r^2 = \frac{D^2}{p_A(1-pA)p_B(1-p_B)}$$

Linkage disequilibrium is what makes GWAS possible without the need for whole genome sequencing: it is not necessary to include all of the millions of possible causal variants in the genome in a genotyping panel, as given appropriate panel design and sufficient density of coverage, there is a high chance that one or more variants measured will be correlated with the true causal variant.[32] Linkage disequilibrium does however complicate the interpretation of genetic association data, as we cannot determine which variant at a disease-linked lo-

cus, or even in some cases which gene, is the true causative factor. This is the case for all types of genetic association analysis, including analyses such as expression quantitative trait locus (eQTL) analysis, in which genetic variation is associated with tissue gene expression. Associating a disease with a linkage disequilibrium block is an invaluable first step, narrowing down our search space from a genome of over three billion base pairs to a smaller region tens or hundreds of kilobases long, but causal inference requires additional experimental or statistical evidence.

1.3.3 Statistical approaches to causal inference in observational genetic data

While traditionally intervention has been considered necessary to prove causation, as exemplified by the superiority of randomised controlled trials over observational studies such as cohort studies in clinical evidence-based medicine, a number of statistical frameworks have been developed to approach causal inference in observational genetic studies, to elucidate elements of the causal pathway from primary causal genetic variant, via causative genes and other mediators, to the observed phenotype.

At the variant level, Bayesian analysis of fine mapping data has been harnessed to estimate the probability that individual variants at a locus could contribute causally to the phenotype association observed in a GWAS study, or to produce credible sets of variants with a specified probability of containing the causal variant. The variant with the lowest p-value in a locus will in many cases not itself be the causative variant, due to effects of experimental noise and local correlation structure. This is especially the case for rare alleles, low effect sizes and small samples sizes.[33] A variety of algorithms are available, but the general approach is to start with an assumed set of prior probabilities (for example that each variant is equally likely to be causal) and a specified model for causality at the locus (for example that there is exactly one measured causative variant, and that the other variants have no effect on the phenotype except through LD), and to compute the posterior probability of that model being correct for each variant given the observed genotype data.[34] Causality in this context refers solely to a non-zero effect size; as reverse causation is not possible, this does

provide supportive evidence of causality, but only subject to certain assumptions which may not hold true, including that the true causal variant is included in the measured set and that there is no remaining confounding. This method narrows down the set of likely candidates for the true causal variant at a locus, but will rarely implicate a single variant, and will not be able to discriminate between variants in perfect or near-perfect linkage disequilibrium. Integration with other data on the nature of the variants, such as whether they affect the protein sequence or coincide with known regulatory elements, may help refine these predictions further.

Moving further along the causal pathway, the absence of reverse causation and the random segregation of alleles in meiosis allow genetic data to be used to inform causal inference on the effect of downstream mediators on disease, or of causal relationships between one genetically-determined trait and another. This is the principal behind Mendelian Randomisation (MR), a technique which uses genetic variants as a randomising instrument to distinguish cause from effect when a biological variable is associated with disease, and thus provide evidence equivalent to a randomised controlled trial from purely observational data.[26,35] If a risk factor such as the concentration of a mediator is associated with disease status, it could be causative, reactive, or an incidental result of a common causative factor. If the concentration of that same mediator is robustly associated with a genetic variant, the mediator concentration can only be a consequence and not a cause of the genetic variant (or another variant in LD). Subject to certain assumptions, this variant may be used as an instrument to effectively randomise mediator concentrations across the population independent of any reactive change from disease or other extrinsic factors, and thus association between the variant and the disease can be construed as evidence of causation of the mediator, bypassing the need to measure the mediator itself. This technique has for example been used to show that circulating concentrations of low-density lipoprotein cholesterol are causative for coronary heart disease rather than a consequence of a shared environmental or dietary risk factor.[36] The principal assumptions for this analysis include lack of pleiotropy for the instrumental variant and lack of association with confounding factors, or that there is no alternative pathway by which the instrument could affect the outcome other than via the risk factor under investigation. This is easier to establish if the variant is for example located in the gene encoding a protein under

investigation, than for a variant where the effector gene is unclear or located further upstream from the mediator in question in the relevant metabolic pathway. These assumptions may not be perfectly met in very case and can be difficult to test, as unknown, unmeasured confounders may be present, but nonetheless this technique provides useful information on likely causation which can be therapeutically actionable. Latent causal variable analysis is an alternative method for causal inference between pairs of genetically correlated traits where the traits are highly polygenic and pleiotropy would invalidate the assumptions of Mendelian Randomisation: this relies on the principle that if one trait causes another, variants affecting that first trait will also affect the second trait, but not vice versa.[37]

Methods have been proposed to extend genetically informed causal inference to more complex biological pathways, for example with Bayesian networks. These establish networks of dependencies (which could be potential causal relationships) between various combinations of variants, genes and phenotypes, by assessing conditional independence between variables, i.e. at a simple level, whether any association between a pair of variables disappears after adjustment for an intermediate variable.[38,39] Network approaches can be powerful, and can for example overcome problems such as pleiotropy which would invalidate MR. They are however still limited by the completeness of the data used, and in most cases results of these analyses are considered to suggest or support, rather than to prove, causation. Direct experimental evidence will be needed to confirm such findings.

1.3.4 Combining experimental and observational evidence to demonstrate causation

The theoretical basis for proving that a genetic variant causes disease has been compared to Koch's postulates, a framework proposed in the 19th century to demonstrate the causative role of a specific pathogen in infectious disease. Koch's original postulates were:

- The microorganism must be found in all cases of disease and absent in unaffected individuals.

- It should be possible to isolate the microorganism in pure culture from affected individuals.
- Introducing the microorganism to healthy individuals must cause the disease.
- It should be possible to re-isolate the same microorganism from experimentally induced cases.

It quickly became apparent even within Koch's lifetime that these were not sufficient to cover the full range of naturally-occurring pathogen-disease associations, such as asymptomatic carrier states, or the roles of oncogenic viruses, but this nonetheless provided an initial logical basis for the investigation of host-pathogen interactions. A partially analogous set of postulates have been suggested for determining the causal effect of a genetic variant on a disease[40]:

- A genetic variant should be enriched in the diseased population compared to the control population.
- The variant must be functional and pathogenic.
- Introduction of the variant in a suitable model should cause a similar disease.
- Removal or silencing of the variant should reverse the phenotype.

Variants which fulfill all of these criteria, such as polymorphisms in *TRIM63* implicated in hypertrophic cardiomyopathy[41] can be considered to have a high standard of evidence of causation.

There are clear challenges in applying this framework to genetic causation in human disease. First, the enrichment of the variant in the diseased versus control population, the *sine qua non* without which the other postulates become irrelevant, must be robust and free from confounding. This is easier to achieve for variants with large effects, such as those underlying Mendelian disease, than for the multiple small effects that comprise the genetic basis of complex phenotypes. Such variants require adequately powered genetic association studies that are appropriately designed to remove any confounding effects such as population stratification, ideally with additional replication in different populations

or validation via meta-analysis. It will be much harder to obtain the required evidence for rare variants, which are typically excluded from GWAS analyses: family studies may help if the variants have high penetrance and a Mendelian inheritance pattern, but even in monogenic conditions the effect of a rare spontaneous mutation may have to be inferred by analogy with other known mutations causing the same disease.

Demonstrating that a variant is functional, except for variants causing large structural protein changes such as frame shift mutations, premature stop codons or deletions of large portions of the coding sequence, requires availability of a suitable model. While this is conceptually straightforward, especially with increasing availability of gene editing technologies, it requires some knowledge of expected function. For a protein-coding variant such as a missense mutation, while the protein to investigate is obvious, the function of that protein with respect to that disease must be known, and a suitable functional assay must be available in a physiologically relevant cell type, such as ability of an enzyme to cleave a specific substrate. This will be more problematic for proteins with unknown function, partially known function (such as orphan receptors where the signalling pathway may be assumed by analogy with other proteins but the ligand is unknown), or multiple functions. As a hypothetical example, the receptor CD55, discussed later in this thesis, has known functions in complement regulation, intercellular signalling, and as an entry factor for certain viruses, all mediated by different regions of the protein. Demonstrating that a coding variant modifies one of these functions would only be sufficient to suggest causation if it could also be demonstrated that modifying that function could modulate the disease, independently of the other functions. For non-coding putative regulatory variants, demonstration of a change in expression may be considered sufficient evidence of functionality, but this is still far from straightforward in many cases: as interactions with target genes can occur at a considerable distance it will not always be evident which gene or genes to assess, and as activity of regulatory elements such as enhancers can be highly tissue-specific or perturbation-specific, prior assumptions as to the context of the gene's role in disease are required to select a suitable cellular model.

The need for demonstration of biological relevance of a functional gene perturbation leads to the next proposed steps, the demonstration of induction of a comparable phenotype by introduction of the variant in a suitable model, and

conversely the reversal of the phenotype by removal of the variant. In studies of animal genetics, this may be possible to achieve directly in the species of interest, depending on the practicalities of gene editing and any ethical concerns related to the target species. In human-oriented studies, an animal model may be used, as in the *TRIM63* study cited above.[41] This however is still fraught with problems. No animal model is perfect, but even if a suitable animal model of the disease exists and the gene of interest has a comparable expression pattern, validation at the variant level depends on homology between species. Although conservation between species tends to be high in regulatory regions of the genome as well as in coding sequences[42], sequence differences mean that it is unlikely that a small polymorphism such as a single nucleotide polymorphism (SNP), even if it results in a missense mutation in the protein, would have identical effects in rodents and humans. It may therefore be necessary to take a more step-wise approach: to demonstrate the effect of the variant in a human cellular model, to demonstrate the effect of the whole gene on phenotype in an animal model (if possible controlling for any alternative functions of the gene), and if possible confirming the relevance of animal model findings to humans with additional observational data or *in vitro* modelling. In some cases sufficient data may be obtainable without animal models where there is sufficient prior knowledge on gene function. For example, if a gene implicated in severity of viral disease is known to have a direct effect on viral replication, attempting to recapitulate the effect of a variant in the gene in an animal model, with a different genetic background and different susceptibility to the virus, will add little compared to demonstration of an effect of the variant in viral replication in human cells.

A framework for integrating observational and interventional data from animal and human studies, to support the hypothesis of disease causation by a genetic variant in humans, is shown in Figure 1.3. This cannot be considered proof of causation, as while it links the effect of the variant on the gene, the effect of changes in gene function on biological processes, and the effects of both gene and biological process on the disease, there is no direct interventional observation of the effect of the variant on the disease. Consequently, various assumptions are required about exchangeability between models, for example whether an animal model accurately recapitulates human disease or whether a change in gene expression in a cellular model is of sufficient magnitude to be

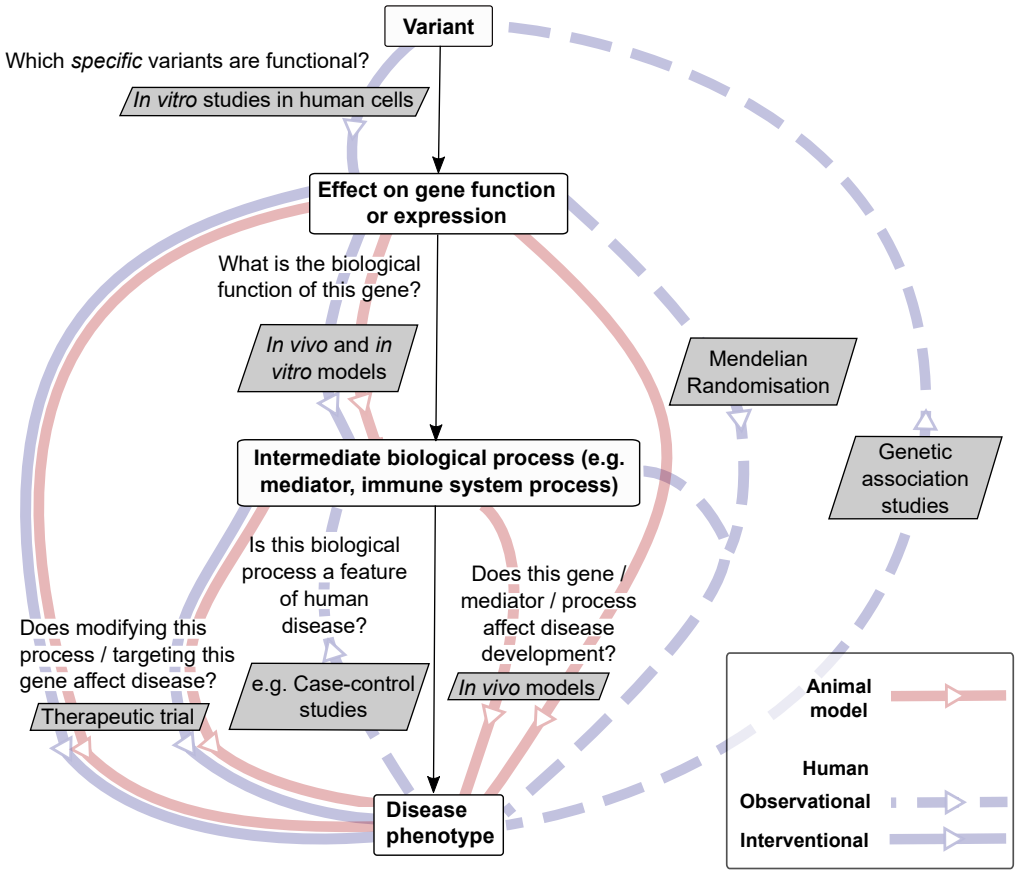

Figure 1.3 – A framework for integrating observational and interventional data to assess genetic causation in human disease. Data types assisting inference between steps of a hypothetical directed acyclic graph, addressing key questions regarding which specific variants are functional, and which biological processes are linked both to these variants and to disease, are shown in parallelogram boxes. Open arrowheads indicate conceptual direction of information from independent (manipulated or selected) to dependent (outcome) variables, although these are to a certain extent interchangeable in purely observational studies.

comparable to that used in whole-animal models. However, when supported by genetic association data (and even more so if there is additional Mendelian randomisation data for downstream mediators), to link the variant to the phenotype in the correct species, albeit via observational data, this framework will provide sufficiently strong evidence of causation to justify investigation of the gene as a possible therapeutic target.

Both statistical and experimental approaches to genetic causation of disease

tend to assume a relatively simple model of causation, whereby a small number of genes are responsible for the majority of the genetic component of natural variation in disease risk or phenotype. An alternative model recently proposed is the 'omnigenic' or 'infinitessimal' model of genetic causation. This model proposes that due to the complexity and inter-relatedness of biological networks, each expressed gene (and all associated functional variants) in a cell type or tissue will directly or indirectly affect the expression or function of all other expressed genes, and thus all will contribute to causation. While this model is not universally accepted as yet, it is consistent with the observation that as GWAS scale and power has increased, it has become increasingly apparent that for many complex traits the genetic association signal is spread over a large part of the genome, rather than being restricted to a small number of discrete loci as originally predicted.[43]

1.3.5 Transcriptional regulation as a potential mechanism underlying the effects of causal variants

Simple monogenic diseases with Mendelian inheritance are often caused by polymorphisms affecting protein structure. These include single missense mutations in the coding sequence, mutations at canonical splice sites, unstable trinucleotide repeats, premature stop codons and frame shift mutations. For some Mendelian diseases such as phenylketonuria, a large number of different mutations in a single gene can cause the same phenotype, with the severity of clinical signs corresponding to the degree of residual activity in the mutated protein.[44] Autosomal recessive conditions are typically characterised by a reduction in protein function, while autosomal dominant conditions sometimes involve gain of an abnormal function or loss of a regulatory function, such as the failure of inactivation of the voltage-gated sodium channel caused by point mutations in the *SCN4A* gene which characterises hyperkalaemic periodic paralysis in humans and horses.[45]

Polymorphisms underlying Mendelian disease have a very large effect size but are typically rare in the population as a whole. This contrasts sharply with the genetic basis of complex phenotypes such as psychiatric disorders or susceptibility to infectious disease. In these diseases, loss-of-function variants with

large effect sizes occur but are rare, while much of the genetic risk comes from common variants with small individual effect sizes, typically distributed over many genes. Analysis of single nucleotide polymorphisms (SNPs) reaching the threshold for genome-wide statistical significance (typically $p < 5 \times 10^{-8}$) across sets of GWAS studies has shown some enrichment of nonsynonymous sites[46], but this is insufficient to explain the majority of the observed signal. In a systematic study of 5,386 genome-wide significant SNPs associated with 679 disease or traits across 920 publications[47], only 4.9% were in coding regions, but 76.6% of the remaining non-coding SNPs were either within a region of DNase hypersensitivity (a marker of putative regulatory regions), or in complete LD with another variant within a DNase hypersensitivity site, a significant enrichment (compared to random samples of SNPs matched by various location parameters) that increased in strength in studies with higher sample sizes or in results with external replication. Of those non-coding variants in DNase hypersensitivity sites within gene bodies, only 10.9% were in strong LD with a coding SNP, suggesting that the non-coding SNPS are not just proxies for a coding variant. This is supported by fine mapping studies, in which credible sets of candidate causal variants often contain few if any likely coding variants.[34,48] Enrichment of lead SNPs in regulatory regions has been a repeatable observation using different data sets and enrichment analyses, and using different features, such as chromatin immunoprecipitation sequencing (ChIP-Seq) peaks or bidirectional transcription as well as DNase hypersensitivity.[49,50] It is thus hypothesised that the majority of causal variants in genotype-phenotype associations with low effect sizes in polygenic traits arise from effects on gene regulation.

Regulatory genomic features, which could contain causal variants, include promoters, enhancers, silencers and insulators. Promoters (Figure 1.4A), located immediately adjacent to the transcription start site of their target gene, are responsible for recruiting the RNA polymerase and initiating transcription. They consist of a core promoter region which contains the primary binding sites for the pre-initiation complex (containing the RNA polymerase and associated general transcription factors) and a proximal promoter region that binds other activating transcription factors. Enrichment of SNPs in promoter sequences has been widely recognised in GWAS results[46,50], and promoter variants both in the core and proximal promoter regions can even in some cases be responsible for Mendelian phenotypes such as mild forms of β thalassaemia.[51]

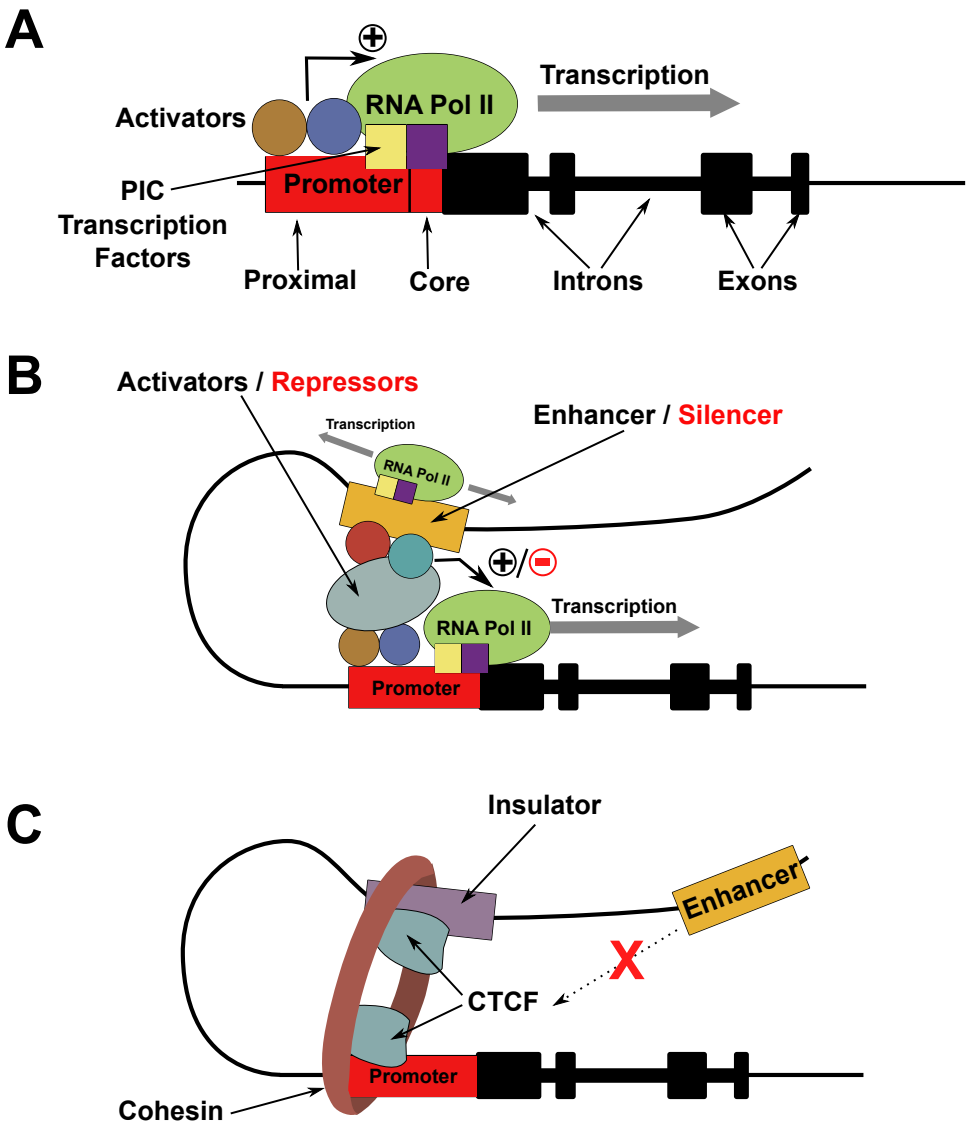

Figure 1.4 – Transcriptional regulatory elements. A: Promoters recruit the transcriptional machinery to the transcription start site and initiate transcription. B: Enhancers interact with promoters via specific transcription factors, and in some cases additional mediating proteins, to increase transcriptional activity, while silencers reduce transcription in a similar manner but via repressive interacting proteins. In some cases a regulatory element may be either an enhancer or a silencer in different contexts, depending on the transcription factors bound. C: Insulators physically isolate enhancers from target promoters, via CTCF/cohesin-mediated chromatin loops.

Enhancers (Figure 1.4B) are short regions, around 200 base pairs in length, that enhance gene transcription. Enhancers can act at a long distance from their target gene, and the most widely accepted model of enhancer function

involves physical interaction with the promoter via chromatin looping and interactions between specific transcription factors bound to the enhancer and proteins bound at the promoter. These interactions sometimes involve other proteins responsible for regulating three-dimensional chromatin structure such as CCCTC-binding factor (CTCF) and cohesin, or elements of the transcriptional machinery such as the Mediator co-activator complex. Unlike promoters, the effect of enhancers is independent of their orientation.[52,53] Enhancers are themselves transcribed, although the whether the resulting RNA is functional remains unclear. A number of common features help to identify enhancers in the genome, including open chromatin, binding of the histone acetyltransferase p300 and multiple transcription factors, and flanking regions of specific histone marks (monomethylation of lysine residue 4 on histone 3, H3K4me1, or acetylation of lysine 27, H3K27Ac).[54–56] Enhancer variants have been implicated both in Mendelian disease and in complex traits: for example, variants such as rs9930506 in an intronic enhancer within the *FTO* gene are among the strongest contributors to risk of polygenic obesity, and this is thought to be via long-range interactions with the transcription factor-encoding gene *IRX3*.[51,57]

Silencers inhibit gene expression. Their mechanisms of action are less well understood than those of enhancers, and may not be the same for all silencers. One model proposes that their chief mechanism is via physical isolation of their target promoter from regions of high transcriptional activity.[58] At least a proportion function via chromatin looping and physical interactions with promoters, in the same manner as for enhancers but with repressive rather than activating transcription factor interactions (Figure 1.4B). Indeed, there is evidence that some regulatory elements can be bifunctional, acting as either enhancers or silencers in different contexts.[59] Similarly to enhancers, silencers are characterised by open chromatin, but although some chromatin features such as high H3K27me3 enrichment may help with identification, there is currently no consensus as to other identifying features, and different screening methods often have limited overlap.[56,60] As an example of a candidate causal variant within a silencer, variant rs75915166 has been proposed to mediate an effect on risk of oestrogen-receptor positive breast cancer by introducing a GATA3 transcription factor binding site into a silencer for *CCND1*, encoding the cell cycle regulatory protein cyclin D1, increasing the suppressive effect of the silencer.[61]

Insulators do not directly suppress promoter activity, but prevent the interaction

between an enhancer and its target promoter, and can have functions including prevention of off-target activation of adjacent genes by an enhancer, or control of alternative promoter usage. Their effects are principally mediated by the effects of CTCF and cohesin on chromatin looping (Figure 1.4C).[62] Mutations causing loss of CTCF or octamer-binding transcription factor binding sites in the silencer between the *IGF2* growth factor gene and the *H19* long non-coding RNA are among the recognised genetic causes of the foetal growth disorder Beckwith-Widemann syndrome.[63]

As the example of *IRX3* regulation in obesity illustrates, the potential distance between interacting pairs of regulatory elements such as enhancers and promoters can complicate identification of the target gene of a putative causal regulatory variant, as it cannot be assumed that effects are mediated via the closest gene to the lead variant. Approaches to identifying these interactions include analysis of chromatin contacts and eQTL colocalisation analysis. Chromatin contacts can be identified by three-dimensional chromatin capture techniques. A number of variations of these techniques are available, but the principle involves formaldehyde cross-linking of chromatin regions in direct contact, shearing the DNA and ligating the cross-linked segments, and then sequencing the ligated segments to identify contacts.[53] If a SNP in a regulatory region is causal, it (and any variants in LD with it) should be an eQTL, i.e. a locus for which the genotype has a measurable effect on gene expression. Comparison of the location and spread of the GWAS signal at the locus with the location and spread of the genotype-expression association signal at that locus in publicly available databases such as GTex[64] can help to identify the target gene, and to clarify whether the two signals are likely to have the same origin.[65] This could however be limited by the highly context-specific nature of some regulatory interactions, as most available eQTL data sources are based on data in unperturbed cells, in a limited range of tissue types. Furthermore, eQTL analysis will be unable to distinguish direct from indirect regulatory effects, for example where an effect on gene transcription is mediated by protein-mediated feedback effects involving the true target gene of the locus.

Altered transcription factor binding, either by loss or gain of a binding site, is the likely mechanism of action of most causal variants lying in regulatory regions.[66] Binding motifs for physiologically relevant transcription factors are enriched in disease-associated variants in GWAS results.[47] The effect of a se-

quence change on transcription factor binding can sometimes be predicted, but this is difficult where the target motif of the transcription factor in question is not clearly defined. Differential transcription factor binding can be demonstrated directly with techniques such as electrophoretic mobility shift assays (as used in the GATA3/*CCND1* example above[61]) or by ChIP-Seq.[66] Identifying the likely transcription factors involved, as well as the target genes, will give us important information about the physiological context of a disease-causing variant.

1.3.6 Other approaches to screening for genes involved in disease development

The beauty of the GWAS as a means for discovery of genes involved in disease development is that it allows us to detect genes which are important in the naturally-occurring disease in the correct species, in a context involving all physiological interactions between cell populations, and without any prior assumptions as to disease mechanisms. This gives us unprecedented opportunities to find novel information, without the restrictions of attempts to model the disease *in vitro*, and may give us clues about which gene products could have the most therapeutic impact if targeted.[9] The major limitation of genetic association studies is that they are constrained by the extent of naturally-occurring genetic variation, and are usually restricted to common variants. This approach may thus miss promising therapeutic targets for which few polymorphisms in the population have a substantial effect on gene expression or protein function, for example if the gene is essential to another cellular or developmental process, or to resistance to another disease; selection pressure will keep the frequency of functional variants low in the population.

Aside from hypothesis-driven testing of single genes based on prior knowledge of function and disease pathogenesis, a variety of methods are available to screen for genes essential to specific disease processes in cellular models at genome or sub-genome scale. These include screens based on gene knockdown via RNA interference, gene knockout with the CRISPR (Clustered Regularly-Interspaced Short Palindromic Repeats) / Cas9 system, or plasmid-based overexpression of sets of genes of interest. These techniques require a relevant cellular model, and will be unable to discover genes essential outside

the confines of that model: for example, CRISPR screening for host genes essential for viral infection in an *in vitro* epithelial infection model will be able to identify critical genes involved in viral replication or in the innate, cell-intrinsic interferon response, but will be unable to identify genes involved in the wider immune response, which could be equally important to disease pathogenesis and as therapeutic targets.[67] Integration of data from different sources, including genetic studies, *in vitro* screens and other sources such as transcriptomic studies, may therefore be necessary to give the fullest picture of which genes are involved in disease pathogenesis.[68]

1.4 Genetic risk factors for influenza susceptibility in humans

In Chapter 3 of this thesis, the potential biological relevance of a novel genetic association with severe influenza will be investigated. The balance of evidence suggests that, like other infectious disease, mortality from influenza is likely to be at least partly heritable, although the direct evidence may not be as strong as for some other infectious disease, due in part to the ubiquitous nature and low case fatality rate of the disease. The development of screening techniques as discussed in section 1.3.6, combined with other molecular virology approaches, has given us a rapidly expanding knowledge base on the host factors involved in IAV replication.[67,69] However, despite the evidence for a genetic contribution to risk of severe disease after infection with influenza A virus, and the long history of influenza research, our knowledge of specific host genetic variants that underlie variation in individual risk in human patients remains limited. To date there have been few genome-wide association studies of influenza severity or susceptibility, and the few reported have all had small sample sizes and limited statistical power.[70–73] The majority of the data we have come from candidate gene studies, with varying degrees of experimental validation, although in some cases the candidates have been informed by prior pilot genome-wide association studies. Rare inborn errors of immunity in individual clinical cases can also give us clues as to genetic causation, especially if supported by experimental functional validation of the variants discovered, although proving that any single defect is the primary cause of the phenotype is difficult in such cases. Genetic variation could affect infection susceptibility, i.e. the probability of becoming infected, or severity of disease after infection, phenotypes which will overlap to an extent and are variably distinguished in association studies, depending on case and control definitions. Influenza risk could be modulated by genes involved in the viral life cycle or in the immune response, modulating the efficiency of viral replication and/or clearance, or the degree of host tissue damage, for example related to an excessive inflammatory response.

The following section will review our current knowledge on specific genes implicated in influenza susceptibility and severity in humans in genetic studies. Genetic variants associated with influenza, and their underlying evidence base, are summarised in Table 1.1.

Table 1.1 – Human genetic variants associated with influenza susceptibility or severity, and their evidence base.

Gene	Lead variant	IAV Strain	Severity / susceptibility	Evidence type				Functional variant?	Refs
				GWAS	CGS	LOF	AM		
IFITM3	rs12252	H1N1, H3N2, H7N9, Seasonal	Severity > Susceptibility	×	✓	×	✓	Function unclear	74,75
	rs34481144	H1N1, H3N2, Unspecified	Severity	×	✓	×	✓	CTCF-binding / transcription regulation*	76
GLDC	rs1755609	H1N1, H7N9	Both	(✓)	✓	×	×	Unknown	77
TMPRSS2	rs2070788	H1N1, H7N9	Both	(✓)	✓	×	✓	In LD with enhancer variant rs383510, ↑*TMPRSS2**	78
RPAIN	rs8070740	H1N1	Severity	(✓)	×	×	×	In LD with other genes	71
ST3GAL	rs113350588/ rs1048479	H1N1	Severity	×	✓	×	×	Splicing effects?*	79
DDX58	p.R71H + p.P885S	H1N1	Severity	×	×	✓	×	Missense variants, reduce IFN induction*	80
TLR3	p.F303S, p.P554S, p.P680L	Seasonal, unspecified	Severity	×	×	✓	✓	Missense variants, reduce IFN / NFκB induction	81,82
	rs5743313	H1N1, H7N9	Severity	×	✓	×	✓	Unknown	83
	p.L412F	Unspecified	Severity	×	✓	×	✓	Missense, function unknown	84
IRF7	p.F410V, p.Q421X	H1N1	Severity	×	×	✓	✓	Missense affecting nuclear translocation, ↓ I/III IFN production*	85

Continued on next page

Table 1.1 – *Continued from previous page*

Gene	Lead variant	IAV Strain	Severity / susceptibility	Evidence type				Functional variant?	Refs
				GWAS	CGS	LOF	AM		
IRF9	c.991G>A, c.577+1G>T	H1N1, unspecified	Severity	×	×	✓	×	Truncated protein, ↓ ISG production*	86,87
IFNAR2	rs1131668	Seasonal	Severity	×	✓	×	✓	Missense, function unknown	88
UBXN11	rs189256251	H7N9	Susceptibility	(✓)	✓	×	×	Missense, function unknown	89
DHX33	rs3744714	H1N1	Severity	(✓)	×	×	×	Unknown, in LD with other genes	71
IL1A	rs17561	H1N1	Susceptibility	×	✓	×	✓	Missense, function unknown	90,91
IL1B	rs1143627	H1N1	Susceptibility	×	✓	×	✓	Promoter variant, ↓c/EBPβ affinity, ↑*IL1B*, ↓*IL1A**	90,91
IL1B	rs3136558	H1N1	Susceptibility	×	✓	×	✓	Unknown	90,92
TNFA	rs361525, rs1800629, rs1800750	H1N1, Seasonal	Severity > Susceptibility	×	✓	×	✓	Promoter variants, transcription regulation*	92–96
IL6	rs2066992	H1N1	Susceptibility	×	✓	×	✓	Unknown	92,97
IL10	rs1800872, rs1800896	H1N1, Seasonal	Both	×	✓	×	×	Promoter variants, ↓*IL10**	96,98
LGALS1	rs13057866, rs4820294, rs2899292	H7N9	Susceptibility	(✓)	×	×	×	Transcription regulation?	72

Continued on next page

Table 1.1 – *Continued from previous page*

Gene	Lead variant	IAV Strain	Severity / susceptibility	Evidence type				Functional variant?	Refs
				GWAS	CGS	LOF	AM		
SFTPA2	rs1965708, rs1059046	H1N1	Severity	×	✓	×	×	Missense, function unknown	99
SFTPB	rs1130866	H1N1	Both	×	✓	×	×	Missense, function unknown	100
HLA locus	Various, inconsistent	H1N1	Susceptibility	×	✓	×	×	Coding variants	89,101,102
KIR locus	Various	H1N1	Severity	×	✓	×	×	Coding variants	103
GATA2	Various	H1N1	Severity	×	×	✓	×	Loss of function*	104,105
CCR5	CCR5Δ32	H1N1	Severity	×	✓	✓	✓	Loss of function*	106–108
FCGR2A	rs1801274	H1N1	Severity	(✓)	×	×	×	Missense, function unknown	71
C8	rs1960384	H7N9	Susceptibility	(✓)	×	×	×	In LD with other genes	72
C1QBP	rs3786054	H1N1	Severity	(✓)	×	×	×	In LD with other genes	71
CD55	rs2564978	H1N1	Severity	(✓)	✓	×	✓	In LD with rs3841376, ↓*CD55**	70,73,83

Evidence types: GWAS - genome wide association study or large sub-genome-scale screen (e.g. whole exome sequencing); CGS - candidate gene study; LOF - individual cases of rare loss-of-function mutations; AM - animal model of gene function in influenza. (✓) indicates nominal significance below the genome-wide significance threshold for GWAS. 'Severity' indicates comparisons between severe and mild influenza cases, or between severe cases and population controls, while 'susceptibility' indicates comparisons between mild cases or cases of unspecified severity and healthy or population controls. * indicates that functionality of at least one variant has been demonstrated directly *in vitro*. Reported functions relate to the 'risk' allele/haplotype.

1.4.1 Genetic risk factors associated with the viral life cycle

Influenza A virus is a negative sense RNA virus with an eight-segmented single-stranded genome. Like all viruses, influenza relies on the host machinery for replication. The viral life cycle, and the steps in the cycle where identified host genetic variants may have an impact, are shown in Figure 1.5.

1.4.1.1 Influenza A virus replication

Viral entry starts with binding of the viral haemagglutinin (HA) to sialic acid on host cell surface glycoproteins. This triggers endocytosis, after which low pH inside the endosome triggers a conformational change in the HA to expose the residues necessary for fusion, and activates the M2 channel, leading to fusion between the viral envelope and the endosome membrane and release of viral ribonucleoprotein (vRNP) into the cytoplasm. Import of vRNPs into the nucleus occurs via the host's nuclear import machinery, with the aid of nuclear localisation signals on the viral nucleoprotein. To produce viral proteins, viral RNA (vRNA) is first transcribed to messenger RNA (mRNA), assisted by 'cap-snatching' endonuclease activity, a process whereby the viral polymerase cleaves the 5' cap of a host mRNA to use as a primer for transcription. The viral mRNA is exported to the cytoplasm for synthesis of viral proteins. The viral polymerase and nucleoproteins are imported back to the nucleus for viral genome replication and assembly of new vRNPs. To replicate the viral genome, negative sense viral RNA is first transcribed to positive sense complementary RNA (cRNA), in an unprimed process, by the viral polymerase, and then new vRNA is transcribed from the cRNA. After complexing with the viral proteins, vRNPs for the eight viral genome segments are exported to the cytoplasm by a mechanism involving viral NS2 (nuclear export protein) and M1, and are trafficked towards the plasma membrane. Viral envelope proteins (HA, neuraminidase and M2) are targeted to the endoplasmic reticulum after translation and then on towards the plasma membrane via the Golgi apparatus. Viral assembly and budding take place at specific regions of the apical plasma membrane, and the viral particles are released at the cell surface by cleavage of sialic acid residues by the viral neuraminidase. The viral HA has to be primed by cleavage into subunits HA1 and HA2 to enable its fusion function in the next round of the life cycle - this

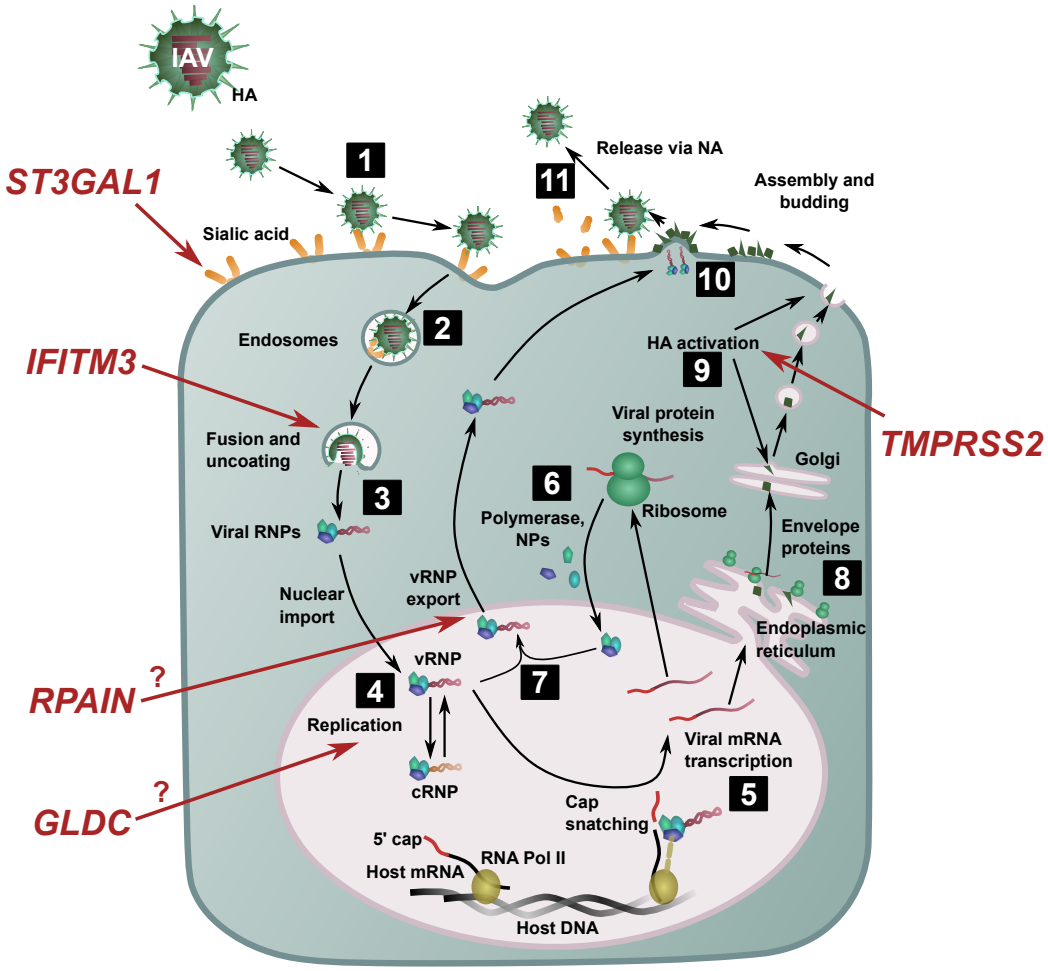

Figure 1.5 – Host genetic variants implicated in the Influenza A Virus life cycle. 1. Viral attachment via sialic acid residues. 2. Endocytosis. 3. Membrane fusion and vRNP release into the cytoplasm. 4. Nuclear import of vRNPs and replication of viral genome (via cRNA) by the viral RNA-dependent RNA polymerase. 5. Viral mRNA translation, using viral polymerase endonuclease activity to 'snatch' host mRNA 5' caps to prime viral mRNA transcription. 6. Translation of viral polymerase and nucleoproteins in cytoplasmic ribosomes, followed by re-import to the nucleus. 7. vRNP assembly and export. 8. Translation and processing of viral envelope proteins in the endoplasmic reticulum, for export via the Golgi. 9. HA priming by host enzymes. 10. Virion assembly and budding at the cell surface. 11. Virion release by NA-mediated cleavage of sialic acid residues. Host genes for which genetic variants have been associated with influenza susceptibility or severity are shown in red.

occurs either in the Golgi by protease enzymes such as transmembrane serine protease 2 (TMPRSS2), or at the plasma membrane (for example by human airway trypsin-like protease).[109]

1.4.1.2 *IFITM3* variants associated with influenza

Influenza severity or susceptibility has been associated with genetic variants in a small number of host genes that are involved in, or interfere with, steps in this life cycle. Of these, the strongest evidence base is for interferon-inducible transmembrane protein 3 (*IFITM3*). This was originally recognised as a viral restriction factor in an RNA interference screen[110], and has been subsequently confirmed to block viral entry via inhibition of fusion pore formation, and thus release of viral nucleoproteins into the cytoplasm after endocytosis.[111] An *in vivo* murine model has also confirmed the importance of this protein in influenza, as *Ifitm3-/-* mice show increased susceptibility to disease and increased viral loads.[74] A synonymous *IFITM3* variant, rs12252 allele C, which is common in Asian populations but rarer in Europeans, was enriched in hospitalised patients during the 2009-2010 H1N1 pandemic compared to ancestry-matched controls in the 1000 Genomes database, in a small candidate gene case-control study.[74] This variant was initially thought to modify a splice acceptor site, leading to production of a truncated protein, but subsequent investigation showed that production of the full length protein was unaffected by this polymorphism, and so the underlying mechanism is still unclear.[112] An *IFITM3* promoter variant associated with reduced expression, rs34481144 allele A, which is common in European populations but rare in East Asian populations, has also been associated with increased influenza severity in adults, IAV replication speed in experimental challenge of healthy adults, and mortality in paediatric influenza, in three separate cohorts.[76] The mechanism of action of this variant is clearer, as the variant decreases binding of the transcription factor IRF3 and increases CTCF binding to the promoter, possibly introducing an insulator function to disrupt expression regulation of surrounding genes. The risk variant is also associated with reduced accumulation of $CD8^+$ effector T cells.

A number of studies have attempted to replicate these findings in different populations. Three studies assessing rs34481144 found either no association with severe disease, or even a contradictory possible protective effect of the ‘risk’ A

allele under a dominant model in H1N1 cases in a Portuguese population.[113–115] Studies of rs12552 have had varying results, in part because very low minor allele frequencies have hindered attempts to demonstrate associations in populations with European ancestry. A recent meta-analysis of 12 studies confirmed a significant association between the C allele and severe influenza in both allelic and recessive genetic models, albeit with statistically significant heterogeneity between studies. For mild disease, the C allele was significantly over-represented overall compared to healthy or population controls, but only in white populations in subgroup analysis. The findings suggested that the locus has a greater effect on severity of disease after infection than on susceptibility to infection.[75]

1.4.1.3 Variants in other viral life cycle-associated genes

Associations with variants in other genes involved in the life cycle have been detected in single studies only. Variants increasing expression of *TMPRSS2*, which is essential for HA activation in H1N1 and H7N9 but not H3N2 strains, have been associated both with severity of pandemic H1N1 influenza and with susceptibility to H7N9 infection (comparing clinical cases to unaffected poultry workers with similar exposure risk). Enhancer variant rs383510 had *in vitro* functionality in a luciferase reporter assay and is a plausible candidate causal variant for this association.[78,116] Interestingly, a missense *TMPRSS2* mutation in LD with the influenza-associated variants (r^2 = 0.36, D' = 1.0 in a Southern Han Chinese population), has also been associated with COVID-19 severity. The non-risk allele reduces catalytic activity, suggesting an alternative (or additional) mechanism underlying the association with both diseases.[117] Genetic variants in the gene *GLDC*, encoding glycine decarboxylase, were associated with H1N1 severity and H7N9 severity in the same cohorts. The function of this gene in the influenza life cycle is not well defined, but increased lung expression with the risk haplotype could enhance viral replication either by increasing pyrimidine availability or by suppressing the interferon response.[77,118]

The enzyme ST3 beta-galactosidase alpha-2,3-sialyltransferase, encoded by *STGAL3*, is involved in transfer of sialic acid residues to galactose-containing substrates (via an alpha-2,3 linkage). As 2009 pandemic H1N1 strains of IAV show increased binding to alpha-2,3-linked sialic acid, in addition to the alpha-

2,6 binding that is more typical of human-associated strains, modification of this process could conceivably modulate viral entry for related H1N1 strains. A candidate gene study in a Brazilian cohort found an increased frequency of specific haplotypes of two synonymous SNPs in *STGAL3*, with putative functions in differential splicing, in patients dying of H1N1 influenza versus non-hospitalised or surviving hospitalised patients.[79] A SNP in the 3'-untranslated region of *RPAIN*, encoding RPA-interacting protein, has been associated with disease severity in a case-control study of patients with severe pneumonia after H1N1 IAV infection versus household controls, using a sub-genome scale SNP array. RPA-interacting protein is known to interact with viral NS2 and may be involved in nuclear export of vRNPs. However, the variant is in strong linkage disequilibrium with variants in two other genes (*DHX33* and *C1QBP*) that were similarly associated with disease severity in the same study, and so it is unclear which (if any) of these genes could be responsible for the phenotype effect.[71] Neither of these associations has as yet been replicated.

1.4.2 Genes involved in pathogen pattern recognition and the interferon response

1.4.2.1 Detection of IAV infection by pattern recognition receptors

A number of complementary detection mechanisms initiate the early innate response to IAV infection in epithelial and immune cells, especially in mononuclear phagocytes, after recognition of pathogen-associated molecular patterns (see Figure 1.6). Retinoic acid-inducible gene I (RIG-I) is among the most important of these in infected cells. This cytoplasmic receptor detects viral RNA in antiviral stress granules. Although the precise ligand specificity of RIG-I remains controversial, it has a preference for short double-stranded RNA products that are tri-phosphorylated at the 5' end - for single-stranded RNA viruses such as IAV this can include single strands with a looped secondary structure. Activation by ligand binding leads to association with the mitochondrial antiviral signalling protein complex. The resulting signalling cascade activates transcription factors including NF-κB and the interferon regulatory factors IRF3 and IRF7, leading to secretion of type I and type III interferons and other cytokines.[119,120]

Toll-like receptors (TLRs) 3 and 7 can also respond to viral nucleic acids after phagocytosis of virions or infected cells, without the requirement for active viral replication. TLR3 detects double-stranded RNA in endosomes, but may also recognise other unidentified RNA structures from IAV-infected cells phagocytosed by macrophages or other immune cells. It signals via the TRIF (TIR-domain-containing adapter-inducing interferon-β) adapter protein, inducing activation of IRF3 and NF-κB to trigger transcription of type I and type III interferons (interferon-α, -β and -λ) and cytokine genes respectively. TLR7 detects single-stranded RNA in endosomes of plasmacytoid dendritic cells and possibly in other immune cell types, and signals primarily via the MyD88 (myeloid differentiation factor 88) adapter protein to activate IRF7 and NF-κB. A third mechanism of virus recognition in a number of cell types, including macrophages, involves inflammasome activation in response to a combination of signals related to both active viral infection and cell stress. Inflammasome components, including NOD-like receptor NLRP3 and precursors to IL1B, IL18 and caspase-1, are expressed in response to TLR and cytokine signalling, but activation (to allow cleavage of precursors by caspase-1 to form active cytokines) requires the additional signals of IAV M2 ion channel activity in the Golgi and accumulation of IAV polymerase component PB1-F2 in lysosomes.[119]

Secreted interferons act in an autocrine or paracrine manner. On binding to their specific receptors, signal transduction via the JAK-STAT (Janus-associated kinase / signal transducer and activator of transcription) pathway leads to binding of phosphorylated STAT1/2 to IRF9 to form the interferon-stimulated gene factor 3 (ISGF3) complex. This translocates to the nucleus and binds to interferon-sensitive response elements in the genome, leading to expression of interferon-stimulated genes (ISGs), which have a range of effects to restrict viral entry and replication. Products of ISGs include proteins that restrict viral entry (e.g. IFITM proteins), destroy viral RNA (e.g. oligoadenylate synthase and ribonuclease-L) or reduce viral gene expression (e.g. zinc finger antiviral protein).

1.4.2.2 Genetic variants in pattern recognition receptors and interferon regulatory factors

Genetic susceptibility to influenza has been associated with variation in the pattern recognition receptors themselves, elements of their signalling pathways,

and in the end products (cytokines and ISGs). A pair of pathogenic variants in the *DDX58* gene (encoding RIG-I) has been identified by whole exome sequencing in a single severely affected patient without underlying risk factors. These variants affected the protein sequence in the RNA-binding and caspase-recruitment domains, and reduced type I interferon induction in response to RIG-I ligands *in vitro*.[80] Similarly three heterozygous missense mutations in *TLR3*, associated with reduced interferon and NF-κB induction *in vitro*, have been reported in individual children with influenza-associated encephalopathy or severe influenza pneumonitis.[81,82] At the population level, missense variant p.L412F and intronic SNP rs5743313, both of uncertain function, have been associated with influenza severity in single reports, without successful replication as yet.[83,84,121] The role of TLR3 in pathogenesis of severe influenza is uncertain given discrepant results in mouse models with different challenge strains.[122]

Other rare inborn errors of immunity have been reported in *IRF7* and *IRF9*. A child developing acute respiratory distress syndrome after primary H1N1 infection was found to have two simultaneous heterozygous loss-of-function mutations in *IRF7*, which caused an almost complete loss of function, with severely reduced production of type I and type III interferons. The two mutations were both missense mutations that seemed to interfere with protein function in different ways, by preventing correct nuclear import or export of the transcription factor.[85] Two examples of homozygous *IRF9* mutations have been reported in young children (of consanguineous parents) with influenza pneumonitis. Both caused a loss of function by aberrant splicing and production of a truncated protein, severely reducing (but not completely abolishing) induction of interferon-stimulated genes *in vitro*.[86,87] The effects of loss of IRF function on other viruses seemed to be variable in these cases: whereas some individuals suffered from recurrent severe viral infections or illness following live attenuated vaccine administration, others had been able to clear common viral infections normally.[87]

1.4.2.3 Variants in interferon signal transduction mediators and interferon-stimulated genes

Severe disease with loss of IRF function indicates that the type I and type III interferon responses are indispensable for control of IAV infection. Consequently, other components of the interferon signalling cascade, such as the interferon re-

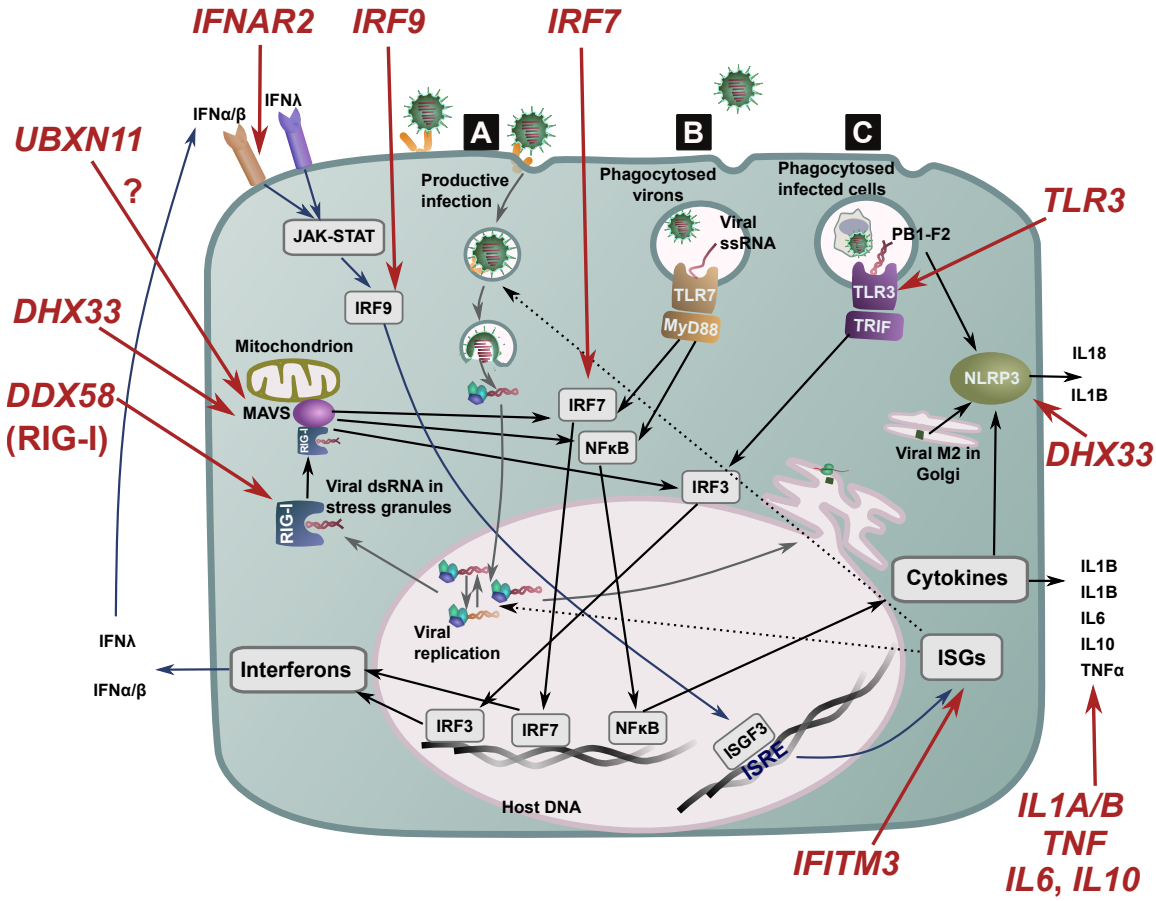

Figure 1.6 – Host genetic variants implicated in pathogen molecular pattern recognition and the interferon response. Different mechanisms of pathogen recognition operate with productive infection (A, grey arrows), versus non-productive uptake of virions in endosomes (in which fusion and vRNP release does not occur) (B), or uptake and destruction of infected cells by phagocytes (C). Blue arrows indicate the interferon response. MAVS: mitochondrial antiviral signalling complex. ISGs: interferon-stimulated genes. ISRE: interferon-sensitive response element. Host genes for which genetic variants have been associated with influenza susceptibility or severity are shown in red.

ceptors and elements of the JAK-STAT pathway would be expected to influence influenza severity. Loss-of-function mutations in these signalling components have been implicated in other viral diseases, but evidence in influenza is currently limited. An association has been reported between missense rs1131668 variant in *IFNAR2* (interferon alpha and beta receptor subunit 2) and hospitalisation due to seasonal influenza in a single study in a small European cohort.[88]

Given the importance of the interferon pathways, interferon-stimulated genes (ISGs) are expected to be important in determining disease outcome. Interferon-stimulated gene products include RIG-I, and IFITM3 as discussed above. There is as yet limited evidence for a role of genetic variants in other ISGs in influenza severity in humans. A loss-of-function mutation in ISG *Mx1*, which can restrict viral replication, is a key determinant of IAV susceptibility in laboratory mice[123], but there have been no reports to date of a genetic association with influenza for the homologous *MX1* gene in humans.

Other influenza-associated genes with possible roles in pathogen recognition and the interferon response, but for which the exact mechanism of action is unclear, include *UBXN11* and *DHX33*. A rare missense variant in *UBXN11* has been associated with H7N9 susceptibility. Type I interferon production was reduced in patient cells carrying the risk allele. The authors hypothesised that the gene product could have similar functions to the better known ubiquitin-binding protein UBXN1, which blocks RIG-I signalling by preventing assembly of the mitochondrial antiviral signalling protein complex.[89] DEAH-Box Helicase 33 (encoded by the gene *DHX33*) is a protein that has only been studied to a limited extent in the context of influenza, but may be involved in activation of both the inflammasome and the mitochondrial antiviral signalling complex after binding viral RNA products cleaved by RNAse-L.[124] An intronic variant in this gene was one of the cluster of variants in high LD on chromosome 17 (along with *C1QBP* and *RPAIN*) associated with severe influenza in a GWAS in a Mexican population, but could be merely a proxy for a variant in one of the other genes.[71]

1.4.3 Genes involved in other aspects of the immune response to infection with influenza A virus

1.4.3.1 Overview of the integrated innate and adaptive immune responses to IAV infection

Beyond the interferon response, a wide range of innate and adaptive processes work in concert to prevent infection with influenza A virus, or to eliminate infection once established.[125] Mononuclear phagocytes play key roles in many of these processes. Genetic associations with genes involved in the broader

immune response are shown in Figure 1.7.

Initial barriers to infection include airway mucus, neutralising antibodies (from previous exposure) such as immunoglobulin A in nasal secretions, and lectins. Lectins are soluble pattern recognition receptors which bind to glycoproteins on the virion surface, neutralising the virus and enhancing phagocytosis and complement activation.[126] Virions can be phagocytosed by a variety of immune cells including alveolar macrophages, which play a key role in airway surveillance and may help increase resistance of adjacent epithelial cells to infection[127], or neutrophils, which can contribute to host damage and disease severity as well as aiding pathogen clearance.[128–130] Cytokines and chemokines expressed in response to pathogen-associated molecular pattern receptor activation, particularly in epithelial cells and macrophages (see section 1.4.2.1), help to recruit and activate neutrophils and other immune cells.

The complement system interacts with both the innate and adaptive immune responses, and can be triggered by bound antibodies (the classical pathway), binding of complement component C3 to a microbial surface (the alternative pathway), or via lectin binding to carbohydrate surfaces on microbes (the lectin pathway). The resulting signalling cascades converge, and produce a number of active components including C3a and C5a, which have chemotactic and immunomodulatory functions, and the membrane attack complex which can lyse infected cells. Beneficial effects of complement in influenza include virus neutralisation (which can be mediated by naturally-occurring non-specific IgM even without prior exposure), promotion of phagocytosis (by opsonisation and via activation of complement receptors) and augmentation of virus-specific B and T-cell responses. While complement activation may contribute to host damage, deficiency of complement components (especially C3) or complement receptors reduces viral clearance and increases mortality in mouse models of influenza.[131–133]

Before development of the adaptive response, infected cells can be lysed by natural killer (NK) cells, which bind to viral HA on the cell surface via the NK-p44 or NK-p46 receptors. Other populations of innate-like lymphoid cells, such as mucosal-associated invariant T cells, natural killer T cells and gamma-delta T cells, are also involved in cell clearance or immune activation via cytokine secretion.[134–136]

The adaptive immune response is initiated by antigen presentation via the major histocompatibility complex (MHC) in dendritic cells and other antigen presenting cells, which migrate to local lymph nodes after antigen exposure. Viral peptides are presented to $CD8^+$ T cells via MHC class I (primarily for endogenous or cross-presented antigens), or to $CD4^+$ T cells via MHC class II (primarily for exogenous antigens). When the T-cell receptor recognises its target antigen mounted on the MHC, naïve T cells undergo clonal expansion and activation to an effector phenotype, with the aid of additional costimulation via receptors such as CD28. $CD8^+$ cells become cytotoxic T lymphocytes, which migrate to the lung (via downregulation of chemokine receptor CCR7 and upregulation of CXCR3 and CXCR4), and destroy infected cells by mechanisms including induction of apoptosis by perforin/granzyme or Fas-ligand binding. $CD4^+$ T cells can differentiate into a number of effector phenotypes, particularly T helper type I (Th1) cells in the context of IAV infection, which secrete interferon-gamma (IFNγ) to activate macrophages and promote cytotoxic T-cell function, but also other phenotypes such as regulatory T cells to prevent excessive immune activation. B lymphocytes are activated later in the course of disease and play a greater role in prevention of reinfection. Antigen binding to the B-cell receptor, without the need for presentation on MHC, triggers clonal expansion and antibody secretion. For full activation, B cells require help from T cells (especially Th2 and T follicular helper cells) and pathogen-associated molecular pattern receptor activation, and conversely can efficiently present antigen to T-cells via MHC II. Functions of secreted antibodies include virus neutralisation, opsonisation with enhancement of phagocytosis, and promotion of antibody-dependent cell-mediated cytotoxicity. B cells and both T-cell types can gain a memory phenotype to enable a rapid response to reinfection.[125,137]

1.4.3.2 Cytokine gene variants

Cytokines secreted as a result of activation of pattern recognition receptors, especially from mononuclear phagocytes, play key roles in coordinating the ensuing innate and adaptive responses. A number of candidate gene studies have examined polymorphisms in cytokine genes including *IL-1A*, *IL-1B*, *IL-6*, *IL-10* and *TNFA*. Interleukin-1 (A and B) and TNFα are involved in the development of acute lung damage following IAV infection, but also have beneficial effects on

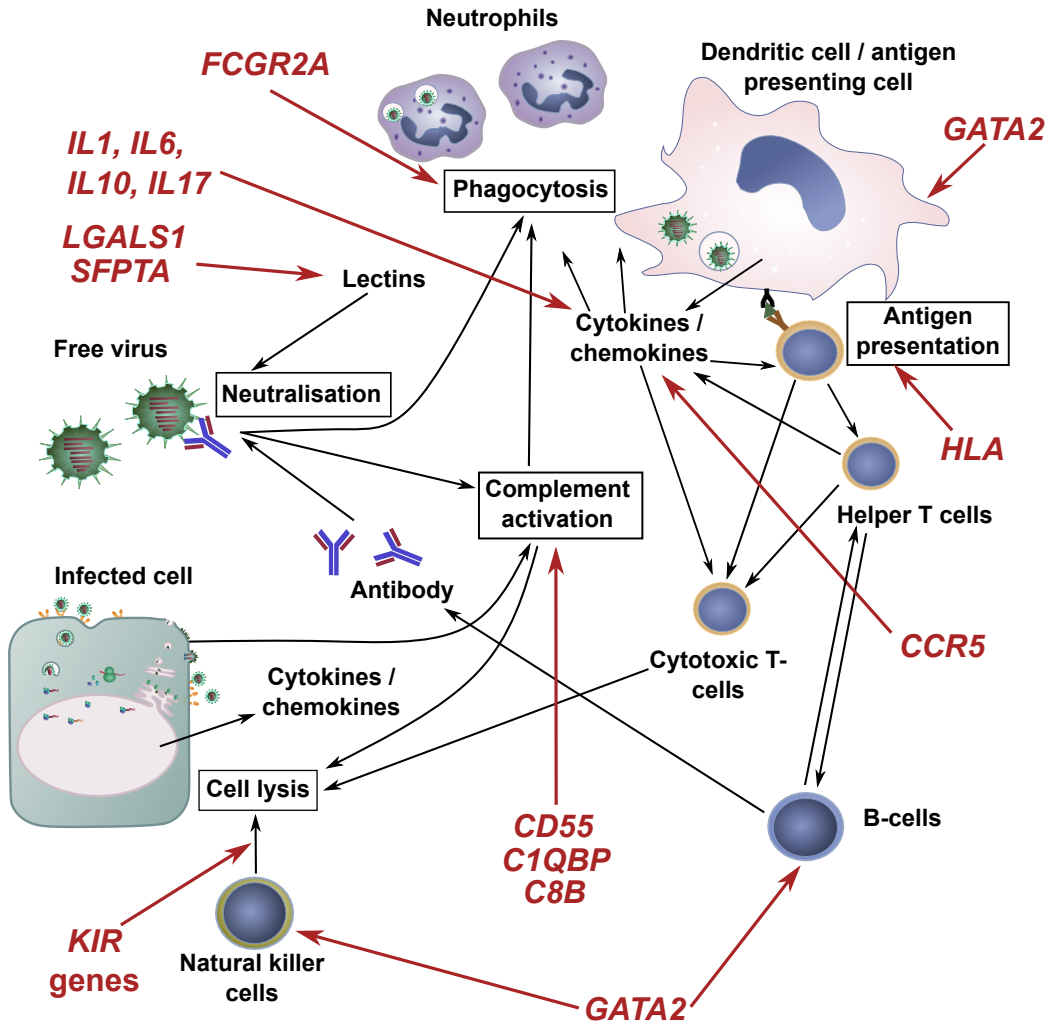

Figure 1.7 – Host genetic variants implicated in wider immune response to influenza A virus. Host genes for which genetic variants have been associated with influenza susceptibility or severity are shown in red.

viral clearance or immune regulation, and their absence exacerbates disease in mouse models.[90,93] Circulating IL-6 concentrations tend to be increased in severe versus mild cases of influenza[92], but this cytokine also seems to be necessary for viral clearance in mouse models.[97] The anti-inflammatory cytokine IL-10 may help to limit a potentially damaging excessive inflammatory reaction, but may also have detrimental effects in influenza by limiting protective adaptive immune responses.[98] Single unreplicated associations with disease susceptibility (but not severity) have been reported for intronic, promoter and missense

variants in *IL1A*, *IL1B* and *IL6* (see Table 1.1), with a possible protective effect of *IL1A* or detrimental effect of *IL1B*, while other studies have found no effect of variants in these genes.[91,92,96] Two promoter polymorphisms associated with lower promoter activity and *IL10* expression, rs1800872 allele C and rs1800896 allele A, were associated with hospitalisation due to 2009 pandemic H1N1 infection[96], suggesting a protective effect of IL-10 in natural infections, in contrast to effects predicted from knockout mouse models.[98]

For the *TNFA* gene, three promoter variants have been assessed in a number of different populations, with limited replication. The minor allele of variant rs361525, a variant which increases promoter activity and has been previously associated with sepsis and other inflammatory disorders, was associated with increased risk of IAV-associated pneumonia in a Greek population.[95,138] The association was replicated in two Mexican cohorts, either on the basis of increased risk of influenza versus population controls, or in terms of increased disease severity[92,96], with an additional significant effect of promoter variant rs1800629[96]. It was not, however, replicated in a study of influenza-associated deaths in children and young adults in the USA, which found no differences compared to ancestry-matched controls except for a more rapid course of disease for promoter variant rs1800750.[94]

Together, these results provide some limited support for a role of cytokine gene variants in disease development. There is some inconsistency as to whether the effects are primarily on susceptibility to infection or predisposition to severe disease, the latter of which could be more consistent with expectations based on animal models. This distinction is often blurred in case-control studies, as for example, if cases are defined on the basis of a positive test result without regard to disease severity, severe cases could still be over-represented due to biases in the population of patients presenting for testing.

1.4.3.3 Genetic variants in lectins and surfactant proteins

Lectins that can bind IAV envelope glycoproteins include surfactant proteins A and D, mannose-binding lectin and galectin-1. No association has been found between genetic variants in the mannose-binding lectin *MBL2* gene and influenza.[94,99] Galectin-1, encoded by *LGALS1*, is upregulated in the lung in re-

sponse to IAV infection and has both virus-neutralising and anti-inflammatory effects.[139] A number of variants around the *LGALS1* promoter have been associated with increased susceptibility to H7N9 influenza and reduced *LGALS1* expression *in vitro*. While a precise causal variant was not identified, this is consistent with the expected protective effect of galectin-1. This is one of the strongest associations for an unbiased genome-wide study of influenza to date, albeit in a single study only.[72] Two missense variants in *SFTPA2*, encoding surfactant protein A2, have been associated with disease severity in pandemic 2009 H1N1 patients, as assessed by parameters including need for mechanical ventilation and development of acute respiratory distress syndrome, but no association has been found with genetic variants in other lectin-like surfactant proteins.[99,100] A missense mutation in the unrelated *SFTB* gene, which encodes a non-lectin surfactant protein, has also been associated with H1N1 severity (comparing severe to mild cases) and susceptibility (comparing cases to population controls) - an effect on severity could plausibly be related to its function in reduction of lung surface tension and maintenance of alveolar inflation.[100]

1.4.3.4 Genetic variants affecting adaptive immune functions

Despite the importance of the adaptive immune response in clearing IAV infection and in preventing re-infection, surprisingly few genetic associations have been found for genes involved in adaptive immune cell function. There is little evidence of a predisposition to severe influenza in people with severe primary immunodeficiencies such as severe combined immunodeficiency or agammaglobulinaemia, conditions associated with a lack of T cells or B cells respectively. One possible exception identified to date is inherited deficiency of the transcription factor GATA2, which causes deficiency in a number of immune cell types including NK cells, B cells, monocytes and dendritic cells.[104] People affected by this condition have increased susceptibility to a range of viral, bacterial and fungal infections, and death from influenza has been reported in some individuals.[104,105] The complex nature of this phenotype makes it difficult to infer which aspects of the immune deficiency contribute most to the observed influenza susceptibility, and consequently this condition provides little illumination as to critical elements of the immune response to IAV infection in the wider population.

Efficiency of antigen presentation for different antigens is determined in part by haplotype at the *HLA* locus. Certain combinations of *HLA* alleles and IAV peptides are known to efficiently stimulate $CD8^+$ T cells in a manner that is conserved across IAV strains, such as the combination of HLA-A*02:01 with IAV M protein residues 158 - 66.[140] Conversely, other alleles such as the HLA-A*24 allele group have consistently low targeting efficiency at least for H1N1 strains, and population allele frequency has been correlated with country-level H1N1 pandemic mortality.[141] Case-control studies of 2009 pandemic H1N1 influenza or H7N9 influenza cases in Mexican, Indian and Chinese populations have found over-representation or under-representation of certain allele groups and haplotypes in influenza patients.[89,101,102] There is remarkably little consistency between studies, and reported results include both associations consistent with *in vitro* observations (e.g. reduced HLA-A*02 efficiency in Indian cases) and counterintuitive observations such as reduced HLA-A*24 frequency in Mexican influenza cases. Differences between studies could arise both from methodological and population differences. While a documented genetic association could facilitate prediction of individual risk within a population, variability between populations limits external validity.

Haplotype at the *HLA* locus interacts with genetic variation at the killer-cell immunoglobulin-like receptor (*KIR*) locus to determine affinity of NK-cell interactions with MHC molecules, in which binding of predominantly inhibitory KIRs on the surface of NK cells helps to prevent inappropriate destruction of healthy 'self' cells. Similarly to *HLA* genes, *KIR* genes are encoded by a hypervariable locus, with differences between individuals not only in polymorphisms within genes, but also in presence or absence or copy number of specific receptor genes. A study of *KIR* genes in 2009 pandemic H1N1 influenza patients compared to blood donor controls in a Mexican population found over-representation of specific genes (such as inhibitory receptor gene *KIR2DL5*), gene groups and haplotypes in influenza patients, and an apparent increase in haplotypes bearing greater numbers of inhibitory receptors. This could reflect varying ability of natural killer cells in patients with these genotypes to bind to and respond to MHC class I molecules presenting IAV antigens.[103] As for the *HLA* locus, extrapolation to other populations is difficult, and there was little overlap with *KIR* subtype enrichment in influenza patients in a Canadian First Nations population.[142]

Chemokine receptor 5 (CCR5) is principally expressed in immune cells includ-

ing macrophages, dendritic cells and T cells, and is likely to be involved in cell migration. A relatively common deletion mutation, CCR5Δ32, which has a carrier frequency of approximately 11% in Europeans, renders this receptor non-functional. Homozygosity for this loss-of-function mutation has been observed in fatal cases of influenza without pre-existing comorbidities.[107] Although there has been some conflicting evidence in small underpowered population studies[143,144], mortality from H1N1 influenza was increased in CCR5Δ32 carriers in a Spanish population.[108] Increased mortality with CCR5 deficiency would be consistent with observations of increased influenza severity, associated with aberrant macrophage function and excessive pulmonary macrophage accumulation, in *Ccr5-/-* mice.[106]

Fc-gamma receptors on mononuclear phagocytes and granulocytes play an important role in antibody-mediated immunity, binding to the Fc (fragment crystallisable) region on the stalk of bound immunoglobulin-G antibodies to facilitate phagocytosis or antibody-dependent cell-mediated cytotoxicity. A missense mutation in the *FCGR2A* gene, which encodes one such receptor, was one of the variants most strongly associated with severe H1N1 IAV-associated pneumonia (compared to household controls) in a small GWAS.[71] This was not corroborated in a larger candidate gene study, which found no difference in allele frequencies between hospitalised and non-hospitalised H1N1 influenza cases.[145] This variant has, however, been associated with a number of other inflammatory conditions such as ulcerative colitis and rheumatoid arthritis, and thus a biologically relevant role in modulation of inflammation is plausible.

1.4.3.5 Genetic variants in *CD55* and complement system genes

A number of studies have found evidence for an effect of genetic variants in genes encoding complement or complement regulatory proteins on influenza susceptibility or severity. An intronic variant in the *C8B* gene, encoding a component of the complement membrane attack complex, was the strongest single-SNP association in a small GWAS of H7N9 infection susceptibility.[72] Similarly, an intronic variant in the *C1QBP* gene, associated with increased expression, was the SNP most strongly associated with severe H1N1 influenza pneumonia in a Mexican GWAS.[71] This gene encodes complement component 1 Q subcomponent-binding protein, a cell surface receptor that binds to comple-

ment component C1q and suppresses activation of the classical complement pathway, although the receptor may also have other pro-inflammatory functions such as promotion of macrophage chemotaxis - both could plausibly increase influenza severity. However, as discussed above, the effect of this variant could not be distinguished from these in neighbouring genes in high LD.

The strongest evidence for a genetic association with a complement-associated gene is for that encoding the complement regulatory receptor CD55 (also known as Decay Accelerating Factor or *DAF*). This receptor degrades the classical and alternative C3 convertases to prevent activation of the downstream complement cascade. Additional recognised functions include regulation of T-cell-mediated immunity via complement-dependent and independent mechanisms, and a possible role in leukocyte adhesion and migration.[146–148] An association between promoter variant rs2564978 (genotype T/T) in *CD55* and influenza severity was first detected in study of severe versus mild cases among 2009 pandemic H1N1 patients in a Chinese population, in which results of a small pilot GWAS in 51 patients were confirmed by targeted genotyping of an additional 374 patients.[70] The risk genotype for this variant, for which the genetic evidence fits best with a recessive effect, was associated with lower *CD55* expression in patient monocytes (but not T cells), although a luciferase assay suggested that the true causal variant is most likely to be 21-nucleotide indel rs3841376, with which it is in complete LD.

The association between rs2564978 and influenza has been replicated to a certain extent in two subsequent studies. A small GWAS comparing 49 severe and 107 mild influenza cases in a Spanish population, while not evaluating rs2564978 directly, provided indirect supportive evidence, with a significant ($p < 0.01$) association of a nearby variant in strong linkage disequilibrium (r^2=0.88 and D'=1.0).[73] In a candidate gene study in 275 Chinese patients diagnosed with either H7N9 or 2009 pandemic H1N1 influenza, the T/T genotype was associated with increased risk of hospitalisation with H1N1 infection, but not with acute respiratory failure with either strain, nor with mortality across both strains.[83] Although the evidence in each individual case is not especially strong, this is among the most replicated of genetic associations with influenza, and one that to date has been exclusively associated with severity, rather than susceptibility to infection. This could be consistent with a role of CD55 to protect host cells from complement attack, as suggested by the authors first reporting the as-

sociation, who demonstrated that an anti-CD55 neutralising antibody enhanced complement-mediated lysis of A549 cells in an *in vitro* model.[70] This assay was flawed, however, as the results could have been attributable solely to greater activation of complement by bound antibody compared to a non-targeting control antibody. In a mouse model, *Cd55-/-* mice showed reduced disease severity after IAV challenge, an opposing effect to that expected based on observations in humans.[149] This protective effect was IAV strain-specific, depending at least in part on the HA and NA segments, and was dependent on the presence of C3, although counterintuitively C3a concentrations were reduced in *Cd55-/-* mice. Viral NA can furthermore modulate CD55 via cleavage of α-2,6-linked sialic acid residues, which appears to attenuate CD55 function.[149] Modulation of immune responses by altered complement activation, and non-complement-mediated functions of this multifunctional receptor, could have complementary or opposing effects on the host response, and both remain plausible explanations for the effect of genetic variation in *CD55* on influenza severity.

1.4.4 Summary of evidence for genetic associations with influenza

The majority of evidence for specific genetic associations with influenza comes from candidate gene studies and individual cases of inborn errors of immunity, with their associated limitations and biases. The few small GWAS reported have been underpowered, although in some cases they have generated useful hypotheses for testing in larger genotyped cohorts. Determinants of genetic susceptibility to disease are likely to range from rare variants with large effects to common variants with small effects - both ends of the spectrum can be challenging to detect with small sample sizes, and consequently lack of evidence for association in reported GWAS results should not be interpreted as evidence of absence of effect. Lack of replication of the majority of genetic associations with influenza to date does however mean that confidence in most findings remains limited. Even for the genes with the strongest evidence base, such as *IFITM3*, *TLR3*, *CD55* and *TNFA*, findings have not been completely consistent, and plausible causal variants have only been identified in a minority of cases. The only way to increase confidence in these findings will be either to replicate and meta-analyse results in multiple populations, as has been done for *IFITM3*,

or to conduct more unbiased genome-wide studies with increased sample sizes and adequate power.

The spectrum of genetic variants associated with influenza shows a bias towards elements of the innate immune response, including pattern recognition receptors, interferon signalling and inflammatory cytokines. This reflects biases in selection of genes for candidate gene studies for which a role in viral resistance or disease pathogenesis is already known. These biases limit the capacity of such studies to provide us with new information about disease pathogenesis. The spectrum of genes implicated is notably different from those implicated by genetic perturbation screens (such as genome-wide CRISPR knockout or RNA interference screens) for host dependency or restriction factors. While interferon-related genes are also over-represented in such screens, the predominant host factors to emerge are those related to aspects of viral entry, replication, trafficking within the cell and export.[67] Few genes associated with adaptive immunity have been associated with influenza by any method, but it is difficult to establish if this reflects lesser impact on disease severity, or whether this merely relates to selection bias and difficulty in studying adaptive processes in model systems. Unbiased genome-wide studies, which have notably implicated genes with a comparatively broad range of functions (notwithstanding some uncertainty as to which gene in a locus underlies the association), will be the best way to address this question.

1.4.5 Comparison to genetic associations with COVID-19

The impact of the COVID-19 pandemic on global health and the global economy has led to an unprecedented investment into research focused on a single infectious disease. Within a year of the pandemic, data on genetic associations with SARS-CoV2 had far exceeded that available for influenza. Multiple GWAS studies have been performed, including high quality studies with sufficient statistical power to detect relatively small effects, and meta-analyses are also available for associations with disease severity or susceptibility to infection.[9,150] These data have been complemented by a range of candidate gene studies and cases of inborn errors in immunity, as well as a range of genetic perturbation screens, transcriptomic studies, and *in vivo* and *in vitro* studies to address specific aspects of disease pathophysiology.[68,151] SARS-CoV2 and IAV are both single-

strand RNA viruses, but while IAV has a negative-sense genome with nuclear replication, SARS-CoV2 has a positive-sense genome with cytoplasmic replication. Infection with both viruses can lead to acute respiratory distress syndrome in severe cases, but there are important differences between the diseases, such as increased tropism of SARS-CoV2 for endothelial cells, and a prominent T-cell lymphopaenia with severe COVID-19.

Of the genes with the strongest evidence for involvement in COVID-19 on the basis of GWAS results, a number are involved in interferon signalling and have evidence of involvement in influenza. For example, there is evidence of an association with *IFNAR2* for both diseases, and the reported associated variants for the two are in high LD. The locus containing the interferon-stimulated *OAS* genes (the products of which activate RNase L) is strongly associated with severe COVID-19, and these genes are also known to be involved in the response to influenza, although no genetic association with influenza has yet been demonstrated.[9] A similar but broader spectrum of inborn errors of immunity has been reported, with predicted loss-of-function mutations in approximately 3.5% of patients, including mutations in genes associated with influenza (*TLR3*, *IRF7*, *IRF3* and *IFNAR2*).[10] Overlap in candidate genes studies reflects bias in gene selection: associations have been reported with COVID-19 for a number of influenza-associated genes such as *IFITM3*, *TMPRSS2* (which activates the SARS-CoV2 spike protein as well as IAV HA) and *TNFA*, as well as others which could plausibly have roles in both diseases (e.g. *MX1*, *OAS1*, *MBL*).[117,151] As is the case for influenza studies, there has been notable inconsistency in results of candidate gene studies, with few genes implicated by more than one study - this is especially true of the ten studies examining the *HLA* locus, in which no single allele or haplotype association has been replicated.

It is clear from the above similarities, especially those derived from unbiased GWAS results, that there is a 'common framework' of host genes important to the response to infection with the two viruses, and probably with other RNA viruses, principally involving the interferon response. Where evidence is weaker or absent for influenza, it is difficult to exclude the effect of lesser experimental power, but some associations could reflect genuine differences in disease pathophysiology - for example, the chromosome 3 locus strongly associated with COVID-19, which contains a number of chemokine receptor genes as well as other genes of less well characterised function, could reflect the relative im-

portance of key immune cell types depending on these chemokines for effective migration. Similarly, an association with *DPP9*, which is associated with pulmonary fibrosis, could reflect mechanisms underlying pulmonary damage.[9] Genes associated with influenza but not with COVID-19, such as the surfactant proteins, could likewise reflect differences in pathophysiology (for example in this case, differing affinity of lectin binding to the surface of different viruses), although given the lack of replication, spurious results cannot be ruled out. Better understanding of the biological functions underpinning such genetic associations will help us to distinguish the real from the spurious, to rationalise extrapolation between diseases, and to prioritise therapeutic targets.

1.5 Aims and hypotheses

In this thesis, I have taken two contrasting approaches to address the biological relevance of specific genetic associations between complex human disease phenotypes and variants in mononuclear phagocyte-associated genes involved in immune cell interaction and regulation.

First, I investigate the biological relevance of a novel genetic association between adhesion G-protein-coupled receptor CD97 and severe influenza, by exploring the impact of the gene on influenza A virus infection in an animal model. I hypothesise that CD97 is required for an appropriate immune response to IAV infection, and that CD97 deficiency will predispose to severe disease either by reducing the efficiency of viral clearance or by exacerbating inflammation and consequent host tissue damage. The principal aims of these investigations are as follows:

- To investigate whether CD97 deficiency modulates influenza severity in a murine model.
- To investigate immunologic dysfunction associated with *CD97* deficiency *in vitro* and *in vivo*.

Secondly, I address the problem of investigating causation at the single variant level within a disease-associated linkage disequilibrium block. Using a novel association between macrophage-associated gene *SIRPA* and schizophrenia for

proof-of-concept, I have developed a method to interrogate potential regulatory variants using a parallelised CRISPR screen. I hypothesise that targeted mutagenesis of non-coding variants can differentiate true regulatory variants from non-functional variants that are associated only via linkage disequilibrium. The principal aims of this project are as follows:

- To develop methodology for CRISPR mutagenesis screening of cis-acting regulatory variants in a linkage disequilibrium block.
- To prioritise candidate causal variants in a locus associated with *SIRPA* expression and schizophrenia.

Chapter 2

Materials and Methods

2.1 Animals

Wild type C57BL/6J mice were obtained from Charles River Laboratories. *Cd97* knockout mice, strain B6;129P2-*Adgre5*^{tm1Dgen}/J, created by Deltagen Inc. on a 129P3/OlaHsd background, were obtained from the Jackson Laboratory (stock no. 005788). The strain was back-crossed to wild type C57BL/6J mice for three generations by the strain creators, and a further six generations in our laboratory. Mice were housed in individually ventilated cages, in groups of two to five per cage. All animal work was performed in accordance with the Animals (Scientific Procedures) Act 1986 under a UK Home Office Project Licence, subject to local ethical review.

2.2 Virus strains

Influenza A virus H1N1 strain A/WSN/1933(H1N1) (A/WSN/33) was a gift from Dr D. Jackson, University of St. Andrews, and was propagated in MDCK cells. 2009 pandemic H1N1 strain A/England/195/2009(H1N1) (A/Eng/195), derived using an established reverse genetics system (sequence accession numbers GQ166654 - GQ166661)[152], was propagated and kindly provided by Yvonne Ligertwood. Reverse genetics plasmids for 2009 pandemic strain A/California/04/2009(H1N1) (A/Cal/04/09; accession numbers FJ966079 - FJ966086)

were a kind gift from Dr. Daniel Perez, University of Georgia.[153] To rescue the A/Cal/04/09 strain, pDUAL plasmids (which transcribe mRNA from an RNA Polymerase II promoter and negative-strand vRNA-like RNA from an RNA Polymerase I promoter) containing the eight viral segments were transfected into HEK293T cells (ATCC®) using Lipofectamine 2000 (Thermo Fisher Scientific). 250 ng of each plasmid was diluted in a total of 200 µl Opti-MEM® Reduced Serum Medium (Thermo Fisher Scientific) with 4 µl Lipofectamine transfection reagent, and added to 10^6 cells in 1 ml antibiotic-free Dulbecco's Modified Eagle's Medium (DMEM, Sigma Aldrich) containing 10% fetal bovine serum (FBS) and 2 mM L-alanyl-L-glutamine dipeptide (Gibco™ GlutMAX™, Thermo Fisher Scientific), in one well of a six-well plate, with an additional negative control transfection omitting Segment 1. The cells were cultured at 37℃ and 5% CO_2 overnight, after which the growth medium was changed to 2 ml serum-free DMEM supplemented with 0.14% w/v bovine serum albumin (BSA) fraction V (Gibco™, Thermo Fisher Scientific) and 1 µg/ml N-tosyl-L-phenylalanine chloromethyl ketone (TPCK)-treated trypsin (Worthington Biochemical Corporation). After a further 48 hours, the supernatants were harvested and used to infect confluent MDCK-SIAT1 cells (a gift from Prof. John McCauley, The Francis Crick Institute) in a T175 flask, in a total volume of 15 ml of the same serum-free growth medium. Forty-eight hours later, the supernatant was harvested and centrifuged at 1000*g* for 5 minutes at 4℃ to remove cell debris. This viral P1 stock was aliquoted and stored at -80℃. Stock titres for all viruses were determined by plaque assay.

2.3 Viral titration by plaque assay

Plaque assays for viral titration were performed either in MDCK cells (ATCC® CCL-34), or for 2009 pandemic-derived H1N1 strains in MDCK-SIAT1 cells, which have been modified to stably express the human α-2,6-sialyltransferase gene. This modification increases the relative proportion of α-2,6-linked sialic acid receptors, the preferred binding substrate of human H1N1 viruses, improving efficiency of infection with these strains.[154] The cells were cultured in DMEM Complete medium (DMEM with 10% FBS, 2 mM L-alanyl-L-glutamine dipeptide, 100 U/ml penicillin and 100µg/ml streptomycin), at 37℃ and 5%CO_2. Ge-

neticin (Gibco™ Geneticin™, Thermo Fisher Scientific) was added at 1 mg/ml for MDCK-SIAT1 cells as a selection agent. Cells were seeded onto 6-well plates at approximately 2 x 10^6 cells per well, to achieve confluence 24 hours later. Serial 10-fold dilutions of virus stocks, culture supernatants or biological fluids were performed in serum-free DMEM Complete medium, up to 1 in 10^6 or 1 in 10^7 depending on sample type. Culture medium was aspirated from the confluent cells. The cells were washed once in phosphate buffered saline (PBS), and then inoculated with 450 µl of the diluted virus solution. The plates were incubated for one hour at 37℃ with 5% CO_2, or at 35℃ for 2009 pandemic viruses. An overlay solution was prepared consisting of 1 part serum-free DMEM Complete medium, 1 part 2.4% Avicel® microcrystalline cellulose (IMCD), 0.14% w/v BSA Fraction V and 1 µg/ml TPCK-treated trypsin. 2ml overlay was applied to each well, and the plates were incubated for 48 hours for A/WSN/33 or 72 - 96 hours for A/Cal/04/09. After incubation, the overlay was aspirated and the cells were fixed with 5% neutral buffered formalin for 20 minutes at room temperature. The formalin was removed, the plates were washed with PBS, and the cell monolayers were stained with 0.1% Toluidine blue or 1% crystal violet (Sigma Aldrich) in 20% ethanol, for 20 minutes to 1 hour. The plates were washed in water and allowed to dry.

Since the weak cytopathic effect of A/Eng/195 *in vitro* results in poor plaque formation, this strain was titrated using a modified immunostaining protocol. The MDCK-SIAT1 monolayers were inoculated with 1ml diluted virus, and after incubating for 1 hour at 35℃ the inoculum was removed before overlay application. The overlay was modified to contain 1 part 2.4% Avicel to 2 parts serum-free DMEM Complete. The plates were incubated for three days, and fixed for 20 minutes with 2 ml 10% neutral buffered formalin per well before removing the overlay. The plates were washed twice with PBS, and the cells were permeabilised in 0.2% Triton™ X-100 (Bio-Rad) in PBS for 10 minutes at room temperature. After washing twice with PBS, the cells were incubated with monoclonal mouse anti-influenza nucleoprotein antibody (IgG2a clone AA5H; Bio-Rad) at 3.3 µg/ml in PBS with 2% w/v BSA, 0.5 ml per well, for one hour at room temperature on a plate rocker. The plates were washed twice in PBS, and a horseradish peroxidase-conjugated secondary antibody (horse anti-mouse IgG, Cell Signalling Technologies) was added at 0.5 ml per well of a 1:1000 dilution in PBS with 2% w/v BSA. The plates were incubated for one hour at room tem-

perature and then washed three times with PBS, and 0.5 ml TrueBlue™ TMB substrate (KPL Inc.) was added to each well. The plates were incubated at room temperature on a rocker until sufficient colour had developed (10-20 minutes), and deonised water was added to stop the reaction. The plates were washed in water and allowed to dry.

To determine viral titre, positively stained plaques for A/Eng/195, or negatively stained plaques for other strains, were counted. The well with the lowest dilution factor (i.e. highest plaque count) for which plaques were countable and distinct, without coalescence (usually in the range 10 - 200 plaques), was used for calculations. The titre in plaque-forming units (pfu) per millilitre was calculated using the formula:

$$Titre = \frac{plaque\,count \times dilution\,factor}{inoculum\,volume}$$

2.4 *In vivo* challenges with influenza A virus

Viral challenges with H1N1 strains of influenza A virus were performed in female mice aged between seven and 12 weeks of age (with a maximum age range of two weeks for an individual experiment). The calculated dose of virus was diluted in 40 µl serum-free DMEM. Mice were briefly anaesthetised with isoflurane and the dose was administered intranasally, dividing the total volume between the two nares. Mice were weighed and monitored for clinical signs before challenge and daily thereafter, using a semi-quantitative scoring system in which scores of 0 - 3 were assigned for posture, piloerection, mobility and respiratory effort (maximum 12 total). Maximum weight loss of 25% body weight or clinical score of 7/12 were used as the human end-points for all studies. For selected experiments, weights after day 4 post-challenge were obtained by an observer blinded to the genotypes. Mice were culled by exposure to rising CO_2. Dose optimisation experiments were performed in small groups of wild type mice (*n* = 3-4) prior to the genotype comparison experiments, to titrate the challenge to an appropriate level of severity.

2.5 Tissue harvest and processing

2.5.1 Bronchoalveolar lavage

After euthanasia, bronchoalveolar lavage fluid (BALF) was collected by intubating the trachea with polyethylene tubing via a tracheal incision, followed by instillation and collection of three 800 µl aliquots of phosphate buffered saline (PBS) with 3 mM ethylene diamine tetra-acetic acid (EDTA). If less than 1.5 ml of the total infusate was retrieved, the fluid was included in analyses for relative cell proportions, but excluded from analysis of absolute cell counts or protein concentrations. BALF was kept on ice for up to three hours, and then centrifuged at 450*g* for 5 minutes to pellet the cells. Supernatants were stored at -80℃ for later analysis. For flow cytometry of BALF cells, red cell lysis was performed by resuspending the cell pellets in RBC Lysis Buffer (Biolegend) and incubating for five minutes at room temperature, after which ice cold PBS was added to stop the reaction. The cells were washed once in PBS, and then resuspended in 250 µl PBS with 2% FBS (FACS Buffer) for staining, or in Roswell Park Memoral Institute 1640 (RPMI) Complete Medium (RPMI with 10% FBS, 2mM L-alanyl-L-glutamine dipeptide, 100 U/ml penicillin and 100µg/ml streptomycin) for culture. Cells were counted on a haemocytometer, identifying viable cells by exclusion of 0.2% Trypan Blue.

2.5.2 Lung tissue

Lung tissue for flow cytometry was stored in ice cold PBS for up to three hours, and then dissociated by cutting into small sections and incubating in 5 mM Collagenase IV (Roche) at 37℃ for one hour, with agitation. After digestion, the tissue was passed through a 70 µm cell strainer to yield a single cell suspension, and after washing once in FACS Buffer, red cell lysis was performed as described above. The lung homogenates were then resuspended in either FACS Buffer for staining, or in RPMI Complete medium for culture.

Lung samples for RNA analysis were placed in at least 10 volumes of RNA*Later* solution (Ambion) and kept at room temperature for 24 hours, followed by -80℃ for longer term storage.

Lung samples for analysis of virus or protein were frozen on dry ice and stored at -80℃. Lung tissues were homogenised for these analyses by mechanical disruption in a TissueLyser II homogeniser (Qiagen Inc.) in 1 ml serum-free DMEM, for 6 minutes at 28 cycles/second with a 5 mm stainless steel bead. Homogenates were centrifuged at 1000*g* for 5 minutes to remove insoluble material prior to analysis.

Where lungs were required for histology, any lobes needed for other analyses were first removed after ligation of the associated bronchus, and the remaining lung lobes were inflated with 10% neutral buffered formalin. They were then removed and stored in 10% neutral buffered formalin for 6-12 hours, then washed once in PBS and transferred to 70% ethanol solution for storage.

2.5.3 Peripheral blood

Blood was collected by cardiac puncture immediately after euthanasia, using a 23g needle and syringe primed with EDTA solution to give a final concentration of approximately 5 mM. Samples intended for complement analysis were supplemented with protease inhibitor Futhan-175 (BD Bioscience) at 50 µg/ml to prevent *ex vivo* complement activation. Blood was kept on ice for up to three hours. For complete blood counts, blood samples were run on an Advia 2120 automated haematology analyser (Siemens) by an independent diagnostic laboratory. For other analyses, samples were centrifuged at 800*g* for 5 minutes to pellet the cells. Plasma was removed and stored at -80℃. Red cell lysis was performed as described above, with 1 ml RBC Lysis Solution per sample, and cell pellets were washed in PBS and resuspended in PBS with 2% FBS.

2.5.4 Splenocytes

Spleens were removed and stored in ice cold PBS for up to three hours. Spleens were dissociated by crushing between two frosted glass slides, and passed through a 40 µm cell strainer to create a single cell suspension. Red cell lysis was performed as above with 3 ml RBC Lysis Buffer per sample. After washing once in PBS and once in RPMI Complete medium, viable cells were counted in a haemocytometer after staining with Trypan Blue, and then resuspended

to the required concentration in either PBS with 2% FBS for staining, or RPMI Complete medium for culture.

2.5.5 Bone marrow

To harvest bone marrow, femurs and tibiae were removed, stripped of muscle and connective tissue, and immersed first in 70% ethanol to dehydrate residual tissue and reduce bacterial contamination, and then in RPMI Complete medium. The ends of the bones were removed, and the marrow cavities flushed with RPMI Complete medium. The collected bone marrow suspension was centrifuged at 450*g* for 5 minutes, and then red cell lysis was performed as described above. The cells were washed once in PBS and once in RPMI Complete medium and were passed through a 70 µm cell strainer. Viable cells were counted on a haemocytometer.

For cryopreservation, bone marrow was centrifuged again at 450*g* for 5 minutes and resuspended in 90% FBS with 10% dimethyl sulphoxide (DMSO; Sigma-Aldrich). Aliquots of 0.5 to 1 ml, in cryopreservation tubes, were transferred to a Mr. Frosty™ Freezing Container containing isopropanol, to allow constant rate cooling at approximately -1℃/minute in a -80℃ freezer, and were then transferred to a -155℃ freezer the following day for long term storage. For resuscitation of cryopreserved bone marrow, samples were thawed rapidly in a 37℃ water bath, and 10 ml pre-warmed RPMI complete medium was added drop-wise to dilute out the cryopreservation medium. The cells were then centrifuged at 450*g* for 5 minutes and resuspended in RPMI Complete medium for onward use.

2.6 Culture and stimulation of bone marrow-derived macrophages

To differentiate bone marrow cells into macrophages, for each biological replicate, approximately 1 - 2 x 10^7 cells were added to a 10 cm diameter polystyrene bacteriological plate in 10 ml RPMI Complete medium, with 10^4 U/ml recombinant human colony stimulating factor 1 (rhCSF1; a kind gift from Chiron, USA),

and were incubated at 37℃ with 5% CO_2. After three days, 10 ml fresh culture medium (RPMI complete) with rhCSF1 was added. After culturing for seven days, the medium was removed, and the plates were washed gently with PBS to remove non-adherent cells. Adherent macrophages were lifted by incubating in 5 mM EDTA in PBS for 5 to 10 minutes, and were washed and resuspended in RPMI Complete medium. Viable macrophages were counted in a haemocytometer, determining viability by Trypan blue exclusion, and were replated in 24-well plates at 2 x 10^5 cells per well.

After incubating overnight to allow re-adherence, macrophages were stimulated either with lipopolysaccharide (LPS) or IAV. Lipopolysaccharide (*E. coli* O55:B5; Sigma Aldrich) was added to a final concentration of 100 ng/ml, in RPMI Complete medium with rhCSF1. For viral inoculation, IAV strain A/Cal/04/09 or A/WSN/33 was diluted in serum-free RPMI, and 10^6 plaque forming units (pfu) was added per well, to give a multiplicity of infection (MOI) of 5. After incubating for 1 hour, the viral inoculum was removed and the plates were washed once in PBS. The medium was replaced with either RPMI complete medium with rhCSF1, or for viral quantification serum-free RPMI with 0.14% w/v BSA, 2 mM L-glutamine, 100 U/ml penicillin, 100 µg/ml streptomycin, 1 µg/ml TPCK-treated trypsin and rhCSF1. Unstimulated cells, cultured in RPMI Complete medium with rhCSF1, were used as negative controls.

At the specified time points, supernatants were removed, centrifuged at 1000*g* for 5 minutes to remove cellular debris, and stored at -80℃ for later analysis. For RNA analysis, cells were washed in PBS and lysed on the plate in buffer RLT (Qiagen). Lysates were stored at -80℃ for later RNA extraction, as described below. For immunoflourescence, cells were washed in PBS and fixed in 10% neutral buffered formalin for 20 minutes at room temperature. The cells were washed three times in PBS with 2% FBS and stored at 4℃. Staining for immunofluorescence is described below.

2.7 Flow cytometry

Antibodies and viability markers used for flow cytometry, and their working concentrations, are given in Tables 2.1 and 2.2. Fluorochromes and viability dyes

were varied between experiments as required for spectral compatibility with other antibodies in the staining panels.

For staining, cells were suspended in PBS with 2% FBS (FACS buffer) and plated onto 96-well round-bottomed plates. To block non-specific binding, cells were incubated with 5 µg/ml purified anti-mouse CD16/32 antibody (Biolegend) in 50 µl FACS buffer for 1 hour at 4℃. After washing, the cells were then stained with the required panels of fluorochrome-conjugated antibodies for surface antigens, diluted to their working concentrations in 50 µl FACS buffer per well, for 1 hour at 4℃ in the dark. To distinguish live from dead cells, the cells were subsequently washed in PBS and incubated with an appropriate fixable viability dye, diluted to its working concentration (1:1000 - 1:4000) in PBS, at 4℃ for 15-30 minutes in the dark. The cells were washed twice in FACS buffer, and then fixed in 50 µl of a 1:1 mix of FACS buffer and 10% neutral buffered formalin at room temperature for 20 minutes. Finally, they were washed again in FACS buffer, and resuspended in an appropriate volume of FACS buffer for analysis, keeping protected from light.

For spectral compensation, additional control samples were stained with single fluorochrome-conjugated antibodies, or left unstained. Where cells were available in limited numbers, or where some markers were expected to be found in low numbers of cells or at low expression levels, UltraComp eBeads™ compensation beads (Thermo Fisher Scientific) were used in addition to cells to aid compensation. Where appropriate, fluorescence-minus-one controls were used to aid gating, and key markers were compared to a matched isotype control antibody (see Table 2.3), used at the same working concentration, to help distinguish non-specific antibody binding from true antigen expression.

Stained samples were run on a BD LSRFortessa™ or BD LSRFortessa™ X-20 flow cytometer. Data analysis was performed using FlowJo v10 10.5.3 (FlowJo, LLC, Ashland OR). Forward scatter and side scatter characteristics were used to distinguish cells from debris and to exclude doublets, and live cells were identified by viability dye exclusion. Cell surface markers used to identify mouse immune cell subsets are given in Table 2.4. Further details of gating strategies are provided in the relevant Results sections. Relative immune cell populations were calculated as the percentage of live leukocytes (defined as $CD45^+$ cells) or an alternative parent population where specified. Median fluorescence intensity

Target	Antibody	Fluorochrome	Working concentration	Manufacturer
CD45	Rat IgG2b,κ, clone 30-F11	FITC PE/Cy7	1 µg/ml 0.5 µg/ml	eBioscience, Thermo Fisher Scientific
CD11b	Rat IgG2bκ, clone M1/70	APC/Cy7	0.5 µg/ml	Biolegend
CD11c	Hamster IgG, clone N418	PE/Cy7	0.4 µg/ml	Biolegend
F4/80	Rat IgG2a,κ, clone BM8	AF700	2.5-5 µg/ml	Biolegend
Ly6G	Rat IgG2a,κ, clone 1A8	PE	0.67 µg/ml	Biolegend
SiglecF	Rat IgG2a,κ, clone E50-2440	BV421	1 µg/ml	BD Biosciences
CD4	Rat IgG2a,κ, clone RM4-5	PE AF700	1 µg/ml 1 µg/ml	BD Biosciences
CD8	Rat IgG2a,κ, clone 53-6.7	FITC APC/Cy7	5 µg/ml 0.5 µg/ml	BD Biosciences Biolegend
CD19	Rat IgG2a,κ, clone 6D5	APC/Cy7	1 µg/ml	Biolegend
CD3	Recombinant human IgG1, clone REA641	PerCP/Vio700	0.6 µg/ml	Miltenyi Biotec
NKp46	Rat IgG2a,κ, clone 29A-1.4	BD Horizon V450	1-2 µg/ml	BD Biosciences
CD25	Recombinant human IgG1, clone REA568	APC	1.2 µg/ml	Miltenyi Biotec
CCR6	Recombinant human IgG1, clone REA277	PE	1.2 µg/ml	Milteny Biotec
CD97v2	Recombinant human IgG1, clone REA678	APC	1.2 µg/ml	Miltenyi Biotec
CD55	Recombinant human IgG1, clone REA300	APC	1.2 µg/ml	Miltenyi Biotec

Table 2.1 – Surface marker antibodies used in flow cytometry for murine immune cells. PE: phycoerythrin; APC: allophycocyanin; FITC: fluorescein isothiocyanate

Viability dye	Band pass filter	Working dilution	Manufacturer
Zombie Aqua	V525/50	1:4000	Biolegend
Zombie Violet	V450/50	1:1000 - 1:2000	Biolegend
Zombie Yellow	V586/15	1:1000	Biolegend
eFluor™780 Fixable Viability Dye	R780/60	1:4000	eBioscience, Thermo Fisher Scientific

Table 2.2 – Viability dyes used in flow cyometry

Clone	Fluorochrome	Manufacturer
Recombinant human IgG1, clone REA293	APC PerCP-Vio700 PE	Miltenyi Biotec
Rat IgG1,κ, clone RTK2071	PE BV421	Biolegend
Rat IgG2a,κ, clone RTK2758	APC/Cy7 AF700	Biolegend
Rat IgG2b,κ, clone ES26-5E12.4	Vio667	Miltenyi Biotec
Rat IgG2b,κ, clone RTK4530	PE/Cy7	Biolegend
Mouse IgG1κ, clone IS5-21F5	APC	Miltenyi Biotec

Table 2.3 – Isotype control antibodies used in flow cytometry. Isotype control antibodies were used at the same working concentration as the specific antibodies to which they were compared.

was used as a summary statistic for fluorescence intensity comparisons between populations. Absolute cell counts were calculated by multiplying the total absolute viable cell count (from haemocytometer counts for BALF, or total white blood cell counts from an automated haematology analyser for blood) by the cell population of interest expressed as proportion of live single cells (derived from flow cytometry).

Cell type		**Markers**
Non-immune cells		$CD45^-$
Neutrophils		$CD45^+$, $CD11b^+$, $Ly6G^+$
Eosinophils		$CD45^+$, $CD11b^+$, $Ly6G^-$, $CD11c^-$, $F4/80^+$, $SiglecF^+$
Alveolar macrophages		$CD45^+$, $F4/80^+$, $CD11c^+$, $SiglecF^+$
Other macrophages and dendritic cells (DCs)	Monocytes and $CD11c^-$ macrophages	$CD45^+$, $CD11b^+$, $F4/80^+$, $CD11c^-$, $SiglecF^-$
	$CD11c^+$ $F4/80^+$ macrophages and DCs	$CD45^+$, $CD11b^+$, $F4/80^+$, $CD11c^+$, $SiglecF^-$
	$CD11c^+$ $F4/80^-$ cells	$CD45^+$, $CD11b^{\pm}$, $F4/80^-$, $CD11c^{hi}$, $SiglecF^-$
B lymphocytes		FSC/SSC^{low}, $CD45^+$, $CD4^-$, $CD8^-$, $CD19^+$
CD4 T lymphocytes		FSC/SSC^{low}, $CD45^+$ and/or $CD3^+$, $CD4^+$, $CD8^-$, $CD19^-$
CD8 T-lymphocytes		FSC/SSC^{low}, $CD45^+$ and/or $CD3^+$, $CD4^-$, $CD8^+$, $CD19^-$
Natural killer cells		FSC/SSC^{low}, $CD45^+$, $CD4^-$, $CD8^-$, $CD19^-$, $NKp46^+$

Table 2.4 – Mouse immune cell population definitions according to cell surface markers. As many macrophages and other activated immune cells can express CD11c in the lung, dendritic cells cannot be distinguished definitively from other immune cells on the basis of these markers.[19]

2.8 Intracellular cytokine staining

For intracellular cytokine staining, BALF cells or splenocytes isolated as described above were cultured in RPMI Complete medium with 81 nM phorbol 12-myristate 13-acetate (PMA) and 1.3 µM ionomycin (Cell Activation Cocktail, Biolegend), together with 1 µg/ml of the Golgi transport inhibitor brefeldin A (GolgiPlug™, BD Bioscience), for 4 hours at 37℃ and 5% CO_2. Initial surface marker and viability staining were performed as in section 2.7 above. After the viability stain, the cells were washed in FACS buffer, and fixed and permeabilised with 100 µl Cytofix/Cytoperm™ Fixation and Permeabilization Buffer (proprietary formulation containing saponin and 4.2% formaldehyde; BD Biosciences) per well, for 20 minutes at 4℃ in the dark. The cells were washed twice in a saponin-containing wash buffer (BD Perm/Wash™ Buffer, BD Biosciences) to maintain permeability, and stained with anti-cytokine antibodies or matched isotype controls (Tables 2.5 and 2.3) diluted to their working concentrations in Perm/Wash Buffer, in a total volume of 50 µl per well, for 1 hour at 4℃ in the dark. The cells were washed again in Perm/Wash Buffer, and finally resuspended in FACS buffer for analysis as described above.

Target	Antibody	Fluorochrome	Working concentration	Manufacturer
IFNγ	Rat IgG1,κ, clone XMG1.2	PE	2 µg/ml	Biolegend
IL-10	Rat IgG2b,κ, clone JES5-16E3	Vio667 PE/Cy7	3 µg/ml 1.2 µg/ml	Miltenyi Biotec Biolegend
IL-4	Rat IgG1,κ, clone 11B11	BV421	2 µg/ml	Biolegend

Table 2.5 – Antibodies used for intracellular cytokine staining

2.9 Histopathology

Formalin-fixed tissues were processed to paraffin wax, sectioned and stained with haematoxylin and eosin, by a commercial laboratory. Sections were examined and scored while blinded to genotype. A previously described semi-

quantitative scoring system was used[155], based on severity and extent of interstitial leukocyte infiltrate (0-5), perivascular lymphoid accumulation (0-4), haemorrhage (0-3) and fibrosis (0-1), to give a maximum total score of 13.

2.10 Immunofluorescence for viral nucleoprotein

Immunofluorescence was used to detect viral nucleoprotein (NP) in infected BMDMs (see section 2.6) cultured on glass coverslips. After removing the culture medium and washing in PBS, cells were fixed for 20 minutes in 10% neutral buffered formalin at room temperature. Cells were then washed in PBS with 1% FBS, and permeabilised with 0.2% Triton TX-100 in PBS for 5 minutes. The cells were washed again in PBS with 1% FBS, and incubated with mouse anti-influenza nucleoprotein antibody (IgG2a clone AA5H; Bio-Rad) at 3.3 µg/ml in PBS with 1% FBS at room temperature for 1 hour. After further washes, the cells were incubated with the secondary antibody, Alexa Fluor® 488-conjugated goat anti-mouse IgG (Invitrogen), at 2 µg/ml, and with Texas Red-conjugated phalloidin (Thermo Fisher Scientific) at 2.2 µM to stain cellular actin, for 1 hour at room temperature in the dark. To stain the nuclei, the cells were incubated with Hoechst 33342 (Thermo Fisher Scientific) at 20 µM in PBS for 10 minutes at room temperature, after washing with PBS. Cells were mounted onto glass slides with ProLong™ Gold antifade mounting medium (Thermo Fisher Scientific).

Slides were imaged on a Leica DMLB fluorescent microscope. To quantify NP-positive cells, multicolour image stacks of three views per sample were collected at 100X magnification, using standardised exposure settings. Views were selected on the basis of appropriate cell density and absence of major artifacts, with reference only to the nuclear stain channel, to avoid selection bias. Images were analysed in ImageJ version 1.52n (National Institutes of Health). Cells with nuclear staining for IAV NP were counted manually for each image, while blinded to genotype. Nuclei in the same image were counted using an automated macro, to allow calculation of the percentage of positive cells.

2.11 Total protein quantification

Total protein in BALF was quantified using a bicinchoninic acid (BCA) colourimetric assay (Pierce™ BCA Protein Assay, Thermo Fisher Scientific). 25 µl of test fluid or BSA standard (in the range 25 - 2000 µg/ml, diluted in PBS) was added to each well of a flat-bottomed 96-well microplate. Assays were performed in duplicate. 200 µl BCA Working Reagent (prepared according to the manufacturer's instructions) was added to each well and the plates were mixed thoroughly on a plate shaker for 30 seconds. The plates were covered and incubated at room temperature for 90 minutes. Absorbance at 562 nm was read on a SpectraMax® microplate reader (Molecular Devices Inc.). To calculate concentration, a standard curve was constructed by four-parameter curve fit of optical densities for the BSA standards, and concentration of unknown samples derived by interpolation from this curve, using package *SciPy* version 1.3.2 in custom Python scripts. The mean of the technical replicates for each sample was used for analysis.

2.12 Protein quantification by enzyme-linked immunosorbent assay

Concentrations of activated complement component C5a, and the cytokines interferon alpha (IFNα), interferon-gamma (IFNγ) and tumor necrosis factor alpha (TNFα) were measured by enzyme-linked immunosorbent assay (ELISA). Antibodies used and their working concentrations are given in Table 2.6.

Capture antibodies were diluted to their working concentrations in PBS for IFNγ and TNFα, or in bicarbonate buffer (pH = 9.5; ELISA Coating Buffer, Biolegend) for other assays. High-binding 96-well microplates (Nunc Maxisorp™, Thermo Fisher Scientific) were coated by adding 50 µl of the diluted capture antibody to each well and incubating overnight at 4℃. The plates were washed three times in ELISA Wash Buffer, consisting of PBS with 0.05% Tween-20 (Sigma-Aldrich). To block non-specific binding, the plates were incubated with 300 µl per well 1% w/v BSA in PBS for 2 hours at room temperature, and were washed again as above. Standards were reconstituted according to the manufacturers'

instructions, and diluted to the concentration range shown in Table 2.6 by serial dilutions in the assay diluent, which varied according to the manufacturer's recommendations for each assay: the diluents used were an FBS-containing proprietary buffer for C5a (ELISA Assay Diluent, BD Biosciences), 1% w/v BSA with 0.05% Tween-20 in PBS for IFNα, 1% w/v BSA in PBS for TNFα and 0.1% w/v BSA with 0.05% Tween-20 in Tris-buffered saline (20mM Trizma base and 150mM NaCl, pH 7.3) for IFNγ. Test samples were also diluted in the same diluent if necessary to ensure that concentrations remained within the working range of the assay. 50 µl of each sample or standard was added to the plate. Two technical replicates were performed for each assay, and the mean value used for analysis.

Target	**Capture Antibody**	**Biotinylated Detection Antibody**	**Standard range**	**Product / Manufacturer**
C5a	Rat anti-mouse C5a, Clone I52-1486, Working conc: 2 µg/ml	Rat anti-mouse C5a, Clone I52-2778, Workinc conc: 2 µg/ml	39 pg/ml - 10 ng/ml	BD Biosciences
IFNα	Anti-mouse IFNα-2, Working conc: 2 µg/ml	Anti-mouse IFNα-2, Working conc: 0.5 µg/ml	1.95 - 125 pg/ml	Mouse IFNα-2 Matched Antibody Pair Kit, Abcam
IFNγ	Rat anti-mouse IFNγ, Working conc: 4 µg/ml	Goat anti-mouse IFNγ, Working conc: 0.4 µg/ml	15.6 pg/ml - 2 ng/ml	DuoSet®, R&D Systems
TNFα	Goat anti-mouse TNFα, Working conc: 0.8 µg/ml	Goat anti-mouse TNFα, Working conc: 50 ng/ml	15.6 pg/ml - 2 ng/ml	DuoSet®, R&D Systems

Table 2.6 – ELISA reagents.

The plates were covered and incubated for two hours at room temperature, on a plate shaker. The plates were washed three times in ELISA Wash Buffer, and 50 µl biotinylated detection antibody (diluted to its working concentration in the specific diluent for each assay) was added to each well. The plates were covered and incubated for a further one hour (IFNα) or two hours (other assays) at room temperature, on a plate shaker. The plates were washed three times in ELISA Wash Buffer, and 50 µl of streptavidin-conjugated horse radish

peroxidase (Streptavidin-HRP R&D Systems), diluted 1:40 in the specific diluent for each assay, was added to each well. The plates were incubated on a plate shaker for 20 minutes at room temperature, and then washed four times in ELISA Wash Buffer. 100 µl substrate solution (tetramethylbenzidine with hydrogen peroxide, R&D Systems Color Reagents A and B) was added to each well, and the plates were incubated at room temperature in the dark, on a plate shaker, until sufficient colour development had occurred (10 to 20 minutes). 50 µl of an acidic solution Solution (1M sulphuric acid or 0.5M orthophosphoric acid) was added to stop the reaction. Optical density was read at 450 nm on a SpectraMax™ microplate reader, subtracting the optical density at 570 nm to correct for optical variation in the plate. Concentrations were calculated from the optical density readings by interpolation from a standard curve of readings for the standards, as for the BCA assay (see section 2.11).

2.13 RNA extraction

Ribose nucleic acid (RNA) for viral quantification was extracted from cell-free biological fluids using a Viral RNA Extraction Kit (Qiagen), according to the manufacturer's instructions. 140 µl of each sample was added to 560 µl lysis buffer AVL combined with 5.6 µg carrier RNA diluted in buffer AVE, and mixed briefly by vortexing. The samples were incubated for 10 minutes at room temperature to allow complete lysis, after which 560 µl ethanol was added to precipitate the RNA, vortexing again to mix. The solution was applied to a QIAamp Mini spin column, and centrifuged at 6000*g* for 1 minute, to allow the RNA to bind to the column membrane. The flow-through was discarded, and the column membrane was washed with 500 µl Buffer AW1, centrifuging at 6000*g* for 1 minute, and again with 500 µl Buffer AW2, centrifuging at 20,000*g* for 3 minutes, discarding the flow-through at each step. The column was centrifuged again at 20,000*g* for 1 minute in a clean collection tube to eliminate residual wash buffer. The spin column was placed in a clean nuclease-free microcentrifuge tube, 60 µl elution buffer AVE was added to the column membrane, and after incubating for 1 minute at room temperature, the column was centrifuged at 6000*g* for 1 minute to elute the RNA.

RNeasy Mini Kits (Qiagen) were used for extraction where RNA was also re-

quired for quantification of host gene expression. Cultured cells were lysed on the plate in 350 µl buffer RLT, and homogenised by vortexing for 1 minute. For lung tissue, up to 25 mg RNA*later*-stabilised tissue was added to 600 µl RLT with 40 mM dithithreitol and homogenised in a TissueLyser II homogeniser (Qiagen) for 4 minutes at 28 cycles per second with a 5 mm stainless steel bead. The sample was centrifuged at 1000*g* for 5 minutes to remove insoluble material before proceeding with the extraction. For both tissue types, one volume of 70% ethanol was added to the lysate and the solution was mixed well by pipetting. The mixture was added to an RNeasy Spin Column and centrifuged for 1 minute at 10,000*g* to allow the RNA to bind to the column membrane. The flow-through was discarded, and the membrane washed with 350 µl wash buffer RW1, centrifuging at 10,000*g* and discarding the flow-through. On-column DNase digestion was performed to remove residual genomic DNA: 27 Kunitz units DNase I (RNase-free DNase Kit, Qiagen), diluted to 80µl in buffer RDD, was added to the column membrane and incubated for 15 minutes at room temperature. The column membrane was then washed successively with 350 µl buffer RW1 and twice with 500 µl buffer RPE, centrifuging for 1 minute at 10,000*g* and discarding the flow-through at each step. The column was then centrifuged for a further 1 minute at 20,000*g* in a clean collection tube to remove residual wash buffer. Finally, 30 to 50 µl RNase-free water was then added to the column membrane and after incubating for 1 minute at room temperature the column was centrifuged at 10,000*g* for 1 minute to elute the RNA. The eluate was passed through the column a second time to maximise yield.

For both methods, all centrifuge steps were performed at room temperature. After extraction, RNA was kept on ice for immediate use or stored at -80 ℃. RNA was quantified by spectrophotometry (Nanodrop; Thermo Fisher Scientific). Absorbance at 260 nm was used to determine concentration, while the 260/280 nm ratio was used as measures of purity (with $\geq$ 1.8 considered adequate).

2.14 Quantitative RT-PCR for host gene expression

Primers for quantitative reverse transcription polymerase chain reaction (qRT-PCR) were either obtained from published literature or were designed using NCBI Primer BLAST and Primer3.[156] To avoid amplification of genomic DNA, primer pairs were designed either with at least one primer spanning an exon-exon junction, or with the pair spanning a long intron. Target primer pair parameters included a product size of 70-200 base pairs, melting temperature of 57-63℃, absence of off-target matches with fewer than 4 nucleotide mismatches, and minimal 3' self-complementarity. Primer pairs were considered acceptable for use if they produced a single pure product on melt curve analysis, and showed minimal amplification in negative controls (lacking template or without reverse transcription). Primers are shown in Table 2.7.

Target	**Primers (5'→ 3')**
Murine *B2m*	Fwd: CTGCTACGTAACACAGTTCCACC Rev: CATGATGCTTGATCACATGTCTCG
Murine *Hprt1*	Fwd: GAGGAGTCCTGTTGATGTTGCCAG Rev: GGCTGGCCTATAGGCTCATAGTGC
Murine *Gapdh*	Fwd: GTTGTCTCCTGCGACTTCA Rev: GGTGGTCCAGGGTTTCTTA
Murine *Tnfa*	Fwd: CATCTTCTCAAAATTCGAGTGACAA Rev: TGGGAGTAGACAAGGTACAACCC
Murine *Il6*	Fwd: GAGGATACCACTCCCAACAGACC Rev: AAGTGCATCATCGTTGTTCATACA
Murine *Cxcl10*	Fwd: AAAAAGGTCTAAAAGGGCTC Rev: AATTAGGACTAGCCATCCAC
Murine *Il12b* (IL12-p40)	Fwd: GGAAGCACGGCAGCAGAATA Rev: AACTTGAGGGAGAAGTAGGAATGG
Influenza H1N1 Segment 2 (PB1), degenerate primers	Fwd: GGAACAGGATACACCATGGA Rev: AGTGGYCCATCAATCGGGTT
Influenza H1N1 Segment 7 (M)	Fwd: ACCGAGGTCGAAACGTACGT Rev: CCAGTCTCTGCGCGATCTC

Table 2.7 – Primers used in qRT-PCR.

First strand cDNA synthesis was performed using SuperScript™ III reverse transcriptase (Thermo Fisher Scientific) with an oligo(dT)15 primer (Promega Cor-

poration). First, 50 to 200 ng total RNA from each sample was incubated with 1 µl (500 ng) primer, diluted to a total volume of 12.5 µl with nuclease-free water, at 70℃ for 10 minutes, and then for 2 minutes on ice. 12.5 µl reverse transcription master mix was then added to each reaction, consisting of 5 µl 5X First Strand Buffer, 2.5 µl 10 mM deoxynucleotide triphosphate mix (dNTPs), 2.5 µl 0.1 M dithiothreitol, 1 µl Superscript III reverse transcriptase at 200 units/µl, and 1.5 µl nuclease-free water. The reactions were incubated at 50℃ for 50 minutes, and then 70℃ for 10 minutes to inactivate the enzyme. Negative control reactions were performed omitting either the template or the enzyme.

The cDNA generated in the above reverse transcription reaction was diluted tenfold for use in quantitative PCR reactions. 20 µl reactions were set up with 10 µl Applied Biosystems™ Fast SYBR Green MasterMix (Thermo Fisher Scientific), 0.4 µl forward and reverse primers at 10 µM, 4.2 µl nuclease-free water and 5 µl diluted template. Reactions were performed in duplicate or triplicate. Reactions were run in an Applied Biosystems™ 7500 Fast Real-Time PCR Machine, with an initial denaturation step of 20 seconds at 95℃, followed by 40 cycles of 3 seconds' denaturation at 95℃ and 20 seconds at 60℃ for annealing, extension and data collection. A melt curve was performed after the end of the reaction, with an incremental temperature increase from 60℃ to 95℃.

Appropriate placement of the amplification baseline and threshold was verified and adjusted manually if necessary. The melt curve was inspected for evidence of multiple amplification products - a single lower peak consistent with primer dimer formation was considered acceptable if present in negative controls only, but assays were rejected if they had evidence of multiple products in experimental samples. For analysis, the mean cycle threshold (CT) value across all technical replicates was taken for each sample, and normalised to the mean of two endogenous controls, selected on the basis of published data on endogenous control stability in murine macrophages[157]: *B2m* (beta-2-microglobulin) and *Hprt1* were used for LPS-stimulated macrophages, while *B2m* and *Gapdh* were used for IAV-treated macrophages, as preliminary results showed downregulation of *Hprt1* after IAV infection. The resulting normalised values (ΔCT values) were used for statistical analysis, and were normalised to the mean of wild type values at a specified reference time point for presentation as relative expression ($2^{-\Delta\Delta CT}$) values.

2.15 Viral quantification by qRT-PCR

As A/Eng/195 has limited cytopathic effect and thus poor plaque formation *in vitro*, a one-step qRT-PCR assay was used to facilitate measurement of this strain, using primers targeting H1N1 Segments 2 (Polymerase segment B1) and 7 (M gene). While qRT-PCR cannot distinguish between infectious virus and other sources of viral RNA such as defective interfering particles, it can be useful as a preliminary measure for comparisons between host groups where the input virus is the same in both groups, and thus viral sequence is unlikely to influence the relationship between viral titre and RNA concentration.

Primers used are given in Table 2.7. Assays were performed with the GoTaq® 1-Step RT-qPCR System (Promega Corporation). 10 µl reactions were set up, in duplicate or triplicate, consisting of 5 µl 2X GoTaq qPCR MasterMix, 1µl of each primer at 1 µM, 0.8 µl nuclease-free water, 0.2 µl GoScript™ reverse transcriptase mix for 1-Step RT-qPCR, and 2 µl template. Negative control reactions were set up omitting the template or reverse transcription mix. The reactions were run on an Applied Biosystems™ 7500 Fast Real-Time PCR Machine. First, reverse transcription was performed at 37℃ for 15 minutes, followed by reverse transcriptase inactivation and hot-start polymerase activation at 95℃ for 10 minutes. After this, 40 cycles were performed consisting of 10 seconds' denaturation at 95℃, 30 seconds' annealing and data collection at 60℃, and extension for 30 seconds at 72℃. Finally, a melt curve was over an incremental temperature increase from 60℃ to 95℃.

2.15.1 Validation of the qRT-PCR assay for A/Eng/195

For initial assay validation, primer performance was assessed across serial dilutions of pDUAL plasmids encoding the relevant A/Eng/195 segments, in the range 20 - 2 x 10^9 copies per reaction, and melt curve analysis was used to characterise the amplified product for templates including whole-virus RNA extracts and cell-free BALF from infected animals. Melt curve analysis (Figure 2.1A) showed a single clear peak, suggesting amplification of a single pure product. Standard curves with plasmid-encoded sequence (Figure 2.1B) indicated that cDNA could be reliably detected down to the lowest concentration

tested for both segments (20 copies per reaction), with minimal amplification in no-template controls (CT values >37, at least 5 higher than for the lowest plasmid concentration).

To calculate primer efficiency, the best fit line of a plot of concentration versus CT (excluding extremes of the concentration range if deviating from the expected linear-log relationship) was determined using Python package *numpy*, to give an equation in the form:

$$CT = a \ln(concentration) + b$$

The efficiency was calculated using the following formula:

$$Efficiency(\%) = 100(e^{-1/a} - 1)$$

Conversely, concentration of an unknown sample could be calculated by interpolation using the formula:

$$Concentration = e^{(CT-b)/a}$$

Efficiency was greater than 90% for both segments, although the linear-log plot deviated slightly from the predicted linear slope towards the extremes of the concentration range (<200 copies or > 10^8 copies per reaction).

To assess the impact of RNA extraction and reverse transcription efficiency on assay performance, similar standard curves were performed using both extracted viral RNA from a single stock (nominal concentration range 0.08 - 8000 pfu equivalents per reaction) and viral stock serially diluted prior to RNA extraction (1.6 to 200 pfu equivalents per reaction). As expected, there was minimal amplification in controls without the reverse transcriptase. No drop in performance was apparent for serial dilutions of RNA (Figure 2.1C), indicating consistent reverse transcription efficiency, but there was a notable deviation from the expected linear-log relationship when virus was diluted prior to RNA extraction, with more marked loss of performance for the Segment 7 assay (Figure 2.1D). This suggests some variability in extraction efficiency across the concentration range. Overall, the above findings indicated that assay performance was acceptable for semi-quantitative comparisons of viral RNA concentrations be-

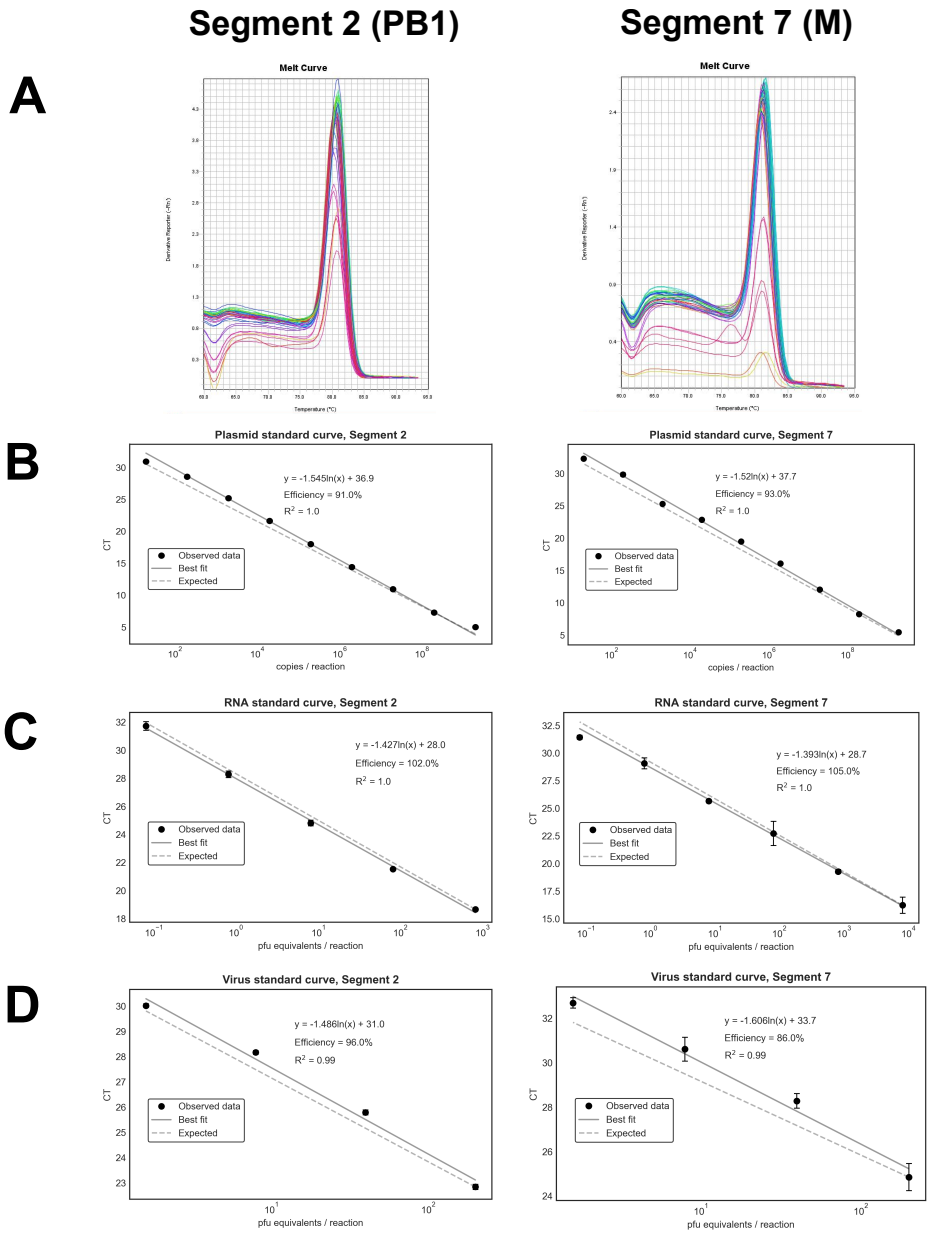

Figure 2.1 – Validation of a qRT-PCR assay for A/Eng/195. A: Melt curves from qRT-PCR performed in BALF show a clear single peak for both segment 2 and 7. B-D: Standard curves performed with serial dilutions of pDUAL plasmid encoding the viral segment (B), RNA extracted from A/Eng/195 stock (C), or whole A/Eng/195 virus diluted prior to extraction (D). Mean ± standard deviation is shown for technical replicates (n=3). The calculated best fit line within the linear range (solid line), and expected curve for an ideal reaction with 100% efficiency (dotted line), are displayed.

tween sample groups, but although performance could be improved with use of a spike-in extraction control, accuracy will be limited for determination of absolute concentrations relative to a standard curve.

2.15.2 Concentration calculation for A/Eng/195 qRT-PCR

For estimation of virus concentrations in cell-free biological fluids, samples were run together with a standard curve of serially diluted RNA from the input virus stock used in the experiment, extracted using the same method as for the samples; a constant volume was used for each extraction to allow comparisons, with the assumption of lack of bias in extraction efficiency between groups. Sample concentrations were calculated by interpolation from the standard curve using the formulae in section 2.15.1 above. The geometric mean was taken of the concentrations for Segment 2 and 7, and concentrations were corrected for dilution factors.

In lung tissue, relative quantification by the $2^{-\Delta\Delta CT}$ method was performed as for host genes. Sample CTs were normalised to the mean of three endogenous controls (*GAPDH*, *HPRT1* and *B2M*). Log-transformed concentrations or -ΔCT values were used for statistical analyses, and were normalised to the mean of wild type values for presentation as relative expression ($2^{-\Delta\Delta CT}$) values.

2.16 Statistical analysis

Statistical analyses were performed in R version 3.5.2 or in *SciPy* version 1.3.2 in Python 3.7.

For simple 2-group comparisons between genotypes, approximation to a normal distribution was verified by a Shapiro-Wilk test and by visual examination of histograms and normal probability (Q-Q) plots, and homogeneity of variances was assessed using Levene's test. Data were log-transformed if appropriate. If data approximated to a normal distribution, a two-tailed, indepent-sample Student's t-test (for equal variances) or Welch's t-test (for unequal variances) was used. Where there was insufficient evidence of normality, a Wilcoxon rank-sum test was used as a non-parametric alternative. Confidence intervals for

non-parametric data were estimated by the Bauer method, using the Hodges-Lehmann location estimator.

Where adjustment for additional factors such as sex, time point or study batch were required, general linear models were used, including these factors (and their interaction when relevant) as fixed effects. Type III Sums of Squares were used to determine significance of factor effects, and confidence intervals for main factors were calculated using the t distribution (with R function *confint*). Where indicated, *post hoc* pairwise comparisons, with confidence interval calculation, were performed using the Tukey method (R package *emmeans*). The assumption of normal distribution of residuals was assessed by visual examination of histograms and normal quantile plots of residuals and with a Shapiro-Wilk test, and homogeneity of variances was assessed by plotting residuals against fitted values and with a Levene's test where appropriate. If necessary, data were transformed either with a logarithmic transformation or an optimised Box-Cox power transformation (R package *MASS*) to fulfill the assumptions.

Time course data such as weight loss curves, where samples were matched between time points, were analysed by mixed models using R package *lme4*, using the restricted maximum likelihood (REML) method. Time, genotype and time:genotype interaction were assigned as fixed effects and subject as a random effect. Where multiple studies were pooled, the study and study:time interaction were included as additional fixed effects. Type III Wald F-tests were used to determine significance of parameter effects, and bootstrapping was used to compute confidence intervals (C.I.) for effect size estimates, using linear combinations of model coefficients (package *glht*) where appropriate. *Post hoc* contrasts were performed to determine the effect of genotype at each time point, using the multivariate *t* method to correct for multiple comparisons.

Except where otherwise specified, data have been displayed as box plots combined with individual data points, where boxes represent the interquartile range and whiskers indicate upper and lower quartiles ± 2 $\times$ interquartile range. The median is represented by a horizontal line and arithmetic or geometric mean (where appropriate) represented by a '+' symbol. Statistical significance has been defined as $p < 0.05$.

Statistical analysis of CRISPR screen data is described separately in section section 2.18.8.

2.17 Genomic data sources

Genomic data, including genome sequences, gene models, chromatin features, expression quantitative locus (eQTL) mapping, tissue expression profiles and genetic association data, were obtained from publicly-available data sources, as detailed in Table 2.8.

Table 2.8 – Genomic datasets used in this thesis

Name	Version / accession nos.	Data type	Accessed via	Refs.
Human genome	GRCh38	Genome assembly	Ensembl genome browser (http://www.ensembl.org), UCSC genome browser (https://genome-euro.ucsc.edu/), downloaded from Ensembl (http://ftp.ensembl.org/pub/release-103/fasta/homo_sapiens/dna/)	
Mouse genome	MGSCv37 (mm9)	Genome assembly	Ensembl genome browser, UCSC genome browser	
RefSeq and GENCODEv32		Gene models	Ensembl genome browser, UCSC genome browser, downloaded from GENCODE portal (https://www.gencodegenes.org/human/releases.html)	158
dbSNP	Build 150 (GRCh38)	Genetic variation	https://ftp.ncbi.nlm.nih.gov/snp/	159
1000 Genomes	Phase 3	Population genetics, linkage disequilibrium	Ensembl Genome Browser	160
Functional Annotation of the Human Genome (FANTOM5)		Transcription start sites, enhancers, tissue expression profiles	Fantom Zenbu browser and download, https://fantom.gsc.riken.jp/5/data	50
Gene-Tissue Expression Project	v8	eQTL mapping, splice variants	GTEx Portal https://www.gtexportal.org/	64
ENCODE Encyclopedia of DNA Elements	Portal accessed 20/01/2021	Chromatin features, candidate regulatory elements, ChIP-Seq data	ENCODE portal (https://www.encodeproject.org/); UCSC Genome Browser ; Factorbook (https://www.factorbook.org)	161–164
GeneHancer regulatory database		Enhancers integrated from various primary sources	UCSC Genome Browser	161,165
PhyloP scores		Conservation scores	UCSC Genome Browser	161,166

Continued on next page

Table 2.8 – *Continued from previous page*

Name	**Version / accession nos.**	**Data type**	**Accessed via**	**Refs.**
Factorbook		Transcription factor binding motifs and ChIP-Seq data	https://www.factorbook.org	[163]
SNP2TFBS		Transcription factor binding motifs, variant impact predictions	https://ccg.epfl.ch/snp2tfbs/snpviewer.php	[167]
Mouse GeneAtlas	MOE430, gcrma (probesets 1418394_a_at, mCD55_-1418762_at)	Tissue expression profiles (microarray)	BioGPS Portal (www.biogps.org)	[168]
Human primary cell atlas	Probesets 202910_s_at, 1555950_a_at	Tissue expression (microarray)	BioGPS Portal	[169]
Immunological Genome Project mouse immune system atlas	GEO Accession no. GSE109125	Tissue expression profiles (RNA-Seq)	NCBI Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/	
NHGRI-EBI GWAS Catalogue	Accessed 08/08/2020	Genetic association	https://www.ebi.ac.uk/gwas/	[170]
Gene ATLAS		Genetic associations in UK Biobank	http://geneatlas.roslin.ed.ac.uk	[171]
COVID-19hg Meta-analyses	Data freeze 18/01/2021	Genetic association	https://www.covid19hg.org	[150]
GenOMICC COVID-19 GWAS	Accessed 15/09/2020	Genetic association	Our group's own data	[9]

2.18 CRISPR screening for regulatory variants

2.18.1 Guide library design

A targeted single guide RNA (sgRNA) library for variants in a linkage disequilibrium (LD) block was designed using custom scripts in Python 3.7.3 (except where specified), adapted from design methods previously described for genome-wide CRISPR screens.[172] The full scripts used are available at: https://github.com/baillielab/locus_crispr.

First, all variants in linkage disequilibrium with the index variant (rs4813319) were identified, with a threshold of $r^2 > 0.2$ in data from a British population from England and Scotland from the 1000 Genomes Project.[160] Variant coordinates and allele data, and chromosome 20 variant flank sequences in FASTA format, were obtained from the dbSNP database, build 150, human genome assembly GRCh38 version.[159] Reference alleles, and flank sequences up to 35 base pairs on either side of each variant in the LD block, were extracted from these data files using script 'flanks_from_dbSNP.py'.

To enable flexible selection of protospacer-adjacent motif (PAM) sequences at later stages of the pipeline, initially all possible 20-mer guides within this region were screened for potential on-target cutting activity, without filtering for PAM sequence. On-target scores were calculated based on sequence features of 30-mer sequences containing the 20-mer sgRNA, PAM sequence and surrounding flank sequences, using the Doench 'Rule Set 2' machine learning-derived algorithm.[173] This was implemented in the script 'score_guides.py' using package *scikit-learn* version 0.16.1, in Python 2.7 to match the version used in creation of the algorithm, with model parameters set to omit PAM audit. Candidate guides were then filtered according to the desired set of allowable PAM sequences, and a FASTA file was created for off-target screening using script 'pandas_to_fasta.py', appending each allowable PAM to the 20mer sequence so that only off-target matches with a relevant PAM were included in scoring. For spCas9, only NGG PAMs were allowed, although off-target matches including NAG PAMs were also included. For xCas9, in the absence of an extensively validated PAM range, all PAMs with a 'PAM depletion score' greater than 0.8 in the PAM depletion assay in the initial describing paper[174] were included to give

a permissive PAM set: NGN, NNG, GAN and CAT.

To estimate off-target effects, candidate sgRNAs with appended PAM sequences were aligned to the human genome (assembly GRCh38) using Seqmap 1.0.8[175], allowing up to 3 mismatches but no indels, and returning all matches. This process was parallelised as necessary for computational efficiency. Potential matches were filtered to remove those with mismatches in the PAM sequence (which would include a large number of duplicates), using script 'remove_pam_-mismatches.py'. To compute summary off-target scores, each off-target match was assigned a score using the formula described by Hsu *et al.*[176], based on a matrix of weights for mismatches at each position in the 20-mer:

$$score_a = 100 \times \prod_{i=1}^{20}(1 - (mismatch_i \times weight_i))$$

where $mismatch_i$ is assigned the value of 0 if nucleotide i matches the target, or 1 for a mismatch. If more than one mismatch is present, this score is further modified according to the number of mismatches and their proximity to each other:

$$score_b = \frac{score_a}{\left(\frac{4\times\left(19-\frac{\sum_{k=1}^{n-1}(pos_{k+1}-pos_k)}{n-1}\right)}{19}+1\right)\times 2^n}$$

where n is the total number of mismatches, and pos_k is the position of mismatch k within the 20mer (if listed sequentially). Finally, all of these individual mismatch scores are aggregated to a single summary score y using the formula

$$y = \frac{100}{\sum_{x\in S} x}$$

for set S of all individual match scores, including one score of 100 reflecting the 'on-target' match to the intended target. This is implemented in the script 'offtarget_analysis.py'. In later iterations of the pipeline, this scoring system was updated to use the Doench cutting frequency determination score for individual off-target matches, which takes into account the identity as well as the position of mismatched bases, and also weights scores by PAM sequence in the off-target match.[173] These were aggregated using the above equation, after converting to a 0-100 scale.

On-target and off-target scores were collated, and additional parameters calculated for each guide including guanine/cytosine percentage, maximum length mononucleotide repeat stretch (e.g. CCCCC), and cutting distance from the target variant, using script 'Ontarget_score_processing.py'. A distance of 0 was defined as a predicted Cas9 cut site immediately adjacent to the variant on either side, or within the reference sequence of the variant for longer indels. Guides were filtered and selected using script 'guide_filtering.py'. First, on-target scores were weighted according to target PAM, using the weights in Table 2.9. Weightings were arbitrarily assigned according to a subjective evaluation of the consistency of evidence and maximal efficiency for each.[174] Calculated on-target scores were multiplied by these weightings, so that guides with high-efficiency PAMs such as NGG were preferred. Next, guides not meeting the minimal requirements in Table 2.9 were removed. Since the screen output (SIRPα expression) was relatively specific, off-target effects at distant sites in the genome were considered less likely to affect outcome than in other screen designs, and so a permissive off-target threshold (allowing one exact mismatch) was selected to improve target coverage. The numeric variables for cutting distance from the variant, on-target score and off-target score were then each divided into three bins, to allow sequential sorting based on these ordinal categories, in the above order: this avoids guides being ranked solely on the first continuous variable selected. Rounded scores to one decimal place were used to break ties where necessary, and the top guides were selected for each target, to a maximum of 14 guides per variant. Duplicate guides were then removed and multiple targets identified where appropriate for targets located less than 30 base pairs apart. Additional human non-targeting guides were selected randomly from a previously published list[177] as negative controls. Finally, cloning flank sequences were added to each guide to allow cloning into the plasmid vector[172]:

```
Left flank: 5'-TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCG-3'
Right flank: 5'-GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGT-3'
```

Parameter	**Criterion**
Cut distance from variant	$\leq$15 bp
Weighted on-target score	$\geq$0.2
Off-target score	$\geq$0.2
Exact off-target matches	$\leq$1
GC%	$\geq$25%
Max. mononucleotide repeat stretch	$<$5
PAM weights: - spCas9: - xCas9:	 NGG: 1.0; others: 0 NGG: 1.0; NGA, NGT: 0.7; NAG, NGC, NCG: 0.6; NTG, GAN, CAT: 0.5; others: 0

Table 2.9 – Guide criteria used in targeted CRISPR sgRNA library design. Weighted on-target scores were calculated by multiplying raw on-target scores by the relevant PAM weights.

2.18.2 Guide library construction and amplification

Designed guide sequences with appended cloning flanks were synthesised as pooled DNA oligonucleotides by Twist Bioscience. The pooled oligo library was amplified by PCR, with forward and reverse primers to recognise the cloning flank sequences (Table 2.10). Reactions were set up with 1 µl template (at 1 ng/µl), 1.25 µl of each primer (at 10 µM), 12.5 µl NEBNext High Fidelity PCR MasterMix (containing Q5 high-fidelity polymerase; New England Biolabs) and nuclease-free water to 25 µl total reaction size. The amplification reaction was performed in an MJ thermal cycler (MJ Research), with an initial denaturation step of 98℃ for 30 seconds, 22 cycles of denaturation at 98℃ for 10 seconds, annealing at 63℃ for 10 seconds and extension at 72℃ for 15 seconds, and finally a two-minute extension step at 72℃.

PCR products were purified using a DNA Clean & Concentrator-5 Kit (Zymo Research). The reaction products were added to five times the volume of DNA

binding buffer, and placed in a Zymo-Spin™ I-C spin column. The column was centrifuged for one minute at 11,000*g*, and the supernatant discarded. The column was washed twice with DNA Wash buffer, centrifuging as above. The DNA was eluted with 8 µl nuclease-free water, and quantified by spectrophotometry (Nanodrop; Thermo Fisher Scientific). A sample of the purified product was run on a 2% agarose gel for 45 minutes at 85 kV, confirming a single pure product of approximately 140 base pairs (bp).

2.18.3 Library cloning into lentiviral plasmids

The lentiCRISPRv2 plasmid (containing *Streptococcus pyogenes*-derived sp-Cas9, an sgRNA scaffold and a puromycin resistance cassette) was a gift from Feng Zhang (Addgene plasmid #52961).[177] LentiCRISPRv2-xCas9 was donated by Tim Regan. This plasmid was created by exchanging the spCas9 sequence in LentiCRISPRv2 with the xCas9 sequence from xCas9 3.7 (Addgene plasmid #108379)[174], and subsequently substituting a synthesised gene fragment for a portion of the xCas9 sequence to remove an unwanted BsmbI restriction site.

The plasmids were digested with FastDigest Esp3I (BsmbI equivalent; Thermo Fisher Scientific). For each plasmid, two 60 µl reactions were set up with 5 µl (approximately 5 µg) plasmid, 3 µl FastDigest Esp3I, 3 µl Fast Alkaline Phosphatase (Thermo Fisher Scientific), 6 µl fast digest buffer, 0.6 µl 100mM dithiothreitol, and 42.4 µl nuclease-free water, with digestion for 30 minutes at 37℃. The digested products were run on a 1% UltraPure™ low melting point agarose gel (Thermo Fisher Scientific), for 90 minutes at 110 kV, together with a GeneRuler 1 kb-Plus DNA ladder (Thermo Fisher Scientific), confirming the expected bands of approximately 13,000 bp and 2000 bp with no evidence of residual uncut plasmid.

The larger band was excised and purified using a Zymoclean™ Gel DNA Recovery Kit (Zymo Research). One volume of agarose gel was added to three volumes of Agarose Dissolving Buffer, and incubated for 10 minutes at 55℃ until completely dissolved. An additional one volume of nuclease-free water was added, to improve DNA recovery for large fragments. The solution was then applied to a Zymo-Spin™ spin column in a collection tube and centrifuged for one minute at 11,000*g*. The flow-through was discarded, and the column was

Primer	Sequence (5'-3')
Olignucleotide pool amplification	
Oligo_Fwd	GTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTG GAAAGGACGAAACACC
Oligo_KO_Rev	ACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCT ATTTCTAGCTCTAAAAC
Illumina library preparation	
NGS-Lib-Fwd-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTTAAGTAGAGGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTATCATGCTTAGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTGATGCACATCTGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTCGATTGCTCGACGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTTCGATAGCAATTCGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-6	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTATCGATAGTTGCTTGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-7	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTGATCGATCCAGTTAGGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-8	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTCGATCGATTTGAGCCTGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-9	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTACGATCGATACACGATCGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-Fwd-10	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCG ATCTTACGATCGATGGTCCAGAGCTTTATATATCTTGTGGAAAGGACGAAACACC
NGS-Lib-KO-Rev- 1	CAAGCAGAAGACGGCATACGAGAT**TCGCCTTG**GTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA
NGS-Lib-KO-Rev- 2	CAAGCAGAAGACGGCATACGAGAT**ATAGCGTC**GTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA
NGS-Lib-KO-Rev- 3	CAAGCAGAAGACGGCATACGAGAT**GAAGAAGT**GTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA
NGS-Lib-KO-Rev- 4	CAAGCAGAAGACGGCATACGAGAT**ATTCTAGG**GTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA
NGS-Lib-KO-Rev- 5	CAAGCAGAAGACGGCATACGAGAT**CGTTACCA**GTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA
NGS-Lib-KO-Rev- 6	CAAGCAGAAGACGGCATACGAGAT**GTCTGATG**GTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA

Table 2.10 – Primers for guide pool amplification and Illumina library construction. Primer sequences are from Joung *et al.* (2017).[172] Barcode sequences are underlined.

washed twice in DNA Wash buffer, centrifuging and discarding the flow-through at each step. To elute the DNA, 20 µl elution buffer (10 mM Tris-HCl, pH 8.5, 0.1 mM EDTA.) was applied to the column and incubated for two minutes, before centrifuging as above. The eluted DNA was further purified and concentrated using a Zymo DNA Clean & Concentrator Kit, as above, and quantified by Nanodrop spectrophotometry.

The amplified oligo pool was cloned into the digested vectors by Gibson Assembly. For each reaction, 330 ng digested plasmid and 110-220 ng pooled amplified oligos were added to 10 µl Gibson Assembly® MasterMix (New England Biolabs), including T5 exonuclease, Q5 polymerase, and a DNA ligase enzyme, and diluted to a total volume of 20 µl with nuclease-free water. Negative control reactions were set up omitting the insert pool. The reactions were performed in an MJ Thermocycler, for 1 hour at 50℃. The reaction products were purified with the Zymo DNA Clean & Concentrator Kit, as described above, eluting in 6 µl nuclease-free water. The purified products were transformed into Endura™ Electrocompetent Cells (Lucigen Inc.). After thawing the cells on ice, 2 µl of purified Gibson product was added to 25 µl of cells, and the mixture was pipetted into a chilled electroporation cuvette with a 1 mm gap. Electroporation was performed in a Gene Pulser II electroporator (Bio-Rad Inc.), at 1800 Volts, 600 Ohms, and capacitance of 10 µF, after which 1975 µl Recovery Medium (Lucigen Inc.) was immediately added, and cells were transferred to a culture tube for recovery in a shaking incubator at 37℃ for one hour. The cells were subsequently inoculated onto LB agar plates containing 100 µg/ml ampicillin: 100 µl of a 1:10,000 dilution was plated onto 10 cm round plates for estimation of total colony count, and the remainder was inoculated onto 265 mm x 265 mm square plates. Comparison of colony counts between complete reactions and vector-only negative controls showed a ratio of 7-10:1, indicating low levels of residual empty vectors or incorrectly ligated plasmid.

A total of 400,000 - 600,000 colonies were harvested per plasmid preparation, by washing off the agar plates with 50 ml LB broth. The plasmids were purified using a Plasmid Plus MaxiPrep Kit (Qiagen), according to the manufacturer's instructions. The bacterial suspension was centrifuged at 3,000*g* for 30 minutes at 4℃, and the supernatant was discarded. The pelleted bacteria were resuspended in 5 ml buffer P1, containing RNase A, and were lysed by adding 5 ml lysis buffer P2, and incubating at room temperature for 3 minutes. Five

millilitres of buffer S3 were added to neutralise the reaction, and the lysate was transferred to a QIAfilter cartridge and incubated for 10 minutes. The lysate was filtered into a clean tube, and 5 ml binding buffer added. The cleared lysate was transferred to a spin column on a vacuum manifold, and drawn through the column with a vacuum pressure of approximately 300 mbar. The plasmid DNA bound to the column was then washed sequentially with 0.7 ml Endotoxin Removal buffer and 0.7 ml wash buffer PE, drawing the fluid through via vacuum for each wash. The spin column was transferred to a collection tube and centrifuged at 10,000*g* for one minute to remove residual wash buffer, and the plasmid DNA was eluted by adding 400 µl elution buffer (10 mM Tris-HCl, pH 8.5, 0.1 mM EDTA) to the column, incubating for two minutes, and centrifuging at 10,000*g* for one minute. DNA was quantified by Nanodrop spectrophotometry.

2.18.4 Cloning for single sgRNAs

To clone single guides into the LentiCRISPR-v2 vector for validation, a complementary oligonucleotide pair was synthesised for each guide, with overhangs for complementarity to the Esp3I restriction sites in the digested vector:

```
5'-CACCGNNNNNNNNNNNNNNNNNNN    -3'
3'-    CNNNNNNNNNNNNNNNNNNNCAAA-5'
```

To anneal the oligonucleotide pairs, 10 µl reactions were set up with 1 µl of each oligonucleotide at 100 µM, 1 µl T4 ligation buffer containing 10mM ATP, 0.5 µl T4 polynucleotide kinase (New England Biolabs) and 6.5 µl water, and incubated at 37℃ for 30 minutes, then 95℃ for 5 minutes followed by a temperature ramp down to 25℃ at a rate of 6℃/minute. Annealed oligonucleotides were diluted 1:200 in nuclease-free water, and ligation reactions were set up with 50 ng Esp3I-digested LentiCrispr-v2 plasmid, 1 µl diluted annealed oligonucleotides, 5 µl Quick Ligase Buffer, 1 µl Quick Ligase (New England Biolabs), and water to 11 µl, at room temperature for 10 minutes. Two microlitres of the ligation reaction was transformed into Stbl3 recombination-deficient bacteria (propagated from stocks originally obtained from Invitrogen / Thermo Fisher Scientific). The bacteria were incubated with the ligation product for 10 minutes on ice, after

which a 30-second heat shock was applied at 42℃. Ten volumes of SOC Recovery Medium (Sigma Aldrich) were added, and after one hour of recovery in a shaking incubator at 37℃, the bacteria were inoculated onto LB agar plates with 100 µg/ml ampicillin.

After overnight culture at 37℃, a single colony was selected and subcultured overnight in 5 ml LB broth with 100 µg/ml ampicillin, and the amplified plasmid was purified with a ZymoPURE™ Plasmid Miniprep kit (Zymo Research). The bacterial culture was centrifuged at 3000*g* for 6 minutes, and after discarding the supernatant, the cell pellet was resuspended in 250 µl ZymoPURE™ P1 resuspension buffer. 250 µl ZymoPURE™ P2 lysis buffer was then added and mixed by inversion, and the lysis reaction was incubated for three minutes at room temperature. 250 µl ice-cold P3 buffer was added to neutralise the reaction, and after gentle mixing the lysate was incubated for 5 minutes on ice, and then centrifuged for 5 minutes at 16,000*g*. 600 µl cleared supernatant was transferred to a clean tube, and 275 µl binding buffer was added. The solution was transferred to a Zymo-Spin™ II-P spin column, incubated at room temperature for two minutes, and centrifuged at 5,000*g* for one minute. The flow-through was discarded, and the column was washed three times, with 800 µl Wash Buffer 1, 800 µl Wash Buffer 2, and 200 µl Wash Buffer 2, centrifuging and discarding the flow-through as above for each wash. The column was centrifuged at 11,000*g* for 1 minute to remove any residual wash buffer, and the plasmid DNA was eluted by incubating the column for two minutes with 25 µl Elution Buffer (10 mM Tris-HCl, pH 8.5, 0.1 mM EDTA), and centrifuging for 1 minute at 11,000*g*. The DNA was quantified by Nanodrop spectrophotometry, and 200-500 ng was sent for Sanger sequencing (Eurofins Genomics) with 2.5 µl 10µM U6 sequencing primer (5'-GACTATCATATGCTTACCGT-3'), to verify that the inserted guide sequence was correct. Following verification, a 50 ml subculture of the original stock was grown overnight (in LB broth with 100 µg/ml ampicillin), and a plasmid maxiprep was performed as described above.

2.18.5 Lentiviral packaging and titration

Plasmids containing Cas9 and sgRNA-encoding sequences were packaged into lentivirus using the LV-Max™ Lentiviral Production System (Gibco). Lentiviral Production Cells were cultured in proprietary low-serum Lentiviral Production

Medium at 37℃ and 8% CO_2, in an orbital shaker at 125 rpm. They were maintained at a concentration between 3 x 10^5 and 5.5 x 10^6 cells / ml, and were used for lentiviral production between passages 7 and 16. On the day of transfection, cells were diluted or concentrated as appropriate to a concentration of 4.7 x 10^6 in 25.5 ml medium, and 1.5 ml of proprietary Lentiviral Production Supplement was added. The lentiviral transfer vector and packaging plasmids were diluted in 1.5 ml Opti-MEM™ reduced serum medium (Thermo Fisher Scientific), in the proportions shown in Table 2.11:

Plasmid	**Quantity**
Lentiviral transfer plasmid: LentiCRISPR-v2 or LentiCRISPRv2-xCas9	35.8 µg
Packaging plasmid: psPAX2 (Addgene #12260)	23 µg
VSV-G envelope plasmid: pMD2.G (Addgene #12259)	16.2 µg

Table 2.11 – Plasmids used for lentivirus packaging. Quantities shown are for a 30ml culture volume, and were scaled up proportionately if necessary for larger-scale packaging. psPAX2 and pMD2.G were gifts from Didier Trono.

LV-Max™ Transfection Reagent was diluted in Opti-MEM™ medium (180 µl in 1.5 ml), vortexed briefly and incubated for one minute at room temperature, before the DNA solution was added and the two solutions mixed gently. The combined solution was incubated for 30-45 minutes at room temperature. The complexed DNA and transfection reagent mixture was added to the diluted cells and mixed by shaking. The cells were incubated under the above conditions, with addition of 1.2 ml proprietary Lentiviral Production Enhancer after 5-11 hours. Lentivirus was harvested after approximately 48 hours: the cell suspension was centrifuged at 1300*g* for 15 minutes, and the supernatant was passed through a 0.45 µm low protein-binding filter to remove cellular debris. The filtered supernatant was stored at -80℃ until needed for transduction.

As accurate titration in the target cell population, THP-1 cells, is difficult due to non-adherent growth and a narrow range of concentrations in which growth

rate remains predictably log-linear, initial titration of the lentivirus stock was performed by transduction of HEK293T cells (ATCC®). The cells were plated in 24-well plates at 125,000 cells per well, in DMEM complete medium. Lentivirus supernatant was added to wells in duplicate, with 0, 15.6, 31.25, 62.5, 125 or 250 µl per well, in a total of 725 µl medium per well, with polybrene (Sigma-Aldrich) at 8 µg/ml. 'Spinfection' was performed, by centrifuging the plates at 1000*g* for two hours, at 32℃, after which the cells were resuspended to allow even seeding and cultured on at 37℃ and 5% CO_2. The media was changed after 24 hours to DMEM complete medium without polybrene, and changed again after a further 24 hours to DMEM complete medium, with 1 µg/ml puromycin (Thermo Fisher Scientific) in one of each pair of wells for a specific virus dilution. After all cells in the untransduced, puromycin-treated wells were dead and non-adherent (but before other wells had reached confluence), the cells were lifted with TrypLE™ Express (Thermo Fisher Scientific) and counted in a haemocytometer after staining with 0.2% Trypan Blue (Sigma Aldrich) to exclude dead cells. The multiplicity of infection (MOI) was calculated as the ratio of cell counts in the puromycin-treated to untreated wells for each concentration, and the titre of the virus stock was determined with the formula $Titre = \frac{starting\ cell\ no. \times MOI}{volume}$. Titres were reconfirmed in THP-1 cells, as described below.

Titration using luminescence-based adenosine triphosphate (ATP) quantification as a surrogate for cell viability, with CellTiter-Glo® (Promega Corporation), was used as an alternative method in some cases. HEK-293T cells were seeded at 3,125 cells per well in a 96-well white-walled plate, and transduced by spinfection as described above, with serial ten-fold dilutions of lentivirus stock at 50 µl per well. Transductions were performed in triplicate. A standard curve of serially diluted, untransduced cells was set up on the same plate. Puromycin was added to all transduced wells, but not the standard curve, after 24 hours. The medium was changed after three days, and viability assessed five days after addition of puromycin. To measure viability, the plate was equilibrated to room temperature, medium was partially removed leaving 100 µl per well, and 100 µl CellTiter-Glo®Reagent was added to each well. The plate was placed on an orbital shaker for two minutes to induce cell lysis, and incubated for 10 minutes at room temperature. Luminescence was measured using a GloMax® Multi plate-reading luminometer (Promega Corporation), with a 0.5 second integration time. Background luminescence was subtracted based on readings

from medium-only wells, and the proportion of surviving cells (and hence MOI) for each virus concentration was calculated from the standard curve of values in untransduced cells.

2.18.6 Lentiviral transduction and pooled CRISPR screening in THP-1 cells

THP-1 cells were obtained from stocks propagated in the David Hume laboratory, originally from ATCC®. The cells were cultured in antibiotic-free RPMI complete medium with 10% FBS, 2 mM L-alanyl-L-glutamine dipeptide (Gibco™ GlutaMAX™, Thermo Fisher Scientific), and non-essential amino acids (Gibco™ MEM Non-Essential Amino Acids Solution, Thermo Fisher Scientific) at 37℃ and 5% CO_2. The cells were maintained between a concentration of 10^5 and 1.2 x 10^6 cells / ml to avoid altered growth kinetics or differentiation at low or high concentrations.

For targeted CRISPR screening in THP-1 cells, 16.7 x 10^6 cells were plated at 833,000 cells per well in 1 ml medium in 6-well plates. The cells were transduced with the lentiviral library at MOI 0.3, to give 5 x 10^6 transduced cells (5,000 per sgRNA): the lentivirus stock was diluted to 125,000 transducible units per ml in complete culture medium, and 2 ml added to each well. Polybrene was not used due to toxicity in this cell line. Additional wells were set up as untransduced or transduced controls. Spinfection was performed by centrifuging at 32℃ for one hour at 1000*g*, after which cells were resuspended and returned to the incubator. The culture medium was changed after 24 hours, and puromycin was added at 1 µg/ml after 48 hours, except in the relevant control wells. At this time a standard curve was set up of transduced cells to confirm the calculated MOI, without assumptions as to growth kinetics: transduced wells were diluted to 10%, 30%, 50% or 100% of their cell concentration, and cultured without puromycin. Three days after addition of puromycin, when no viable cells remained in untransduced, puromycin-treated wells, viable cells (identified by Trypan Blue exclusion) were counted and the actual screen MOI calculated by comparison to the standard curve of transduced cells without puromycin. The remaining cells were cultured for seven days of puromycin selection, passaging as necessary, after which 10^8 cells were harvested for SIRPα staining.

The harvested cells were centrifuged at 450*g* for 5 minutes, and resuspended in 4 ml PBS with 2% FBS. The cells were incubated with recombinant anti-SIRPα antibody (clone REA144, Miltenyi Biotec) conjugated to phycoerythrin (PE), at 0.6 ng/ml, for 1 hour at 4℃. Additional cells were stained with a PE-conjugated isotype control antibody (clone REA 293, Miltenyi Biotec), as a negative control. After surface staining, the cells were washed in PBS and stained with viability dye eFluor™ 780 (eBioscience™, Thermo Fisher Scientific), at a dilution of 1:4000 in PBS, for 15 minutes on ice. The cells were then washed again and resuspended in PBS with 2% FBS, and were filtered through a 70 µm cell strainer and stored on ice in the dark prior to sorting. Low (5%), high (5%) and mid-range (20% of the cell population) SIRPα-expressing cells were isolated by fluorescence-activated cell sorting in a FACSAria™ cell sorter (BD Bioscience), after gating on live single cells. A total of 5,000 to 10,000 cells per guide were collected.

The sorted cell pools were centrifuged at 1,000*g* for 5 minutes, and genomic DNA was extracted from the cell pellets using a Quick-DNA™ Midiprep Plus kit (Zymo Research). The cells were resuspended in 1 ml DNA elution buffer, and 1 ml BioFluid and Cell Buffer was added with proteinase K at 0.3 mg/ml. The samples were vortexed briefly and incubated at 55℃ for two hours, and then 2 ml Genomic Lysis Buffer was added. The lysate was transferred to a Zymo-Spin™ V-E spin column in a 50 ml conical tube and centrifuged at 1,000*g* for 5 minutes to allow genomic DNA to bind to the column. The flow-through was discarded, and the column was washed successively with 9 ml DNA Pre-Wash Buffer and 7 ml g-DNA Wash Buffer, centrifuging as above for each wash. The spin column was transferred to a collection tube and washed with a further 200 µl wash buffer, before centrifuging again at 12,000*g* for one minute to remove residual wash buffer. DNA was eluted by adding 200 µl Elution Buffer (10 mM Tris-HCl, pH 8.5, 0.1 mM EDTA) at 60℃, incubating for 5 minutes, and centrifuging at 12,000*g* for 1 minute, passing the elution buffer through the column twice to maximise yield. The genomic DNA was quantified by Nanodrop spectrophotometry.

2.18.7 Amplification and sequencing of integrated guide RNA constructs

For next generation sequencing of the sgRNA protospacer sequences integrated into the genome by lentiviral transduction, Illumina libraries were created by PCR amplification with primers recognising the guide flank sequences and containing the necessary Illumina adapters and primer binding sites. For each sample, ten reactions were set up, together incorporating the entire genomic DNA sample, each with a different forward primer (to increase sequence diversity) and a single barcoded reverse primer to allow multiplexing. Primer details are given in Table 2.10. Each 50 µl reaction was set up with 20 µl DNA template, 25 µl NEBNext® High-Fidelity PCR MasterMix, 1.25 µl each of the forward and reverse primers, and 2.5 µl nuclease-free water. Reactions were run on an MJ Thermal Cycler, with an initial 3-minute denaturation step at 98℃ followed by 24 cycles of 10 seconds' denaturation at 98℃, 10 seconds' annealing at 63℃, and 25 seconds' extension at 72℃, with a final extension step of 2 minutes at 72℃.

The ten PCR reactions for each cell pool were combined and purified using the Zymo DNA Clean & Concentrator kit, scaled up proportionately for the higher total sample volume and modified by using a higher binding capacity Zymo-Spin™ V column and 150 µl elution volume to allow for the higher expected yield. Aliquots of 25 µl were run on a 2% UltraPure low melting point agarose gel for one hour at 90 kV, with a 50 bp DNA ladder (Thermo Fisher Scentific). The band at the expected library size of 260-270 bp was excised and purified using the Zymoclean™ Gel DNA Recovery Kit (Zymo Research) as described above, eluting in 20 µl elution buffer. The extracted products were further purified using the Zymo DNA Clean & Concentrator kit, pooling multiple gel extractions for a single cell pool where required. The purified libraries were quantified with Nanodrop spectrophotometry and pooled approximately in proportion to the number of contributing cells. The pooled sample, at a total concentration of 93 nM, was submitted to GeneWIZ Ltd. for next generation sequencing.

After confirmation of library concentration, purity and expected size distribution by Qubit fluorometry (Invitrogen) and TapeStation electrophoresis (Agilent) at the sequencing facility, the prepared libraries were sequenced on an Illumina Hi-Seq sequencer with 20% PhiX spike-in (to improve sequence diversity), at

2 x 150 bp paired end reads to give a total of approximately 350 million reads for two screen experiments. Sequence data were demultiplexed using index barcodes in the reverse primers, and returned as FASTQ files.

2.18.8 Screen Analysis

Analysis of sgRNA enrichment used a modified version of an algorithm previously described for analysis of genome-wide CRISPR screens.[67] First, reads matching screen sgRNAs (when flanked by the expected sequences from the promoter and guide scaffold) were counted in forward read FASTQ files for each cell pool, using the script 'count_sgrnas.py'. To minimise the chance of sgRNA misidentification, a strict quality filter was applied, only including reads with a minimum Phred score of 27 (equating to $<$0.2% chance of an incorrect base call) for all bases in the 20-mer.

Local z-scores were then calculated for each guide, for each pairwise comparison between cell pools, using the script '1_local_zscore.py'. This first normalises count data across all samples using the quantile normalisation function in R package *preprocessCore*, run in R version 3.5.3 (Figure 2.2A). Minimum distance d from the expected line of equality $y = x$ for a pairwise comparison was calculated using the formula $d = \sqrt{(y_n - x_n)^2/2}$ if $y_n > x_n$ or $d = -\sqrt{(y_n - x_n)^2/2}$ if $x_n > y_n$ where x_n and y_n are the normalised read counts for a single sgRNA in the two pools being compared. These were converted to normalised z-scores, $z = (d - \mu)/\sigma$, using mean and standard deviations based on sliding bins of 100 distance values to account for heteroscedasticity across the range of count values: thus for read pairs ranked in order of their nearest position on the $y = x$ line, the z-score for pair rank n from a total of T pairs will be based on pairs 1 to 100 if $n \leq 50$, pairs $(T - 99)$ to T if $n \geq (T - 50)$, or otherwise pairs $(n - 50)$ to $(n + 49)$ (see Figure 2.2B).

To estimate the probability that the observed sgRNA enrichment for each target could have been observed by chance, the mean z-score for all guides predicted to cut within 15 base pairs of each target was first calculated (allowing for some guides to be assigned to more than one target), in script 'crispr_permute_multitarget.py' To test the null hypothesis that the mean z-score for a target is equal to the mean for all guides regardless of target specificity, the absolute value

of this score was then compared to an empirical probability distribution constructed from random permutations of the data. This probability distribution was generated by Gaussian kernel density estimation (in package *scikit-learn* v.0.22.1, using the 'Scott' bandwidth estimation method; see Figure 2.2C) of 10^6 scores calculated as absolute mean values of an equal number of z-scores to that associated with the target in question, randomly sampled without replacement from the full dataset. The p-value was then calculated by integration of this distribution from the observed value to infinity: since the distribution was derived from absolute rather than raw values, this is a two-tailed test. Analysis at the single-guide level was performed in the same way, except that only non-targeting guides were used to construct the reference probability distribution. The Benjamini-Hochberg procedure was used to control the false discovery rate (<0.05): the 'FDR' value reported for individual results (commonly referred to as the 'adjusted p-value') in this text is defined as the minimum false discovery rate threshold that would allow rejection of the null hypothesis for that result.

2.18.9 Flow cytometry for single target validation

To validate single targets, THP-1 cells were transduced with lentivirus containing spCas9 with a single sgRNA construct (in LentiCRISPR-v2), using the transduction methods above with an MOI of approximately 0.5. Negative control lentivirus contained either an empty vector (i.e. with Cas9 but no sgRNA) or a non-targeting sgRNA (see Table 2.12). Transductions were performed in triplicate. The cells were cultured under puromycin selection, passaging as necessary, until all cells were dead in untransduced puromycin-treated controls and viability was greater than 90% in transduced puromycin-treated wells (7-10 days). Cells were harvested for DNA extraction and for flow cytometry.

Surface staining for SIRPα and SIRPβ1 was performed in a round-bottomed 96-well plate. Harvested cell suspensions, of approximately 100,000 cells per sample, were centrifuged at 450*g* for 4 minutes and the culture medium was discarded. The cells were washed in PBS with 2% FBS (FACS buffer), and resuspended in 50 µl FACS buffer with PE-conjugated anti-human SIRPα antibody (REA144, Miltenyi Biotec) and APC-conjugated anti-human SIRPβ antibody (mouse IgG1κ, clone B4B6, Miltenyi Biotec) both at 0.6 ng/ml. Samples

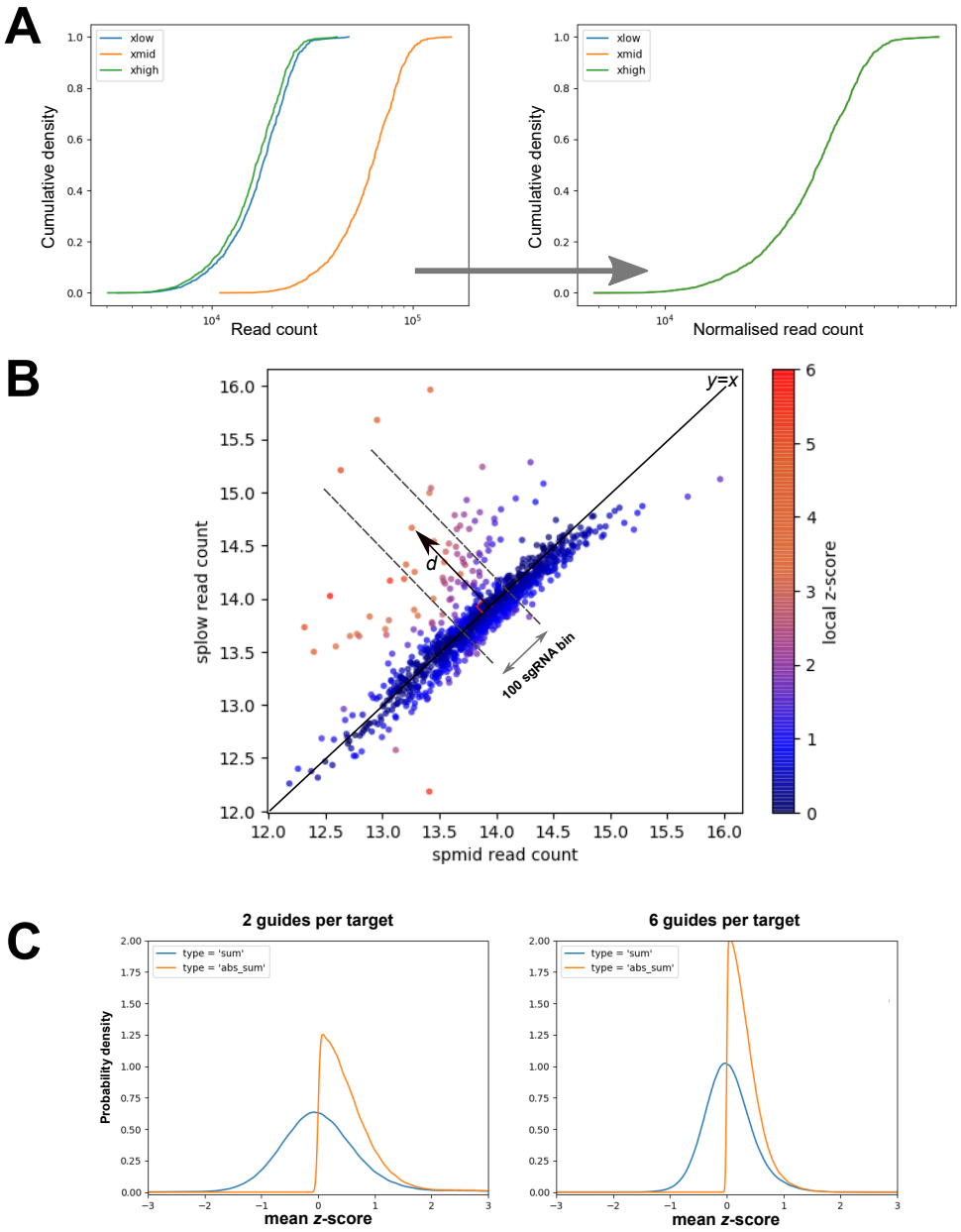

Figure 2.2 – Algorithm for analysis of sgRNA enrichment. A: Quantile normalisation of read counts across three cell pools. B: z-scores are calculated within 100-sgRNA bins, based on distance d perpendicular to line $y = x$, for each pairwise comparison. C: Kernel density estimates for the empirical probability distribution for mean z-score for a target with 2 or 6 guides. The blue ('sum') curve gives the distribution for the raw mean score, whereas the orange ('abs_sum') gives the distribution for the absolute value of the mean score. Single-direction integration of the latter gives a 2-tailed p-value.

stained with matched isotype control antibodies (clone REA293 and clone IS5-21F5, Miltenyi Biotec), with the same fluorochromes and at the same concentrations, were used as negative controls, and additional single-stained controls were used for compensation setup. The samples were incubated with the surface staining antibodies for one hour at 4℃ in the dark, then washed in PBS and stained with fixable viability dye eFluor™ 780 (eBioscience™, Thermo Fisher Scientific), diluted 1:4000 in PBS, for 20 minutes at 4℃. The cells were washed twice in FACS buffer, and fixed in equal parts FACS buffer and neutral buffered formalin for 20 minutes at room temperature. The cells were then washed again and resuspended in FACS buffer for analysis.

Guide / primer	**Sequence (5' - 3')**
sgRNA Targets:	
rs4813319	ATGGAAATGAAACGAGAAGC
Non-targeting 1	ACTGCTCCCGGTCGCCCCTC
Non-targeting 2	CCAGTTATAATTAGGGGTTT
Sequencing primer:	
Human U6 Forward	GACTATCATATGCTTACCGT
Target region PCR primers:	
rs4813319 Forward	TTCCAGTTGGTCCAGTCTCG
rs4183319 Reverse	CACCCAGTCTTGACTCTGGC

Table 2.12 – Protospacer (sgRNA target) sequences and primers used in single target validation.

Flow cytometry was performed on a BD LSRFortessa™ cytometer (BD Biosciences), collecting at least 5,000 events per sample. Cytometry data were analysed in FlowJo version 10.5.3 (FlowJo LLC), using single-stained samples to adjust compensation for spectral overlap, and using isotype control-stained samples to assist with gating. Debris and doublets were excluded based on forward and side scatter characteristics, and dead cells excluded based on eFluor 780 uptake. Surface expression of SIRPα and SIRPβ1 was assessed both on overall distribution and on median fluoresence intensity for the relevant fluorochromes.

2.18.10 Cutting efficiency analysis for single guide validation

To assess Cas9 cutting efficiency for individual sgRNAs, THP-1 cells transduced with a single guide were harvested after puromycin selection as above. DNA extraction and target site PCR, from both these transduced cells and untransduced control cells, were performed with the Phire™ Tissue Direct PCR Kit (Thermo Fisher Scientific). Primers were designed (using CrispOR[178]) to amplify a region of approximately 450 bp surrounding the sgRNA target. Pelleted cells were diluted in 40 µl Dilution Buffer with 1 µl DNA Release Reagent, and heated to 98℃ for 2 minutes to release DNA. The resulting lysate was diluted 1:1 in nuclease free water, and 1 µl was used as the PCR template, in a 20 µl reaction with 10 µl Phire™ Tissue Direct MasterMix and 1 µl of each primer (to a final concentration of 0.5µM). Reactions were run on an MJ Thermal Cycler with an initial 5-minute denaturation step at 98℃ followed by 40 cycles of 5 seconds' denaturation at 98℃, 5 seconds' annealing at a primer pair-specific annealing temperature (65℃ for the primers listed in Table 2.12), and 2 seconds' extension at 72℃, with a final extension step of 1 minute at 72℃. The PCR products were run on a 1% agarose gel at 100 kV for one hour, together with a GeneRuler 1 kb Plus ladder (Thermo Fisher Scientific), and the appropriate band was excised and purified using the gel extraction method described above.

Purified amplified target site fragments from transduced and control cells were submitted for Sanger sequencing (Eurofins Genomics) using the forward PCR primer as a sequencing primer. Cutting efficiency and indel length distribution were estimated using Synthego ICE Analysis v2.0 (Synthego).[179] This algorithm evaluates cutting efficiency by comparing the change in Sanger chromatogram noise before and after the predicted Cas9 cut site, between transduced and control cells, and uses regression analysis of the chromatogram trace to estimate the proportion of each potential repair outcome (related to indel length) contributing to the overall sequence pool (Figure 2.3).

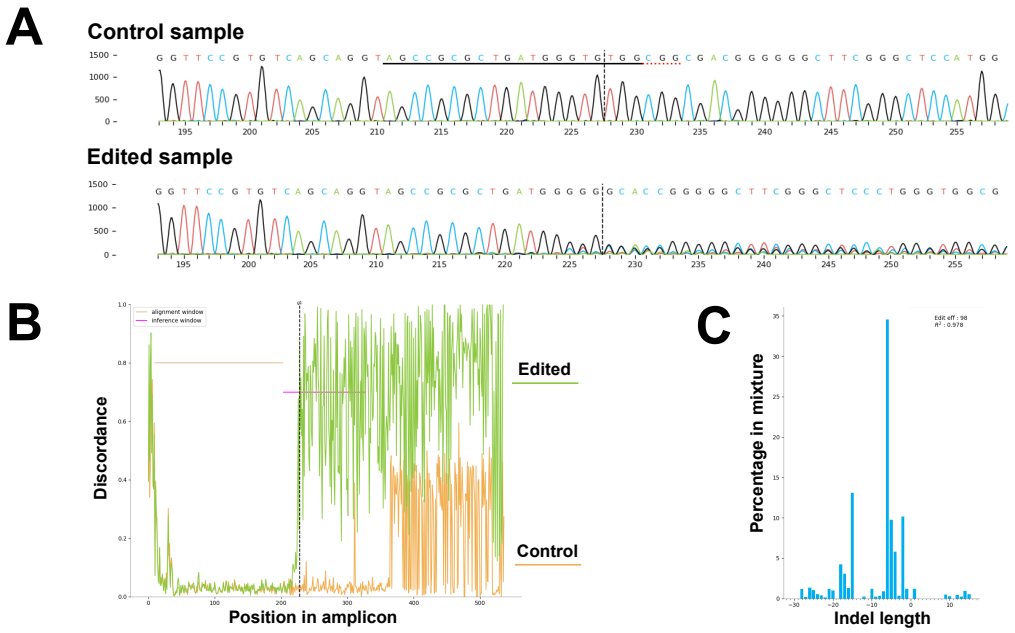

Figure 2.3 – Example of CRISPR cutting and indel analysis from chromatogram data. This example shows analysis with Synthego ICE v2.0 for a highly efficient sgRNA. A: Sanger chromatograms from control and edited cells, showing increased noise close to and downstream of the cut site (vertical dashed line). B: Quantification of trace discordance of the chromatograms in A. (Note - signal discordance from position 360 in the control sample in this example is due to heterozygosity for a naturally-occurring indel variant.) C: Estimated indel length frequency based on regression analysis of the chromatogram.

Chapter 3

CD97 modulates the T-cell response to infection with influenza A virus

3.1 Introduction

CD55 is one of the few human genes for which a genetic association with influenza has been replicated across multiple independent populations. This gene has to date only been implicated in disease severity, rather than susceptibility to infection more broadly, suggesting a role either in modulating host tissue damage or in viral clearance in later stages of infection. Mouse models also support a role in modulation of immunopathology with some IAV strains, without a measurable effect on viral clearance, although in this context CD55 deficiency appears to be protective.[149] While the regulatory effects of CD55 on complement activation have been assumed to underlie its effects on disease severity, supported by data from mouse models, this has by no means been proven, and it has other physiological functions which could be involved. CD55 is a known ligand for CD97, an adhesion G-protein-coupled receptor highly expressed on mononuclear phagocytes and granulocytes.[180] The CD97-CD55 interaction is postulated to have multiple regulatory functions in the immune system, such as in leukocyte adhesion or migration or in T-cell stimulation, independent of complement. In this chapter, prompted both by a novel genetic association be-

tween *CD97* and severe influenza, and by its proposed immune functions and its known interaction with CD55, I investigate whether CD97 itself plays a role in the host response to infection with influenza A virus.

3.1.1 A variant in the *CD97* gene was associated with influenza severity in the 2009 H1N1 pandemic.

Previous work in our laboratory (Baillie, unpublished data) identified candidate genetic variants linked to influenza severity in an exome-wide association study of patients from England (Mechanisms of Severe Acute Influenza Consortium[181]) and Scotland (Genetics of Severe Influenza in Scotland consortium) during the 2009 H1N1 pandemic. Sixty-one previously healthy adults requiring invasive ventilation due to influenza, representing an extreme susceptibility phenotype, were compared to 1092 population-matched controls. An intronic variant in the *CD97* gene (rs2302092 allele G) was one of the most significantly over-represented variants in susceptible individuals, with a minor allele frequency of 20.5% in cases compared to 5.6% in controls (odds ratio 18.4, $p = 4 \times 10^{-14}$). As well as being one of the highest ranked variants (rank 5 by p-value), this was among the most plausible given the interaction between CD97 and CD55.

3.1.2 CD97 structure and ligand binding

CD97 (also known as ADGRE5) is an adhesion G-protein-coupled receptor that consists of a seven-span transmembrane domain and an extra-cellular domain containing a 'stalk' region and a variable number of epidermal growth factor (EGF)-like domains.[182] The stalk contains an autoproteolysis site (the G-protein-coupled receptor autoproteolysis-inducing domain, or GAIN domain), and in the natural state the transmembrane and extracellular domains are linked by a non-covalent attachment: proteolysis at this site seems to be required for full functionality of the protein.[183] Sequence homology between mice and humans is high (75%) in the transmembrane and cytoplasmic regions and 62% in the EGF-like extracellular regions, but substantially lower in the 'stalk' region (46%).[184] At least three isoforms are recognised in both humans and mice, based on alternative splicing of the EGF-like repeats (Figure 3.1). The smallest isoform, CD97-

(EGF1,2,5) predominates in human leukocytes, and relative proportions of the different isoforms are not altered substantially on immune stimulation.[185] The extracellular domain can be shed, yielding soluble CD97 (sCD97), which could have signalling functions of its own including activation of platelets.[182,186] This receptor is closely related to F4/80 (ADGRE1), a similar adhesion G-protein-coupled receptor that is widely used as a pan-macrophage marker in mice, but for which no ligand has as yet been identified.[187]

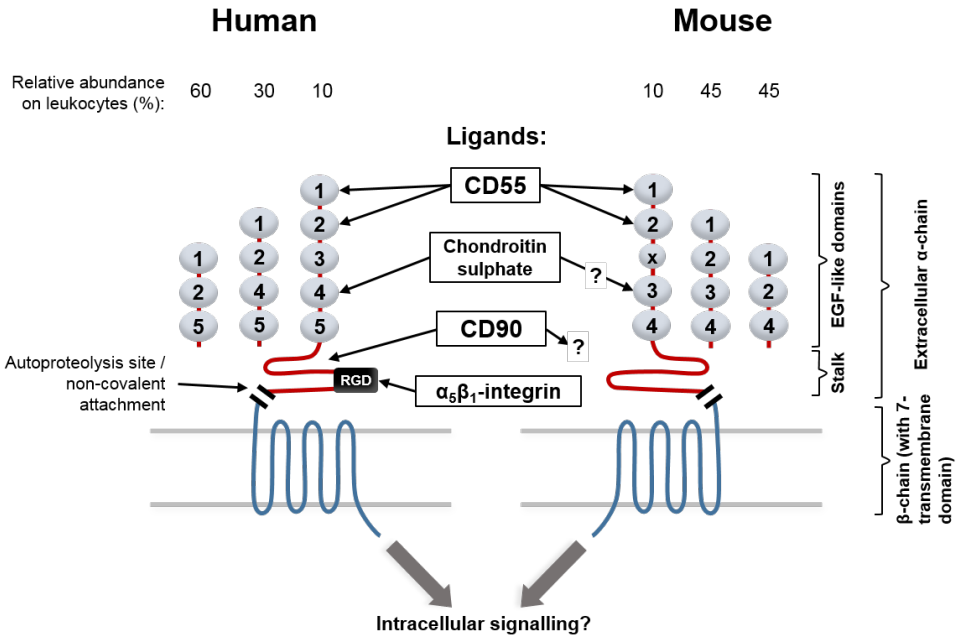

Figure 3.1 – CD97 isoforms in human and mouse

A number of ligands for CD97 have been characterised in humans, most notably CD55. Ligand affinity varies between isoforms. CD55, which binds to the first two EGF-like domains from the N-terminus that are common to all isoforms, binds with greatest affinity to the more common shorter isoforms. Conversely, the extracellular matrix component chondroitin sulphate, which binds the fourth EGF-like repeat in humans, only binds to the longest isoform. CD90 (Thy1) and $\alpha 5\beta 1$-integrin have also been recognised as interacting partners for human CD97, implicated in leukocyte attachment to activated endothelium. Ligand binding is less well characterised in the mouse: the interaction with CD55 has been confirmed, but other ligands have not been fully investigated, and the mouse lacks the Arg-Gly-Asp (RGD) tripeptide in the stalk region thought to be

responsible for integrin binding.[180,184,188–193]

3.1.3 CD97 tissue expression

CD97 is highly expressed in myeloid and lymphoid cells, but is also detectable at lower levels in a wide variety of other cell types, including keratinocytes, fibroblasts, adipocytes, intestinal epithelial cells, smooth muscle and many tumour cell lines.[194–196] In the lung, expression has been demonstrated in endothelial and epithelial cells by immunohistochemistry, but distribution in pulmonary inflammatory cells has not been specifically investigated.[197]

A number of factors are known to modulate CD97 expression, either at transcriptional or protein level. Expression is generally increased during lymphocyte activation, but not after interferon-γ stimulation of monocytes or lipopolysaccharide-mediated activation of endothelial cells.[185,194,198] In a retinal pigment epithelium cell line (ARPE-19), CD97 expression is stimulated by interferon-γ and TNFα, but inhibited by tumour growth factor beta.[199] Vitamin D stimulates expression, and the tumour-suppressor microRNA miR-126 inhibits production at a post-transcriptional level.[200,201] Few details of the molecular basis of transcriptional regulation are known in haematopoietic cells, although binding of the transcription factors Sp1 and Sp3 to a GC-rich region of the promoter has been shown to be a key event in driving expression in HEK293 cells and smooth muscle cell lines.[202] Binding to CD55 leads to shedding of the extracellular domain via a process which seems to depend on shear stress,[203] with subsequent internalisation mediated by phosphorylation by G-protein-coupled receptor kinase-6 and β-arrestin binding to the cytoplasmic domain.[204] The converse effect of CD97-CD55 binding on CD55 expression has not been investigated.

3.1.4 Influence of CD97 on inflammatory responses

The role of CD97 in neutrophil responses has received considerable attention, but prevailing theories as to its function have changed over recent years as transgenic models have become available. Due to its apparent adhesive properties, initial work focused on a potential role in neutrophil migration. *In vitro* assays have suggested that the interaction between CD97 and Thy1 could be

involved in adhesion of leukocytes to activated endothelium in humans, but due to the different expression pattern of Thy1 this may not be important in mice.[193] Studies with monoclonal antibodies against CD97 in mice suggested that it was indeed important in this process, as treatment with neutralising antibodies resulted in reduced granulocyte influx in dioctyl sodium sulphate-induced colitis, *Streptococcus pneumoniae*-induced pneumonia and collagen-induced arthritis, and blocked IL-8-mediated neutrophil mobilisation.[205–207]

Results in CD97-null mice, however, have painted a different picture. Wang *et al.* showed that *Cd97-/-* mice had enhanced circulating and hepatic granulocyte responses, associated with increased resistance to *Listeria* challenge.[208] There was no difference in *in vitro* chemotaxis or in the migratory capacity of *Cd97-/-* and wild type neutrophils in response to intraperitoneal inflammatory stimuli including chemokine (C-X-C motif) ligand 1, granulocyte colony stimulating factor (G-CSF) or *Listeria* in chimeric animals. Similarly, Veninga *et al.* found that while neutrophil migration in thioglycollate-induced peritonitis was inhibited by anti-CD97 antibodies, *Cd97-/-* mice had no impairment either in this model or in *in vivo* neutrophil transmigration in response to leukotriene-B4 or IL-1β.[197] The same group later showed that antibody binding induced Fc-receptor-dependent granulocyte depletion during acute inflammation: the findings of the earlier antibody studies are therefore unlikely to reflect true direct effects of CD97.[209] The contrasting effects of neutralising antibodies and mouse knockout models (for CD97 or CD55) are shown in Table 3.1.

There is still however some evidence that CD97 influences neutrophil homeostasis, independent of antibody-mediated effects. Baseline circulating granulocytosis has been observed in at least a proportion of *Cd97-/-* mice by two authors, using two different knockout strains.[208,210] This is likely to involve the interaction between CD97 and CD55, independent of the effects of CD55 on complement: similar granulocytosis is observed in *Cd55-/-* mice, a difference that was maintained in the absence of C3a and C5a, and that was not increased further in *Cd55-/- Cd97-/-* double-knockout mice.[210] The mechanism of this effect is not well defined: there are no differences in neutrophil margination or circulating morphology (assessed for *Cd55-/-* only), survival and apoptosis *in vivo* or *ex vivo*, or in relative migration from bone marrow. Current evidence suggests that enhanced granulopoiesis in response to G-CSF is most likely responsible: bone marrow of *Cd97-/-* and *Cd55-/-* mice contains increased pro-

portions of cells in the S or G2/M phases of the cell cycle, indicating enhanced proliferation. Baseline and stimulated circulating concentrations of G-CSF and stromal-derived factor-1 (a chemokine involved in haematopoietic cell homing to the bone marrow) are unaffected, but repeated administration of G-CSF results in enhanced granulocytosis in blood and to a lesser extent in bone marrow in *Cd97-/-* mice.[208,210] Modulation of the circulating or pulmonary neutrophil response could affect the degree of tissue injury, and hence disease severity, in influenza, and both number of neutrophils in the lower respiratory tract and markers of circulating neutrophil activation correlate with disease severity in moderate to severe influenza in humans.[129,130] Neutrophils also appear to have protective effects, however, and neutrophil depletion exacerbates disease with virulent strains of IAV such as A/Puerto Rico/8/1934(H1N1) (A/PR/8/34) but not avirulent strains in mouse models.[128] It is therefore difficult to predict the direction of effect of dysregulated neutrophil homeostasis on disease severity.

CD97 has also been implicated in a variety of other immune-related processes. CD97 on antigen-presenting cells can costimulate T cells via CD55, inducing proliferation, enhanced secretion of interferon-γ, GM-CSF and IL-10, and type 1 regulatory T-cell differentiation. These effects are independent of complement: CD97 binds to a different region of the CD55 molecule, and does not interfere with its complement regulatory function.[148,211–213] In human macrophages, CD97 has been shown to negatively regulate NF-κB signalling in response to LPS, via the transcription factor PPAR-γ.[214] Effects on cytokine signalling, T-cell responses and granulocyte homeostasis could all plausibly influence the host response to infection with IAV.

Table 3.1 – Effects of anti-CD97 monoclonal antibodies (mAbs) or gene knockout on *in vivo* murine models of inflammation.

Model	anti-CD97 mAbs	*Cd97-/-*	*Cd55-/-*	Refs.
Listeria monocytogenes infection	Not studied	↑ survival, ↓ bacterial burden, ↑ circulating / hepatic granulocytosis	Not studied	[208]
S. pneumoniae -induced pneumonia	↑ clinical signs, ↑ bacterial load.	No difference in bacterial load in lung, blood or spleen. No effect on neutrophil accumulation in lung at 24 and 44 hours.	↑ survival. ↓ positive blood cultures at 24 and 48h, ↓ bacterial load. No difference in lung histopathology, myeloperoxidase activity or cytokine concentrations (TNF, IL1b, KC, MIP-2).	[197,205,210]
Thioglycollate-induced sterile peritonitis	↓ granulocyte migration to peritoneum.	No difference in peritoneal granulocyte counts or monocyte/macrophage accumulation.	No differences detected.	[197]
LPS-induced acute lung injury (1 mg/kg intraperitoneal)	↓ Neutrophils in lung homogenates (from 16% to approx. 5% at 24h post LPS). ↑ circulating cytokines. No effect on cytokines or neutrophils in *ex vivo* model.	Not studied	Not studied	[209]

Continued on next page

Table 3.1 – *Continued from previous page*

Model	**anti-CD97 mAbs**	***Cd97-/-***	***Cd55-/-***	**Refs.**
Dioctyl sodium sulphate-induced colitis	↓ granulocyte migration, ↑ bacterial outgrowth in lungs, ↓ survival	Not studied.	Not studied	[205]
Collagen-induced arthritis	↓ arthritis scores and bone destruction.	Delayed development of arthritis, ↓ clinical scores.	CD55 – delayed development of arthritis, ↓ clinical scores. Trend towards ↓ erosion / leukocyte infiltration, ↓ IgG2a.	[206,215]
Experimental autoimmune encephalomyelitis	No effect	Not studied	Not studied	[216]
Oxalozone-induced delayed-type hypersensitivity	No effect on granulocyte migration as assessed by ear thickness.	Not studied	Not studied	[216]
Tetanus toxoid delayed-type hypersensitivity	No effect	Not studied	Not studied	[216]
IL-8 or CXCL1-induced mobilization from bone marrow	Completely blocked, ↓ neutrophil compartment, no effect on response to G-CSF	No effect	Not studied	[207,208]

3.1.5 Roles of CD97 outside the immune system

CD97 has been most extensively studied in the context of cancer biology. CD97 is highly expressed in many tumour cell lines, and expression of both CD97 and its ligand CD55 have been associated with invasive and metastatic tumor behaviour and reduced overall patient survival in a number of cancers, including pancreatic cancer, cervical squamous cell carcinoma, hepatocellular carcinoma and acute myeloid leukaemia.[204,217–219] This association could be mediated by effects on tumour cell adhesion and migration. *In vitro* and *in vivo* studies have demonstrated an association between CD97 expression and expression of matrix metalloproteinases 2 and 9, which could assist invasion and metastasis via breakdown of the extracellular matrix. Results from experimental models are not, however, completely consistent: in models of gastric cancer, hepatocellular carcinoma and glioblastoma, CD97 induced an invasive phenotype, while in a fibrosarcoma model, CD97 overexpression inhibited migration by suppressing matrix metalloproteinase activity.[204,220–222] Aggressive behaviour could otherwise arise from CD97-mediated stimulation of angiogenesis, or inhibition of apoptosis in tumour cells.[192,223–225]

CD97 could also play a role in maintaining normal tissue integrity. Although *Cd97-/-* mice have no overt phenotype, subtle changes have been reported in a number of tissues. In the colonic epithelium, lateral adherens junctions (but not tight junctions) are weakened via modulation of catenin expression, and conversely overexpression attenuates experimental colitis but leads to intestinal enlargement.[196,226] Altered sarcoplasmic reticulum structure is seen in the skeletal muscle of knockout mice, but this seems to have no functional consequence.[227] CD97 deficiency has mixed effects on osteoclast activity under different conditions.[228] Effects on lung tissue ultrastructure have not been reported.

3.1.6 Signalling function of CD97

The cytoplasmic domain of CD97 contains a G-protein-interacting region. Evidence of signalling function for a long time remained elusive, and attempts to demonstrate signalling through classical G-protein-associated pathways were

initially largely unsuccessful.[203,229] Recently, however, a variety of direct and indirect observations in studies of cancer cells have shed light on the likely pathways involved, and have allowed us to assemble at least a partial picture of CD97 signalling function. Constitutively active signalling via $G_{\alpha 12/13}$ and Rho-associated kinases was first demonstrated in human prostate cancer cell lines, in part mediated by heterodimerisation with lysophosphatidic acid receptor 1.[230] Further support of this pathway has come from demonstration of direct association of the cytoplasmic domain to G-protein alpha subunits $G_{\alpha 12}$, $G_{\alpha 13}$, $G_{\alpha 14}$ and inhibitory $G_{\alpha z}$, but not to other common subunits such as $G_{\alpha s}$, in a chimeric yeast reporter assay.[231] Transcriptomic analyses and measurements of phosphorylated proteins in key pathways, either by Western blot or phosphorylated protein array, have suggested that in addition to the Rho-associated kinase pathway, CD97 activates downstream pathways including mitogen-activate protein kinase (MAPK) pathways and the phosphoinositol-3-kinase (PI3K) / protein kinase B (Akt) / mammalian target of rapamycin (mTOR) pathway.[224,232] A single study has reported signalling via the JAK2-STAT3 pathway, which could be activated by $G_{\alpha q}$ family subunits such as $G_{\alpha 14}$ and is implicated in immune functions such as dendritic cell activation as well as in cancer biology; this observation has yet to be replicated.[233,234]

The role of ligand binding on receptor function is still not clear. The receptor seems to have a low level of constitutive activity, although this is increased in a truncated protein lacking the extracellular domain.[230,231] A functional autoproteolysis domain seems to be required for effects related to N-cadherin expression and apoptosis.[183,223] CD55 binding can lead to extracellular domain shedding and receptor internalisation under appropriate conditions[203,204], and although one study failed to elicit CD97 activation (as measured by inhibition of cyclic AMP accumulation) with soluble CD55, other studies have found that CD55, with or without a secondary stimulus such as LPS, is necessary for CD97-mediated effects such as modulation of matrix metalloproteinase expression.[204,234] Together, these data suggest a model where ligand-induced shedding of the extracellular domain promotes G_{α} subunit activation by the cytoplasmic domain, triggering downstream signalling events, with termination of the signal by receptor internalisation (Figure 3.2). It is possible that CD97 functions are not solely mediated by intracellular signalling, and distant effects of soluble CD97 or physical effects on cell adhesion could also contribute.

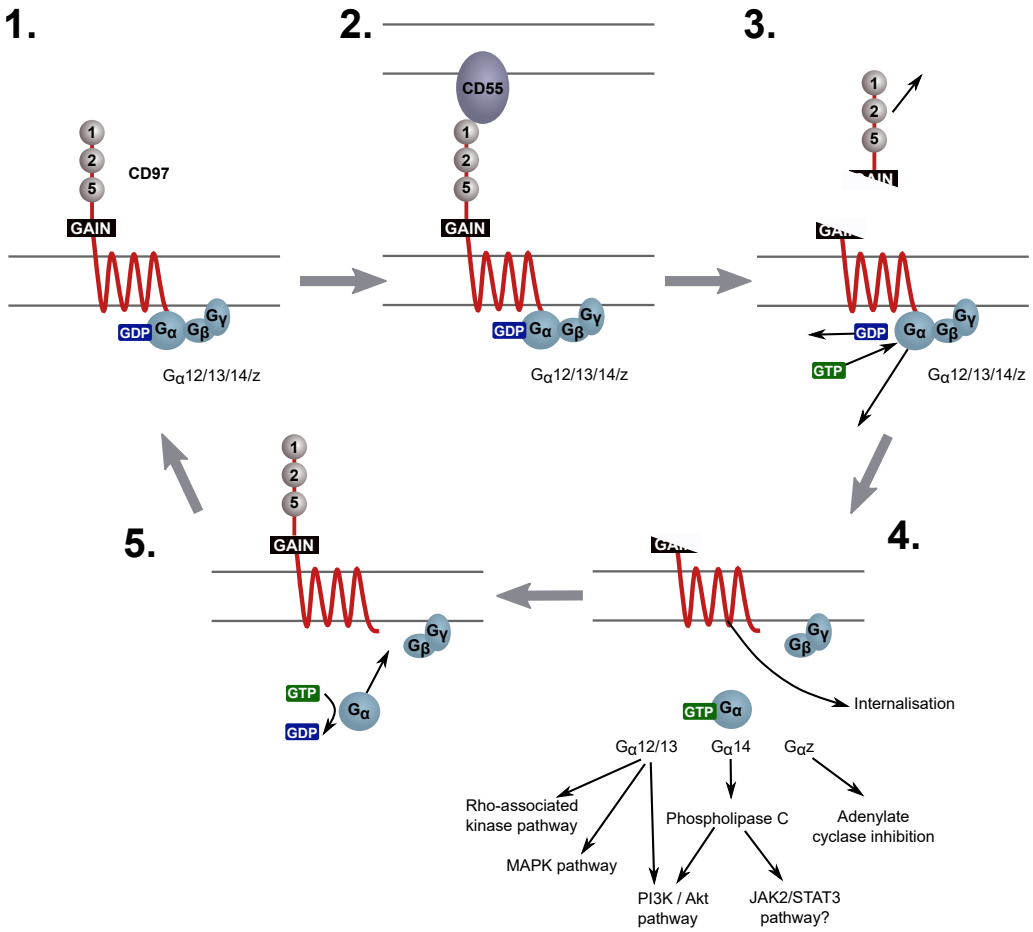

Figure 3.2 – Proposed model of CD97 intracellular signalling. 1. The G_α subunit, as part of a heterotrimeric G-protein, associates with the CD97 cytoplasmic domain and binds guanosine diphosphate (GDP). 2. CD55 binds to the first 2 EGF-like domains. 3. This leads to loss of the extracellular domain via cleavage at the autoproteolysis site, leading to a conformational change in the cytoplasmic domain. This activates the G_α subunit, which exchanges GDP for GTP and dissociates from the heterotrimer. 4. The G_α subunit triggers downstream signalling events, depending on the specific subunit involved. The remaining cytoplasmic domain is internalised. 5. The G_α subunit hydrolyses GTP to GDP, and re-associates with the membrane-anchored subunits, terminating the signal.

3.1.7 Aims and objectives of this chapter

The genetic association between *CD97* and severe influenza warrants consideration of the potential role of this receptor in influenza pathophysiology. Even without this, there is a sufficient basis to justify *a priori* consideration of this receptor as a possible modulator of influenza severity, based on its functional interaction with CD55, a receptor previously implicated in severe influenza, its expression in immune cells and its involvement in multiple biological processes relevant to the immune response. As will be discussed, the disease-associated locus itself is challenging to investigate in humans due to high homology with other adhesion G-protein-coupled receptor genes. Therefore, rather than focusing on the variant itself, the investigations described here aimed primarily to address whether the gene as a whole plays a role in the host response to infection with influenza A virus. Given the wide range of immunological and non-immunological processes in which CD97 has been implicated, an *in vivo* system was the most appropriate approach to model the role of the receptor in the integrated pathophysiological response to a viral challenge.

The primary hypothesis was that CD97 is required for an effective immune response to infection with influenza A virus, and that deficiency will lead to immune dysregulation with either an inefficient or a harmful excessive response.

Specific objectives of this chapter were as follows:

- To characterise the genomic features of severe influenza-associated variant rs2302092, and the likelihood that this variant could be, or could be a proxy for, a functional regulatory variant.
- To characterise the expression distribution of CD97, and its major ligand CD55, in mouse pulmonary and circulating immune cells, as an indicator of the spectrum of immunological processes in which the receptor could be involved.
- To evaluate the impact of CD97 deficiency on influenza severity in a mouse model.
- To evaluate the perturbations in the immune response associated with CD97 deficiency in *in vivo* IAV challenge and in *in vitro* models.

- To assess whether previously reported immune perturbations with CD97 deficiency, such as altered neutrophil homeostasis and increased cytokine production by macrophages, can be replicated in this mouse model.

3.2 Results

3.2.1 Population genetics and functional characterisation of *CD97* variant rs2302092

The genetic association with rs2302092 does not necessarily indicate that this intronic variant is causal or functional. Whole exome sequencing, as used for the association study which implicated rs2302092 in disease severity, allows robust identification of variants within protein-coding regions, but provides limited coverage of intronic and intergenic sequences. It is thus most likely that, assuming the association is not a false positive, rs2302092 is serving as a marker for an unsequenced regulatory variant with which it is in linkage disequilibrium, although it is also possible that it could be a proxy for an unidentified coding variant for which genotyping did not pass quality control. To investigate whether this variant is likely to be, or to be in linkage disequilibrium with, a functional variant that could be a plausible mediator of an effect on influenza severity, publicly available genomic data sources were used to establish the population genetics and presence of regulatory elements at this locus.

3.2.1.1 Population genetics

Data from 2,504 individuals in the 1000 Genomes Project[160] show a minor (G) allele frequency of 12% in individuals from England and Scotland (based on a sample of 91 individuals). The frequency varies in global populations from 2% in Peru to 28% in an African Caribbean population in Barbados, with generally lower frequencies in East Asian and American populations and higher in African populations. In this dataset there is a notable departure from the Hardy-Weinberg equilibrium (χ^2 = 48.6, p = 3 x 10^{-12}) with a reduced frequency of minor homozygotes across populations, which could be indicative of a number of

phenomena including genotyping error, selection or population stratification.[235] In this instance, the most likely cause is mismapping due to almost complete alignment with an intronic region in the *EMR2* gene (also known as *ADGRE2*): mismapping of reads from *EMR2*, with a major C allele at analogous SNP rs35975843, could lead to erroneous heterozygote genotype calls (Figure 3.3). No departure from Hardy-Weinberg equilibrium was present for rs35975843. Difficulties with mapping or with array probe specificity caused by such a high degree of homology could affect many genotyping platforms and will hinder accurate analysis of variant effects in this region; however effects on between-group comparisons (as for a GWAS study) should be limited to a modest bias towards null effect, as the same errors are expected in each group, provided the same platform is used for cases and controls.

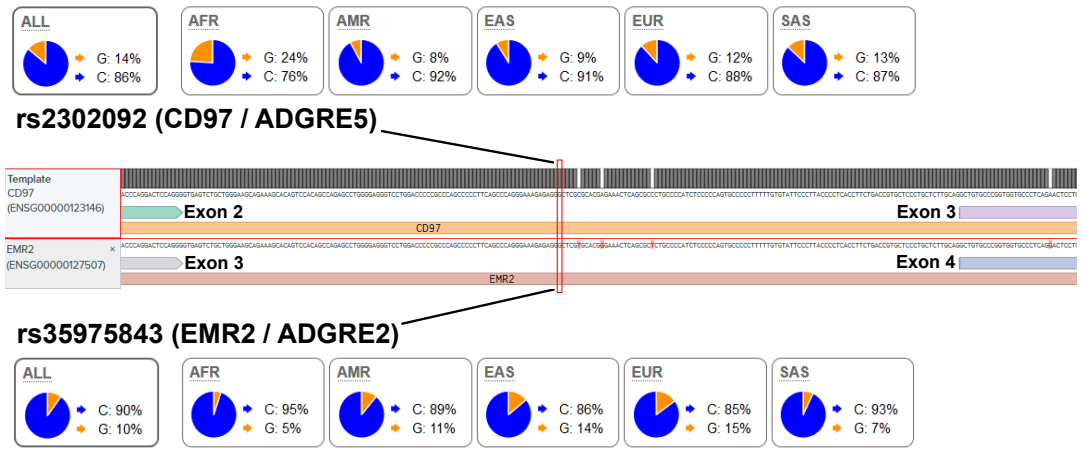

Figure 3.3 – Population genetics of *CD97* variant rs2302092 and homology with the EMR2 gene. Intron 2 of the human *CD97* gene aligns almost perfectly with intron 3 of *EMR2*, and the variant rs35975843 is analagous to rs2302092. The pie charts indicate relative allele frequencies in different geographic populations from the 1000 Genomes project. AFR: African (*n*=661), AMR: American (*n*=347), EAS: East Asian (*n*=504), EUR: European (*n*=503), SAS: South Asian (*n*=489).

Linkage disequilibrium (LD) data from 1000 Genomes, in the same British subpopulation as above, showed only four other variants in moderate to weak linkage disequilibrium with rs2302092, at a threshold of $r^2 > 0.05$ (Table 3.2), all in introns of the CD97 gene. It is however highly likely that over-estimation of major allele frequencies due to the mapping difficulties described above would have hindered accurate determination of LD, and thus it is likely that alleles at other unidentified loci are correlated with this variant.

Variant	Location	Type	MAF	D'	r^2
rs2302093	Intron 2	SNP	0.05	0.41	0.058
rs555717019	Intron 3, in enhancer	2nt deletion	0.12	0.43	0.16
rs10404878	Intron 4	SNP	0.29	1.00	0.061
rs10410071	Intron 4-6 (transcript-dependent)	SNP	0.46	0.536	0.052

Table 3.2 – Variants in linkage disequilibrium with rs2302092. Linkage disequilibrium and minor allele frequency (MAF) data are from 91 English and Scottish individuals in 1000 Genomes.

3.2.1.2 Evidence of local regulatory features

To determine the likelihood that rs2302092 or its linked variants could influence regulation of gene expression, evidence for regulatory variants for *CD97* was first queried in the Genotype-Tissue Expression database (GTex v8)[64], a dataset derived from whole genome sequencing and bulk mRNA sequencing across a range of tissues in 838 subjects. This database was used to identify all cis-acting expression quantitative trait loci (eQTLs) and splicing quantitative trait loci (sQTLs) for *CD97* in humans, and conversely to identify whether rs2302092, or any of the variants with which it is in LD, modulates expression of any other gene.

The majority of eQTLs were located, as expected, within the annotated promoter region upstream of exon 2. However, although none were identified in the intron containing rs2302092, a number of eQTLs and sQTLs were identified a short distance downstream, in the third and fourth introns (Figure 3.4). rs2302092 is an eQTL in adipose tissue for *ZNF333*, a gene also located on chromosome 19, approximately 300 kilobases away from *CD97* and adjacent to *EMR2* (Figure 3.5A). There was no significant evidence of an effect on *CD97* expression. However, the apparent minor allele frequency in this population was very low (2.7%) compared to other populations, which could reflect similar mapping difficulties to those described above, and limits the power to detect genotype-associated regulatory effects. rs2302093, located within the same intron as rs2302092, was a significant eQTL for *EMR2*, with increased expression in the presence of

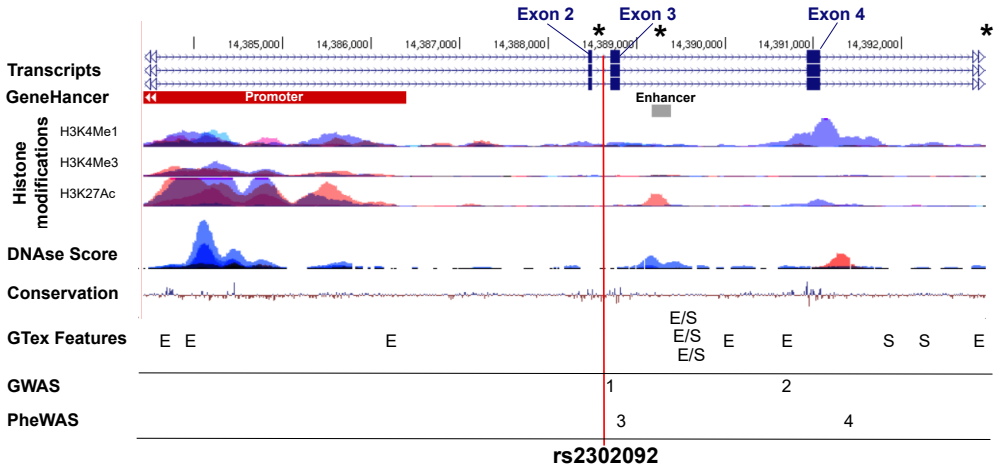

Figure 3.4 – Regulatory features in the vicinity of rs2302092. Genome browser view of the human *CD97* gene in the vicinity of rs2302092 (created with UCSC genome browser, https://genome.ucsc.edu). Transcripts are from RefSeq, with GeneHancer annotated regulatory elements. Locations of variants in linkage disequilibrium with rs2302092 are shown by asterisks (*). Histone modifications shown are from ENCODE CHIP-Seq data from 4 relevant cell lines: K562 (myeloid), GM12878 (B-lymphoblastoid), HUVEC (endothelial) and NHLF (lung fibroblast). The DNase score indicates DNase I hypersensitivity (from ENCODE) in 10 relevant cell types including the above 4 plus A549 cells (epithelial), lymphoid cells and monocytes. The Conservation track indicates PhyloP conservation scores[166] based on 100 vertebrate species: positive values (blue) indicate conserved bases, and negative (red) indicate rapidly evolving sites. All significant GTex eQTLs (E) and sQTLs (S) within this window are shown (false discovery rate <0.05). The GWAS and PheWAS tracks indicate variants significantly associated with phenotypes in the NHGRI-EBI GWAS Catalogue[170] or in UK Biobank (via GeneAtlas[171]) respectively - 1: rs8105300, associated with white blood cell count; 2: rs7260110, associated with blood EMR2 concentration; 3: rs139113505, associated with white blood cell, lymphocyte and monocyte count; 4: rs149215083, associated with reticulocyte count.

the minor allele (Figure 3.5B). Again, although trans-effects are possible given the proximity of the two genes on chromosome 19, this could also be a result of mismapping, and could in fact reflect effects of analogous variants within the *EMR2* gene. Linked variants rs555717019 and rs10404878, and the analogous *EMR2* variant rs35975843, had no available data in GTex. Linked variant rs10410071 was a significant sQTL in whole blood, with increased excision of the intron between exons 5 and 7 (and hence preferential production of shorter isoforms) in the presence of the minor allele (Figure 3.5C).

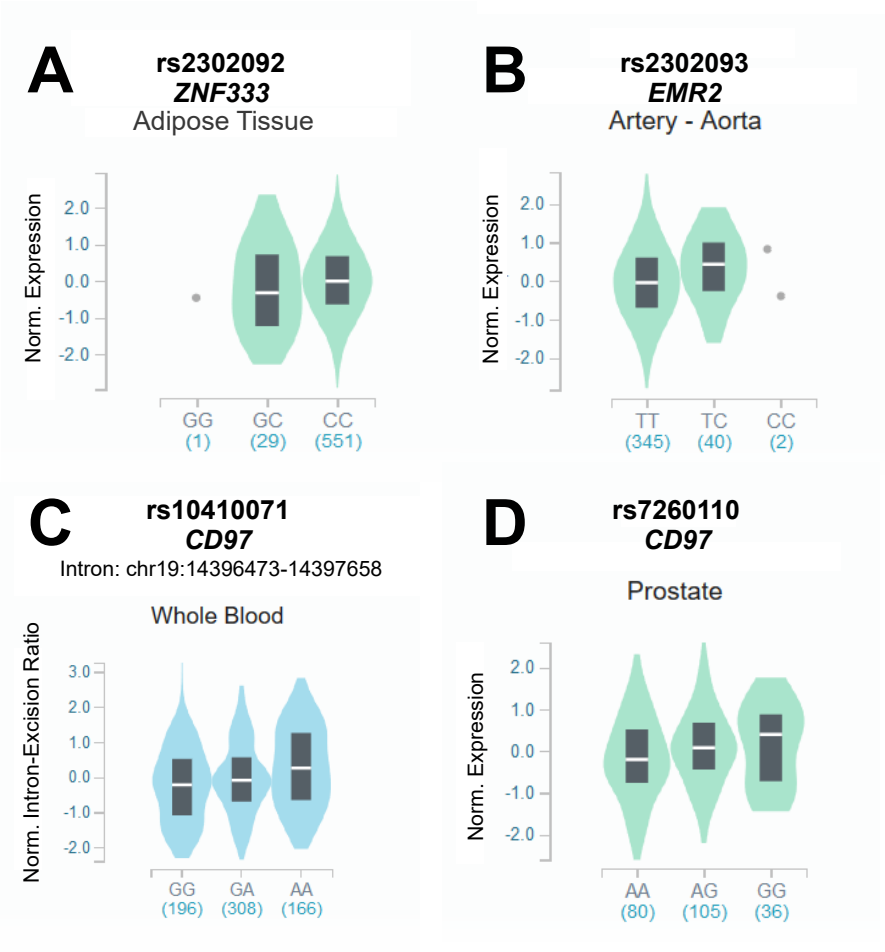

Figure 3.5 – eQTLs in the vicinity of rs2302092. Violin plots showing GTex evidence of regulatory function for selected variants. A: rs2302092 as an eQTL for ZNF333; B: rs2302093 as en eQTL for EMR2; C: rs10410071 as an sQTL for CD97; D: rs7260110 as an eQTL for CD97. Numbers in blue indicate *n*.

Next, known regulatory elements at this locus were identified in the GeneHancer database[165], which integrates data from a variety of sources including the Functional Annotation of the Mammalian Genome (FANTOM) project[50], and for chromatin features suggestive of the presence of regulatory elements using the ENCODE database.[162] An enhancer was identified just upstream of exon 3, associated with a tissue-specific area of DNase I hypersensitivity, indicating the presence of accessible chromatin.[56] A tissue-specific H3K27Ac histone acetylation mark was also present at this site, which is thought to enhance expression via inhibition of suppressive histone methylation, and to distinguish active

from inactive enhancers.[54] Variant rs555717019, in linkage disequilibrium with rs2302092, comprises a 2-nucleotide deletion within this enhancer, and so could plausibly disrupt its function. Additional histone marks and DNase hypersensitivity sites were present further downstream towards the early part of intron 4, suggesting the presence of additional regulatory elements, but there were no chromatin features suggesting the presence of regulatory elements at the site of rs2302092 itself within intron 2.

To complement this, the variants above were cross-referenced with transcription factor binding sites in the Factorbook database and using SNP2TBFS, which models the effect of polymorphisms on likely transcription factor affinity based on pattern weight matrices.[163,167] rs2302092 overlapped with a chromatin immunoprecipitation (ChIP-Seq) peak for RNA polymerase II in spleen, and also overlapped a possible binding motif for interferon regulatory factor 1 (IRF1). If this represents a functional binding site, it could be relevant to regulation of *CD97* expression, although the sequence only loosely conforms to the IRF1 consensus sequence, and the variant is unlikely to affect binding.[236] Of the other variants in LD with rs2302092, rs10404878 overlapped with ChIP-Seq peaks for SMAD5 (involved in transforming growth factor beta signalling) and SUPT5H, and rs555717019 (within the enhancer) overlapped a peak for PCBP2, while both overlapped with peaks for RNA polymerase II. While the variants overlapped a variety of transcription factor binding motifs that could plausibly be involved in *CD97* expression regulation in this context, such as for SP1 and SP2 (for rs555717019), position-weight matrix modelling of transcription factor binding specificity predicted only modest effects on binding affinity.[167]

These results indicate that it is unlikely, although not impossible, that rs2302092 is itself a functional variant, and thus is unlikely to itself be the causative variant for the observed genetic association. There multiple sources of evidence discussed above, however, confirm the presence of regulatory elements and regulatory variants in close proximity in adjacent introns, and in some cases in linkage disequilibrium with this variant. Given that linkage between rs2302092 and other local variants is likely to be underestimated, it is plausible that the observed association between rs2302092 and influenza severity is a proxy for an unmeasured association with one of these other regulatory variants. From the currently available data it is not, however, possible to determine the direction of effect on expression or splicing to be expected with the minor allele of this

variant.

3.2.1.3 Supportive evidence for a role of genetic variants of *CD97* in influenza

Previous genetic association studies have not demonstrated a significant association between influenza severity and any variant in the *CD97* gene. The few genome-wide studies (or large sub-genome scale screens) reported to date have all had low sample numbers and thus limited power, or in one case used a genotyping panel that did not include any markers close to *CD97*.[70,71,73]

To establish whether there is any precedent for an association between *CD97* variants and other forms of infectious or inflammatory disease, or relevant immune traits, the National Human Genome Research Institute - European Bioinformatics Institute GWAS catalogue was used to identify all significant associations with *CD97* in published GWAS datasets (including those associations not reported explicitly in the associated manuscripts). Only four significant associations were identified. Variant rs8105300, 67 bases away from rs2302092 within the same intron, has been associated with circulating white blood cell count.[237] Variant rs7260110, identified as an eQTL for *CD97* in GTex, was associated with blood EMR2 concentrations (see Figures 3.4 and 3.5D).[238] Two other variants in the upstream promoter region have been associated with mean platelet volume, which could be plausible given the suggested function of sCD97 in platelet activation[186], and extreme parental longevity.

Given the recent accumulation of data on genetic susceptibility to COVID-19, available datasets, including those from our own laboratory, were queried to determine if there is an association between *CD97* and severe disease after infection with SARS-CoV2. This is an RNA virus, like IAV, and is therefore expected to share some common aspects of the host immune response, but the spectrum of pathologies produced is quite different. No dataset included rs2302092, and there was no evidence of significant significant associations with other variants in the vicinity of *CD97*. There was limited support for a similar association with variant rs2564978 in the *CD55* gene as has been observed for influenza: in a large GWAS comparing patients with severe disease to population controls in a European cohort, the T allele (associated with severe disease for influenza)

was found more frequently in severe cases (odds ratio 1.1, p = 0.03 under an additive model).[9] Meta-analysis of 35 studies of hospitalised cases versus population controls did not, however, find any significant effect.[150]

Three of the *CD97* associations in the GWAS Catalogue were based on data from the UK Biobank. To complement this, regional phenome-wide association was performed using the GeneAtlas database of genotype-phenotype variants in the UK Biobank[171], incorporating 778 traits and 30 million genotyped or imputed variants, in 452,264 individuals of white British ancestry. Considering all variants within 2 kilobases of the *CD97* gene, with a conservative p-value threshold of $< 1 \times 10^{-8}$, nine associations were found, accounted for by three lead variants. The above association with mean platelet volume was confirmed. A variant in exon 4 was associated with three measures of reticulocyte abundance. The association with white cell count was also confirmed ($p = 3 \times 10^{-14}$), albeit with a different lead variant, rs139113505, a missense variant in exon 3 (see Figure 3.4); this variant was also associated with lymphocyte count ($p = 1 \times 10^{-13}$), monocyte count ($p = 3 \times 10^{-10}$), platelet count ($p = 5 \times 10^{-9}$) and platelet distribution width ($p = 5 \times 10^{-18}$).

The UK Biobank data include a non-specific 'influenza and pneumonia' phenotype, but no more specific phenotype related to influenza severity; moreover our lead variant is not included in the database as it was not in the genotyping array used and cannot be accurately imputed. Lack of evidence of association in this dataset, for this specific variant at least, should not therefore be interpreted as evidence of lack of association. There is no evidence of association between *CD97* and other respiratory diseases (such as asthma, pneumonia or chronic obstructive pulmonary disease) or other viral infections (such as measles), although in the latter case, as for influenza, the recorded trait indicates only historic presence or absence of the disease rather than severity.[171]

The association with total white blood cell, lymphocyte and monocyte counts could reflect an effect on leukocyte development, activation or trafficking, which could be a plausible mechanism for modulation of influenza severity. As this has only been documented in a single population to date, with a large sample size but notable demographic biases, further investigation of CD97 function is required to determine if this gene does influence leukocyte populations *in vivo*, whether this effect extends to the lung as well as peripheral blood, and whether

such effects could modulate influenza severity. Any future genetic studies are likely to suffer from similar limitations to the above studies, with regard to difficulties in genotyping the locus and lack of available proxy variants. To dissect the potential role of CD97 further, particularly given the wide range of immune processes and intercellular interactions in which it could be involved, an *in vivo* model will be required.

3.2.2 Distribution of CD97 and CD55 in human and murine immune cells

CD97 has been previously reported to be highly expressed on variety of leukocyte subsets including neutrophils, and expression has also been demonstrated on pneumocytes, bronchial epithelial cells and a subset of endothelial cells in murine lung.[197] To help establish the possible cellular interactions between CD97 and its major cellular ligand CD55 which could be relevant in the host response to influenza, the distribution of the two receptors in immune cells was characterised in more detail, using both flow cytometry and interrogation of existing data sources. I hypothesised that expression across cell types would be comparable between human and mouse, and thus that a murine model would be appropriate to study the role of the gene in the host immune response.

3.2.2.1 Tissue distribution of RNA expression in published data sources

Human and murine immune cell expression of *CD97* and *CD55* were first compared at the transcript level, as measured by microarray, in a human primary cell atlas[169] and a murine tissue atlas (GeneAtlas MOE430[168]), via the BioGPS portal (Figures 3.6 and 3.7).[239] For additional corroboration, these data were then compared to human transcriptomic data generated by cap analysis of gene expression from the FANTOM5 project[50] (Figure 3.8), and mouse immune cell transcriptomic data derived by RNA sequencing from the Immunological Genome (ImmGen) project[240] (Figure 3.9).

In both species, *CD97* is is broadly expressed across immune cell populations, and expressed at relatively low levels in most non-haematopoietic tissues such as respiratory epithelia. Particularly high levels are seen in granulocytes such as

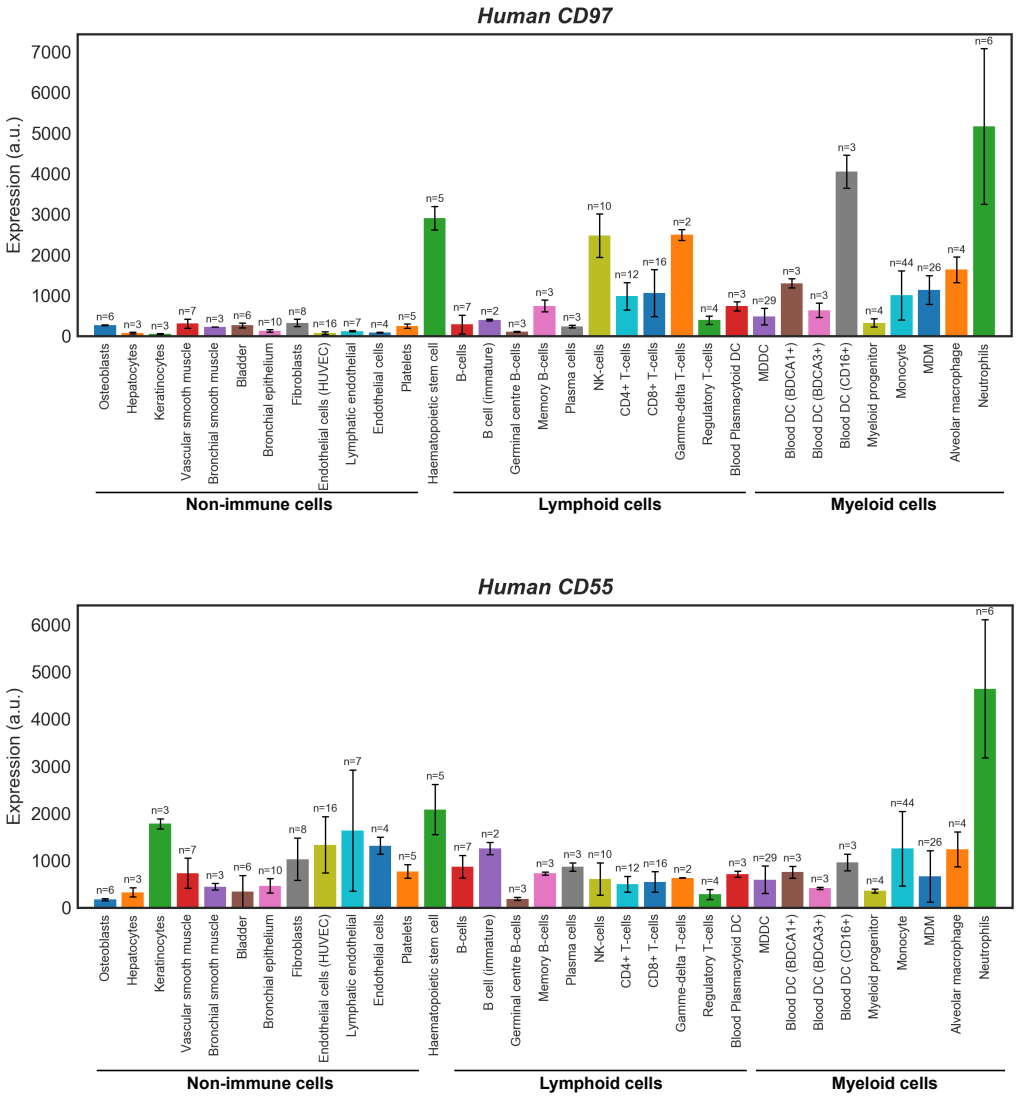

Figure 3.6 – Human *CD97* and *CD55* expression in BioGPS. Bar charts show mean ± standard deviation expression values (arbitrary fluorescence units) for unstimulated cells, from the Human Primary Cell Atlas. Similar cell types (e.g. brain tissues from different areas, fibroblasts from different tissues) have been pooled if appopriate. MDM: monocyte-derived macrophages; DC: dendritic cells; MDDC: monocyte-derived dendritic cells.

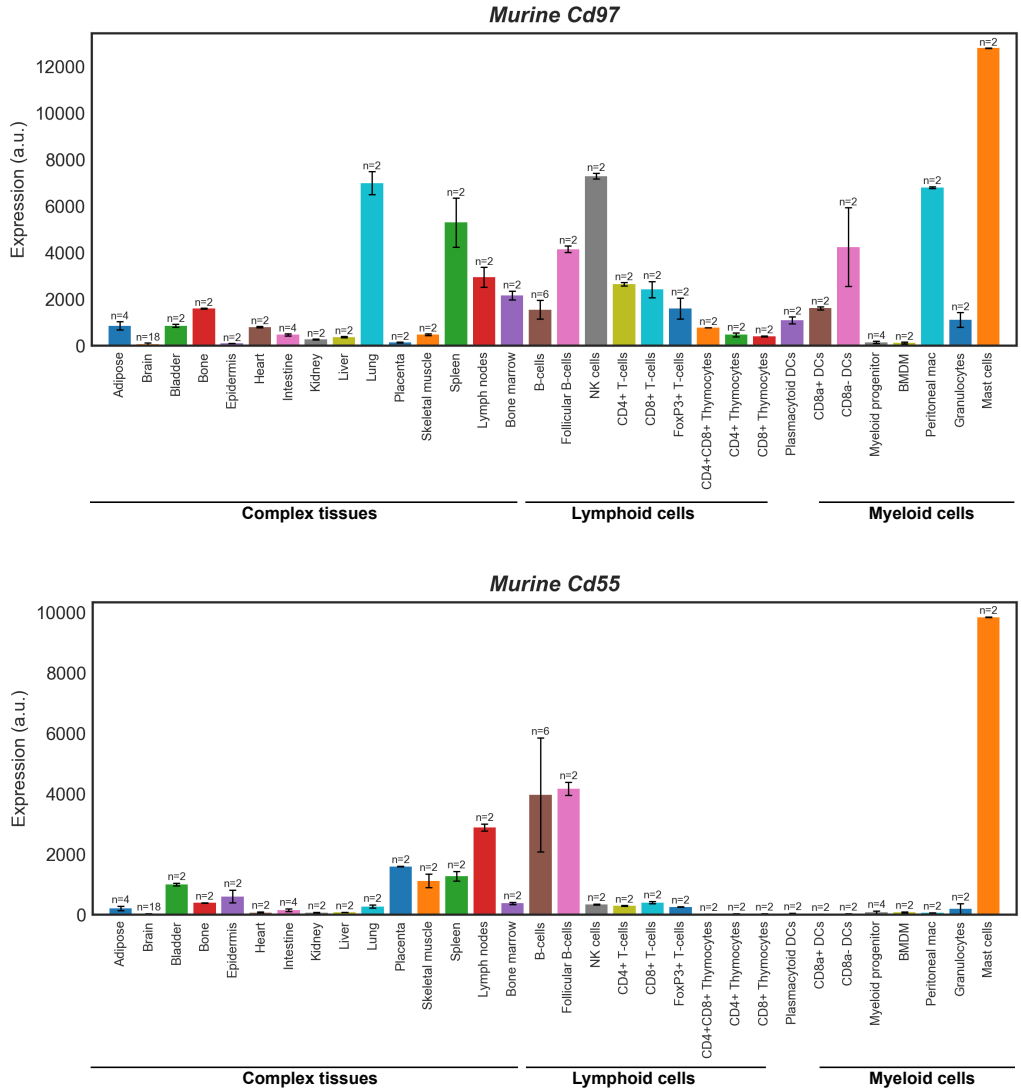

Figure 3.7 – Murine *Cd97* and *Cd55* expression in BioGPS. Bar charts show mean ± standard deviation expression values (arbitrary fluorescence units) for unstimulated cells, from mouse GeneAtlas MOE430. Similar cell and tissue types have been pooled if appropriate. BMDM: bone marrow-derived macrophages; DC: dendritic cells.

neutrophils, eosinophils (although data are lacking in mice) and mast cells, and in natural killer cells. This observation is consistent across datasets and across species. It is also expressed in most monocyte-macrophage and dendritic cell populations, but here there is more heterogeneity. In mice, expression is high in peritoneal macrophages and some dendritic cell subsets, but very low in bone marrow-derived macrophages and absent in alveolar macrophages. This is a notable difference from humans, in which expression in alveolar macrophages is similar to other macrophage populations. In humans, T lymphocytes in general express *CD97* at higher levels than B lymphocytes, with the relative order of T-cell subsets largely consistent between datasets (gamma-delta T cells highest, regulatory T cells lowest). In mice, expression is broadly comparable between lymphocyte populations.

CD55 is broadly expressed across most immune cells and non-immune tissues in humans, with particularly high expression in neutrophils and eosinophils. There is some heterogeneity between datasets, which could reflect differences in methodology or in cell isolation techniques: for example monocyte-derived macrophages and dendritic cells have negligible expression in FANTOM5 (Figure 3.8) but are comparable to other immune cells in the Primary Cell Atlas (Figure 3.6). In mice, *Cd55* expression seems to occur predominantly in lymphatic endothelial cells, B cells and mast cells, with moderate to low expression in other lymphoid cells and granulocytes, but minimal expression in macrophages and dendritic cells.

The above data relate to unstimulated cells. As induced expression may be more relevant during disease associated with inflammatory challenge, *CD97* and *CD55* expression data were extracted from time series experiments in the FANTOM5 database. Although variability between individuals was high, LPS stimulation of human monocyte-derived macrophages resulted in a transient increase in *CD97* expression followed by a return to baseline, while *CD55* expression increased progressively over the time course (Figure 3.10A). Infection of monocyte-derived macrophages with IAV induced more modest expression changes, with a slight increase in *CD55* and a small transient decrease in *CD97* (Figure 3.10A). In mice, fewer relevant datasets were available. Bone marrow-derived macrophages showed little change in expression following interferon-γ stimulation and *Mycobacterium tuberculosis* infection, although this did confirm that BMDMs do dynamically express both genes, which was not clear from the

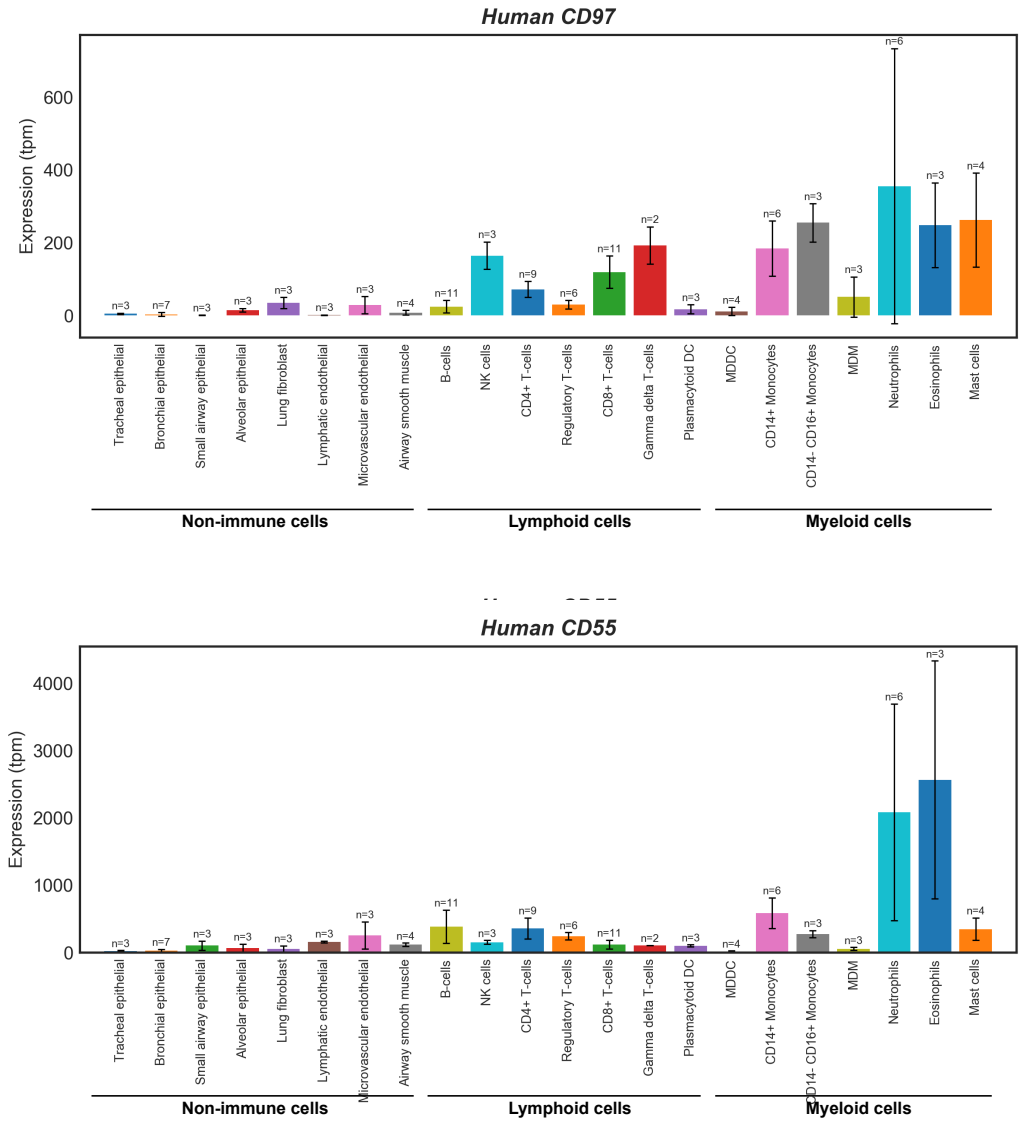

Figure 3.8 – Human *CD97* and *CD55* expression in FANTOM5. Bar charts show mean ± standard deviation expression values (in tags per million, from cap analysis of gene expression) for the primary promoter in unstimulated cells. DC: dendritic cells; MDDC: monocyte-derived dendritic cells; MDM: monocyte-derived macrophages

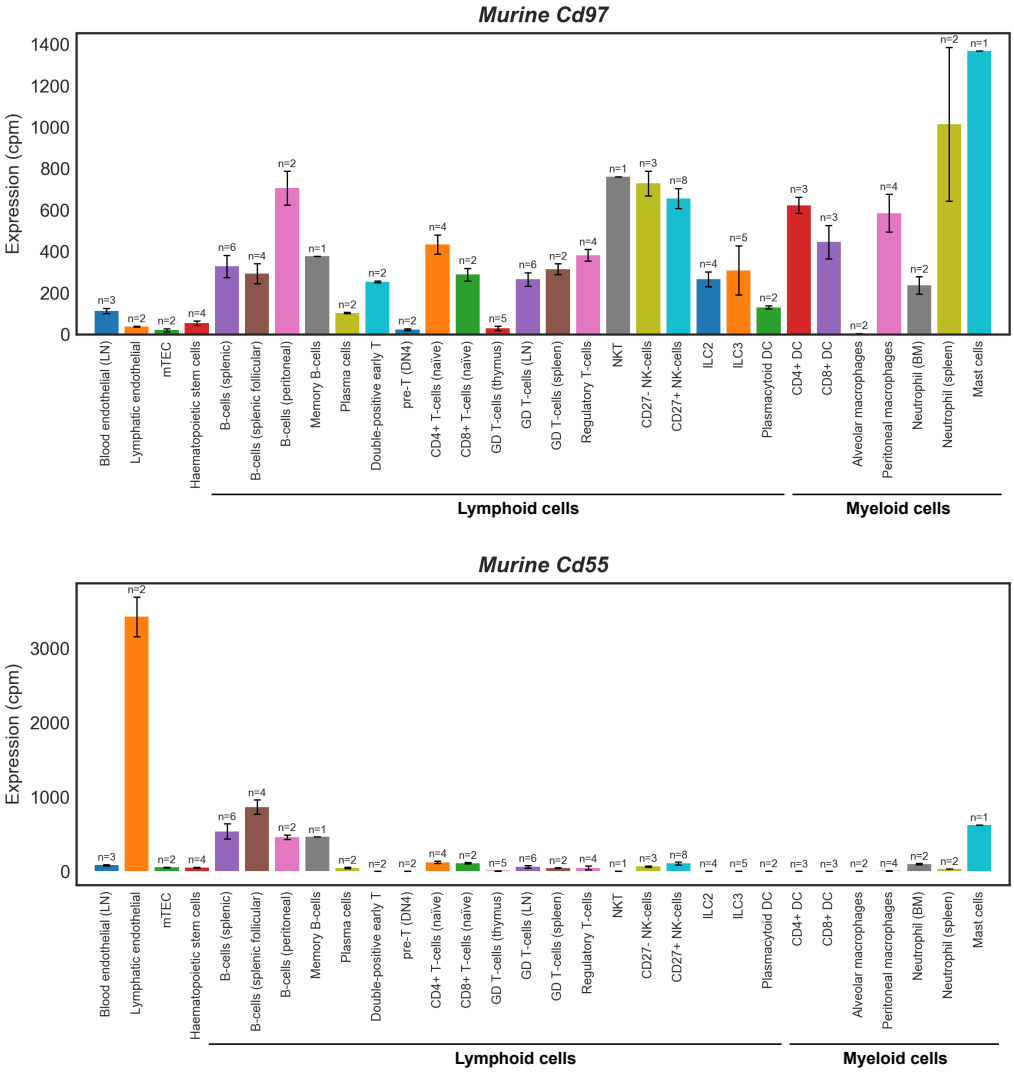

Figure 3.9 – Murine *Cd7* and *Cd55* expression from the Immunological Genome Project. Bar charts show mean ± standard deviation expression values (counts per million) for unstimulated cells. Cells listed were isolated from spleen, bone marrow, peritoneal cavity, lymph node, thymus or lung (alveolar macrophages only). mTEC: medullary thymic epithelial cells; DN4: T-cell precursor, double-negative stage 4; GD T cells: gamma-delta T cells; NKT: natural killer T cells; ILC: innate lymphoid cells; DC: dendritic cells

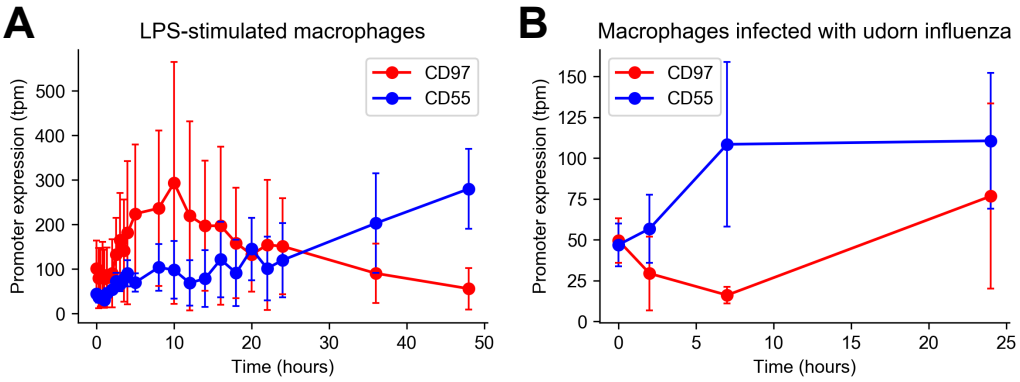

Figure 3.10 – *CD97* and *CD55* expression time courses in human cells in FANTOM5. Plots show mean ± standard deviation expression values (tags per million, summed for all annotated promoters for each gene) for 3-4 biological replicates. A: Monocyte-derived macrophages stimulated with LPS (Salmonella R595 at 100ng/ml). B: Monocyte-derived macrophages infected with H3N2 influenza virus strain A/Udorn/72 at an MOI of 5 pfu/cell.

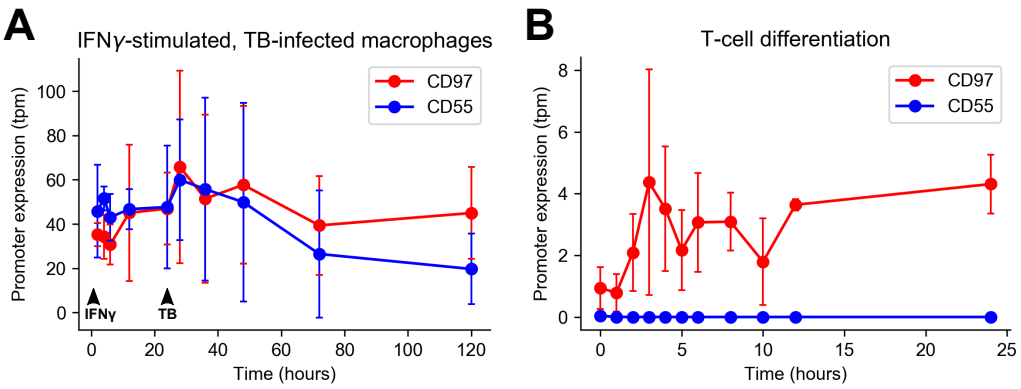

Figure 3.11 – *Cd97* and *Cd55* expression time courses in murine cells in FANTOM5. Plots show mean ± standard deviation expression values (tags per million, summed for all annotated promoters for each gene) for 3-4 biological replicates. A: Bone marrow-derived macrophages, stimulated with 100 U/ml IFNγ, followed by infection with *Mycobacterium tuberculosis* (TB) at 24 hours. B: T-cell differentiation, induced by co-culture with TSt-4/DLL1 stromal cells, of haematopoietic progenitor cells from mice lacking early B-cell factor 1 (EBF1). Complete absence of *Cd55* expression here is not typical for T cells in the FANTOM5 database and could reflect early differentiation state or loss of transcriptional activation by EBF1.

above data in unstimulated cells. Increased *Cd97* expression was also observed in an *in vitro* model of T-cell differentiation.[241]

Although the various datasets above are not completely consistent in the selections of immune cell subsets examined, or in how the cells were defined and isolated, there are clear similarities between the two species with regard to the tissue distribution of these receptors, particularly with regard to expression on neutrophils, natural killer cells and lymphocytes. Although there are some differences which warrant a degree of caution, most notably the lack of expression in murine alveolar macrophages compared to humans and the apparent more restricted expression of *Cd55* in murine antigen-presenting cells, the similarities suggest that CD97 is likely to be involved in similar immune processes, and thus initial investigations in a murine model are justified.

3.2.2.2 Characterisation of CD97 and CD55 expression in murine lung and blood immune cells by flow cytometry

Expression in pulmonary immune cells, and in circulating immune cells trafficked to the lung, is likely to be of prime importance in the host response to infection with influenza A virus. To verify predicted expression patterns in murine immune cells at the protein level, flow cytometry was used to provide a semi-quantitative measurement of CD97 and CD55 expression in major myeloid and lymphoid subsets of lung homogenates and whole blood from 8-10 week-old wild type C57BL/6J mice of both sexes. Cell population definitions are given in table 2.4, and gating strategies are shown in Figures 3.12 and 3.13.

Low fluorescence in cells stained with the isotype control antibody indicated low non-specific binding, although this appeared greater in myeloid cells, especially granulocytes. Non-immune cells in the lung showed a broad range of CD97 expression (Figure 3.12). Strong CD97 expression was seen in most myeloid cell types examined in the lung, including granulocytes, macrophages and dendritic cells (Figure 3.14A). In these cells, fluorescence intensity was significantly ($p < 0.001$) higher than in both isotype and *Cd97-/-* negative controls, and fluorescence in *Cd97-/-* cells was not significantly different from that of wild type cells stained with isotype control antibody, confirming that the protein had been effectively knocked out in *Cd97-/-* mice. Alveolar macrophages were a notable exception, with minimal CD97 expression: wild type and *Cd97-/-* cells were not significantly different, while fluorescence was significantly higher than in isotype controls for both groups. Similarly, there was a significant difference, albeit of

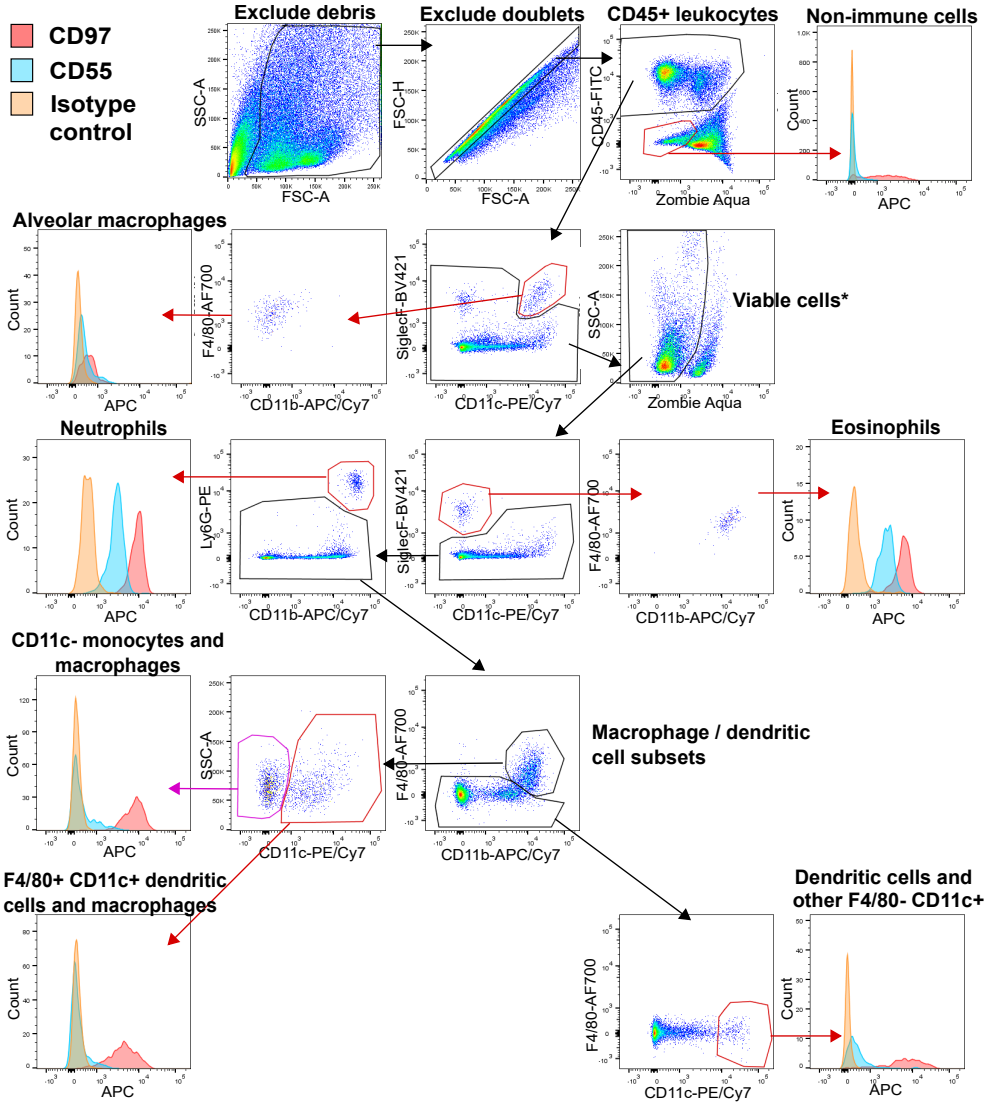

Figure 3.12 – Myeloid cell gating strategy and distribution of CD97 and CD55 in murine lung. The antibody panel for myeloid cell identification was used together with an allophycocyanin (APC)-conjugated antibody for CD97 or CD55 (or isotype control antibody). Zombie Aqua fixable live/dead stain was used to exclude dead cells. The flow cytometry dot plots are pseudocoloured to indicate cell density. Histograms of expression distribution are from a single representative wild type animal. *Alveolar macrophages were identified prior to dead cell exclusion, due to intense autofluorescence overlapping with the viability marker.

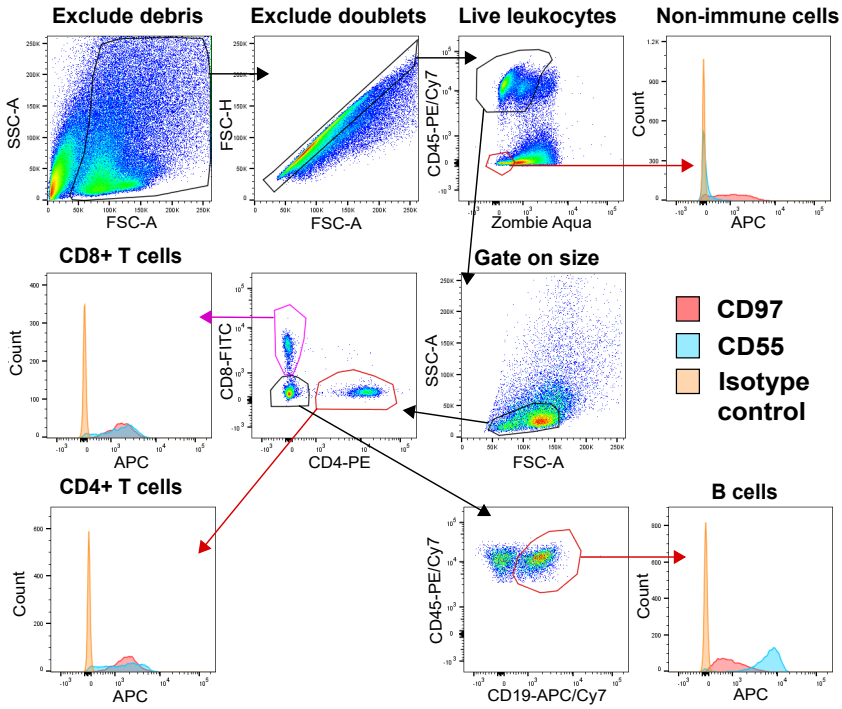

Figure 3.13 – Lymphoid cell gating strategy and distribution of CD97 and CD55 in murine lung. The antibody panel for myeloid cell identification was used together with an allophycocyanin (APC)-conjugated antibody for CD97 or CD55 (or isotype control antibody). Zombie Aqua fixable live/dead stain was used to exclude dead cells. The flow cytometry dot plots are pseudocoloured to indicate cell density. Histograms of expression distribution are from a single representative wild type animal.

extremely low magnitude, between *Cd97-/-* cells and isotype controls for pulmonary lymphocytes (Figure 3.14B). Given the small differences observed and the results in other tissues and cell types of similar lineage, it is unlikely that the increased fluorescence in the *Cd97-/-* cells represents failure of the knockout construct; it is more likely that this either represents technical artefact or weak off-target antibody binding to another cell-specific surface receptor.

Expression in blood myeloid cells (Figure 3.15A) was comparable to that in the lung, with highest median fluorescence intensity in neutrophils. Lymphocytes also showed the same expression pattern in lung and blood, with all subsets tested ($CD4^+$ and $CD8^+$ T cells and $CD19^+$ B cells) showing median fluorescence intensities that, while modest compared to the values observed in the

separate myeloid staining panel, were clearly and significantly higher than in negative controls ($p < 0.001$; Figures 3.13, 3.14B and 3.15B).

CD55 was consistently expressed in neutrophils and eosinophils in both lung and blood. In $CD45^-$ non-immune cells and in other lung myeloid cells (macrophages and dendritic cells) it was either not detectable, or the difference in median fluorescence intensity compared to isotype controls was statistically significant ($p < 0.001$) but small, in some cases reflecting expression in a small subset of cells (Figure 3.16A). Circulating monocytes, on the other hand, had similar expression to granulocytes (Figure 3.17A); this suggests that CD55 is downregulated in cells of the monocyte-macrophage lineage on migration to the lung. CD55 was expressed in both T and B lymphocytes in both lung and blood, with highest fluorescence intensities observed in B lymphocytes (Figs 3.16B, 3.17B).

Sex had a prominent effect on CD97 expression, with significantly higher fluorescence intensity in females in all lung leukocyte subsets except for neutrophils and alveolar macrophages, and in all blood lymphocyte subsets (Figures 3.14 and 3.15). There was little evidence of a sex effect on CD55 expression; there was a small decrease in female mice of both genotypes for circulating B lymphocytes only ($p < 0.05$) (Figure 3.17B).

Interestingly, lung immune cell CD55 expression was altered in *Cd97-/-* mice (Figure 3.16). Fluorescence intensity was markedly reduced in neutrophils ($p = 0.001$) and eosinophils ($p = 4 \times 10^{-5}$) of *Cd97-/-* mice, with smaller but significant decreases in $CD11b^+$ $CD11c^-$ monocytes/macrophages ($p = 0.02$) and $CD8^+$ T lymphocytes ($p = 0.002$). These differences were not so readily apparent in peripheral blood, with a significant decrease for neutrophils only ($p = 0.04$).

The flow cytometry data from lung and blood immune cells confirm, at the protein level, the expression distribution predicted by murine transcriptomic data that is largely derived from other tissues. Apart from higher than expected expression of CD55 in granulocytes, the major features of the expression profile are entirely consistent between methods. These consistent observations include the high CD97 expression in neutrophils, broad expression across lymphoid subsets and the unexpected lack of CD97 in alveolar macrophages, as well as the absence of CD55 in most macrophages and dendritic cells. The flow cytometry results also provide data on cell types that are lacking in the other data sets, such as eosinophils and circulating monocytes.

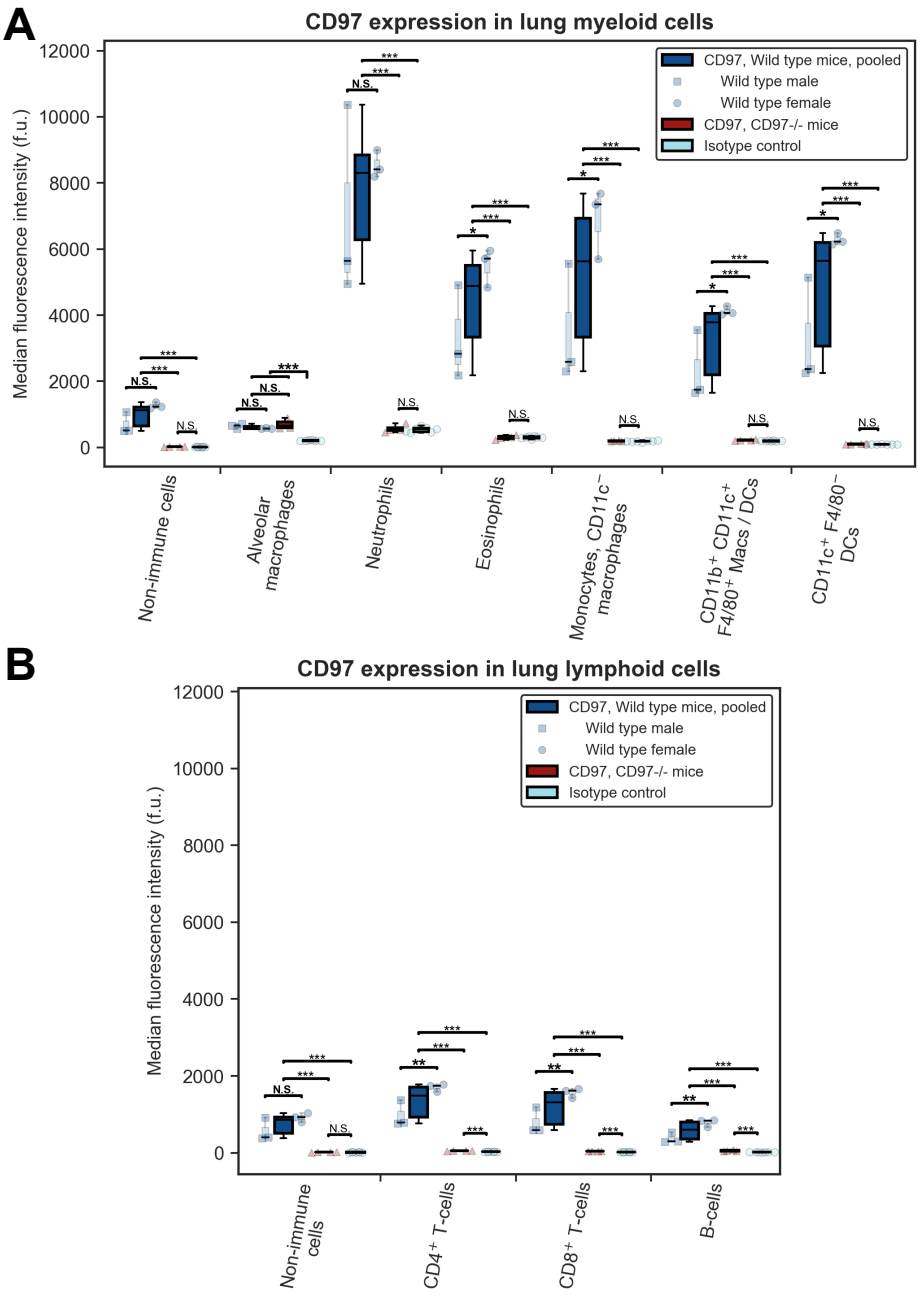

Figure 3.14 – CD97 expression in lung immune cells in healthy mice. CD97 expression was measured by flow cytometry in lung myeloid (A) and lymphoid (B) cells in healthy 8-10 week old male (*n*=3) and female (*n*=3) C57BL/6J mice. *Cd97-/-* mice (*n*=4) and an isotype control antibody (7 mice) were used as negative controls. Differences in median fluorescence intensity were evaluated by general linear models accounting for the effects of sex, antibody, genotype and antibody:genotype interaction. Where the interaction term was significant, Tukey post hoc pairwise comparisons are shown (normal typeface); otherwise, statistical significance of the model main effects is shown (bold font), for sex, genotype, and specific versus isotype control antibody. * $p<0.05$, ** $p<0.01$, *** $p<0.001$, N.S. not significant ($p \geq 0.05$).

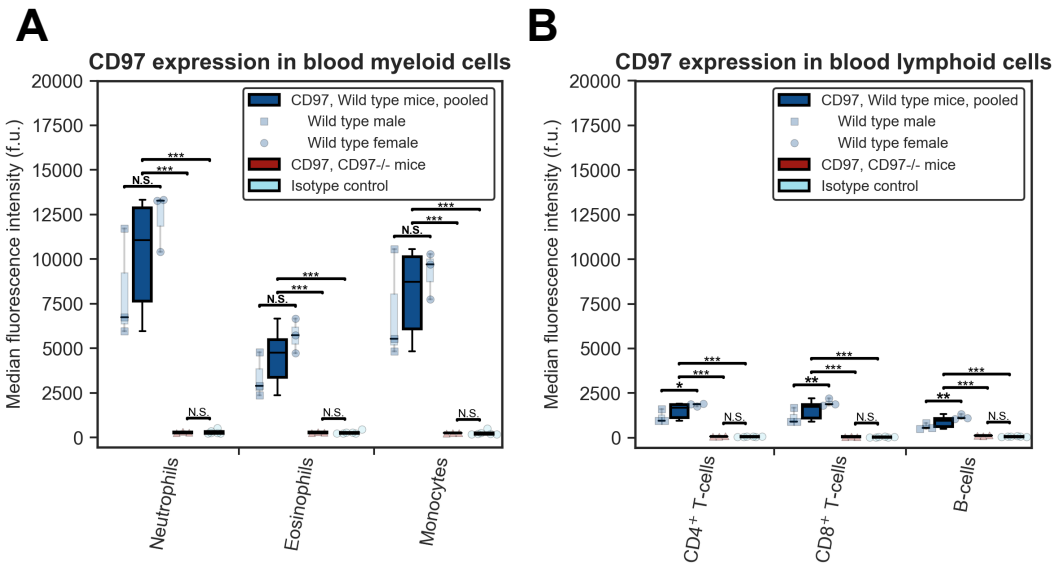

Figure 3.15 – CD97 expression in blood immune cells in healthy mice. CD97 expression was measured by flow cytometry in blood myeloid (A) and lymphoid (B) cells in healthy 8-10 week old male (*n*=3) and female (*n*=3) C57BL/6J mice. *Cd97-/-* mice (*n*=3) and an isotype control antibody (6 mice) were used as negative controls. Differences in median fluorescence intensity were evaluated by general linear models accounting for the effects of sex, antibody, genotype and antibody:genotype interaction. Where the interaction term was significant, Tukey post hoc pairwise comparisons are shown (normal typeface); otherwise, statistical significance of the model main effects is shown (bold font), for sex, genotype, and specific versus isotype control antibody. * $p<0.05$, ** $p<0.01$, *** $p<0.001$, N.S. not significant ($p \geq 0.05$).

Taken together, the RNA and protein-level expression data show a wide range of intercellular interactions in the immune system in which the CD97-CD55 interaction could play a role. These include adhesive interactions between leukocytes and the endothelium (particularly lymphatic endothelium), interactions between antigen presenting cells and lymphocytes, and a wide variety of other contact-dependent interactions between myeloid and lymphoid subpopulations. In mice, CD97 will not be involved in initiation of the immune response to pathogen challenge by alveolar macrophages, although this cannot be excluded as a species-specific function in humans.

Although downregulation of CD97 at the cell surface in response to CD55 binding has previously been described[203], the apparent cell-specific reduction in CD55 expression in neutrophils, eosinophils and T cells (but not B cells) of

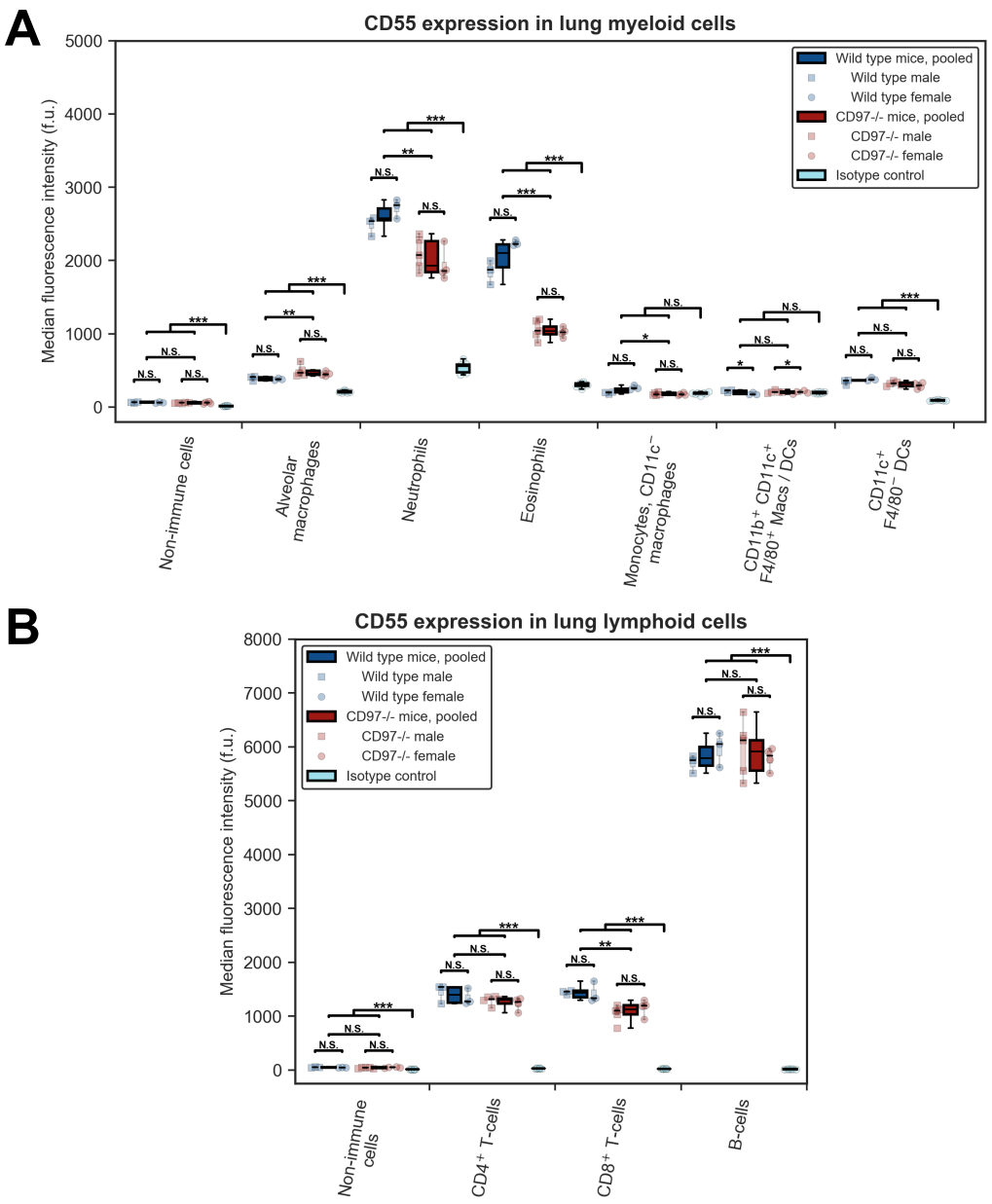

Figure 3.16 – CD55 expression in lung immune cells of healthy mice. CD55 expression was measured by flow cytometry in lung myeloid (A) and lymphoid (B) cells in healthy 8-10 week old wild type (*n*=3 male, 3 female) and *Cd97-/-* (*n*=5 male, 4 female) C57BL/6J mice of both sexes. Differences in median fluorescence intensity were evaluated by general linear models accounting for the effects of sex, antibody, genotype and sex:genotype interaction. Where the interaction term was significant, Tukey post hoc pairwise comparisons are shown (normal typeface); otherwise, statistical significance of the model main effects (genotype, sex, and specific versus isotype control antibody) is shown (bold font). * $p<0.05$, ** $p<0.01$, *** $p<0.001$, N.S. not significant ($p \geq 0.05$)

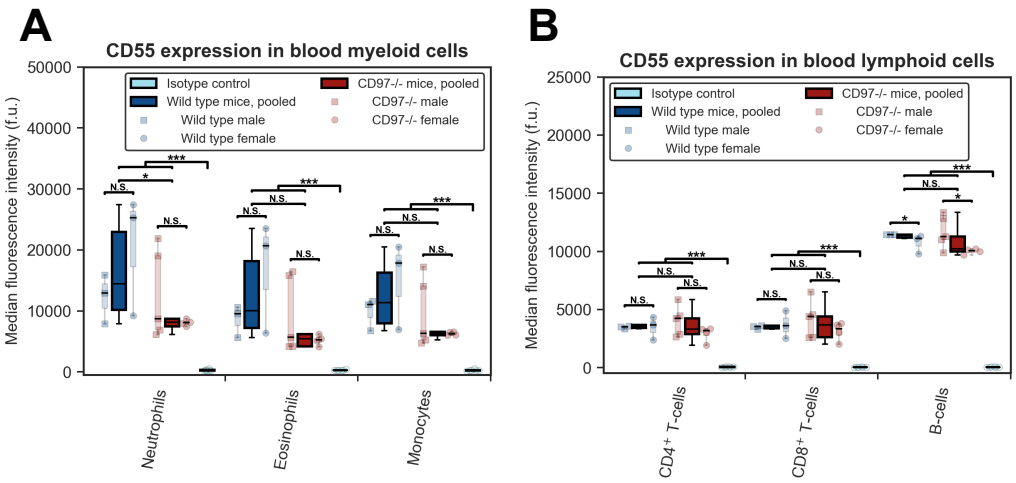

Figure 3.17 – CD55 expression in blood immune cells of healthy mice. CD55 expression was measured by flow cytometry in blood myeloid (A) and lymphoid (B) cells in healthy 8-10 week old wild type (*n*=3 male, 3 female) and *Cd97-/-* (*n*=5 male, 4 female) C57BL/6J mice of both sexes. Differences in median fluorescence intensity were evaluated by general linear models accounting for the effects of sex, antibody, genotype and sex:genotype interaction. Where the interaction term was significant, Tukey post hoc pairwise comparisons are shown (normal typeface); otherwise, statistical significance of the model main effects (genotype, sex, and specific versus isotype control antibody) is shown (bold font). * $p<0.05$, ** $p<0.01$, *** $p<0.001$, N.S. not significant ($p\geq0.05$)

Cd97-/- mice has not previously been reported. The mechanism for this is not yet known, but could reflect reduced activation of these cells: rapid upregulation of surface CD55 has been observed in neutrophils *in vitro* on stimulation with chemoattractants or activators such as ionomycin.[242] It is not yet clear if this reduction in expression could be sufficient to affect the regulatory function of CD55 on complement activation.

3.2.3 *Cd97-/-* mice have no overt baseline phenotype

Given the wide range of immune cells expressing CD97, and the consequent range of cellular interactions which it could influence, an *in vivo* system was necessary to model the effects of CD97 deficiency on the response to infection with influenza A virus. For this purpose, a targeted CD97-knockout mouse strain was used. In the B6;129P2-*Adgre5*^{tm1Dgen}/J strain, one exon common to all isoforms (exon 12/18 for the shortest isoform, or exon 14/20 for the longest) has been replaced with a *lac* operon cassette, such that the C-terminal portion of the protein cannot be synthesised (Fig 3.18A). The manufacturers report no genotype-associated differences in size, fertility, physical examination, gross appearance and histology of multiple organs and tissues on necropsy, bone marrow section evaluation, complete blood count or clinical chemistry, based on evaluation of three to four seven-week old homozygotes of each sex compared to two age-matched controls of each sex (http://www.informatics.jax.org/knockout_mice/deltagen/867.html). In contrast, researchers working with alternative *Cd97-/-* strains, while similarly reporting a lack of gross clinical or pathological changes, have observed an increase in circulating granulocytes in 40% of animals[197,208], without a corresponding change in bone marrow granulocytes. Prior to investigating the response of this strain to viral challenge, it was therefore necessary first to verify whether these mice have any evidence of baseline immune dysregulation in addition to the change in CD55 expression already observed.

First, to confirm that CD97 expression had been effectively eliminated in the mutant strain, lung and blood immune cells were stained with an APC-conjugated antibody raised against the shortest isoform (EGF-1,2,4) of the receptor, the isoform that has the highest affinity for CD55. Analysis by flow cytometry confirmed absence of detectable CD97 on the cell surface (Figure 3.18B, see also Figures 3.14 and 3.15): the fluorescence distribution in *Cd97-/-* mice was indistinguishable from isotype control-stained wild types.

Next, to assess potential effects of CD97 deficiency on immune cell development without pathogen challenge, peripheral blood and pulmonary myeloid and lymphoid populations were compared between *Cd97-/-* and matched wild type C57BL/6J mice.

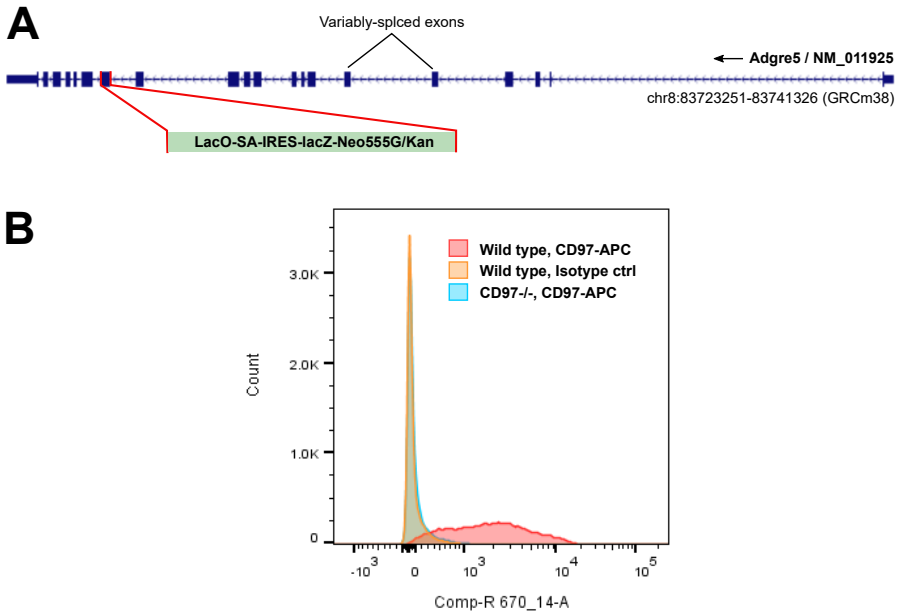

Figure 3.18 – Validation of a *Cd97-/-* mouse strain. A: Details of the knock-out construct. Exon 12 (14 on the longest isoform) has been replaced by the *lac* operon cassette. B: CD97 expression was analysed by flow cytometry in pooled live CD45$^+$ leukocytes from lung homogenates in wild type and *Cd97-/-* (B6;129P2-*Adgre5*^{tm1Dgen}/J) mice, compared to wild type leukocytes stained with an isotype-control antibody.

In the lung, there was no significant difference in relative cell counts between *Cd97-/-* and wild type mice for any of the immune cell populations examined (Figure 3.19). Some sex differences were observed, including relatively higher proportions of neutrophils and CD11b$^+$CD11c$^-$ monocyte/macrophages (a population which includes pulmonary interstitial macrophages) in males, and higher proportions of T cells in females.

In peripheral blood, absolute counts (Figure 3.20) lay within published sex-specific reference ranges for C57BL/6J mice, where these were available, in almost all cases, except for two males (one of each genotype) with marginally elevated eosinophil counts and two with borderline low neutrophil counts (both *Cd97-/-*). There was a significant genotype-sex interaction for total white blood cell count ($p = 0.03$), total lymphocytes ($p = 0.002$), CD4$^+$ T cells ($p = 0.0009$), CD8$^+$ T cells ($p = 0.003$) and B cells ($p = 0.04$). Post hoc pairwise comparisons showed no significant inter-group differences for total white cell count, but for all lymphoid cells the interaction seemed to be driven largely by low counts in the

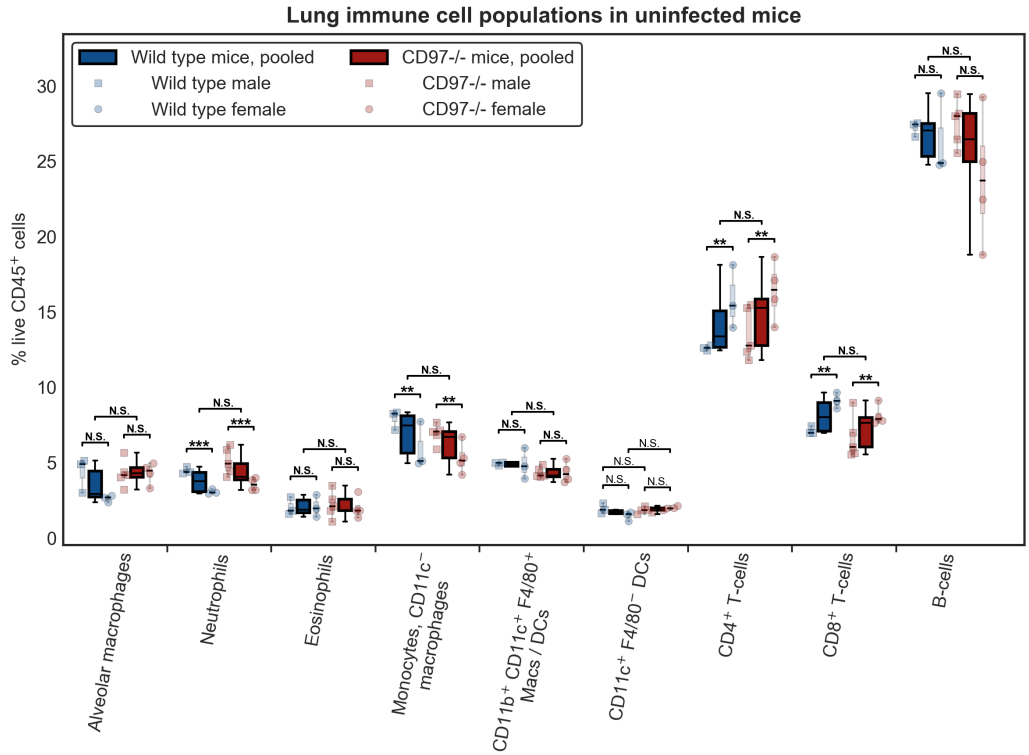

Figure 3.19 – Lung immune cell populations in unchallenged mice. A: Relative populations of immune cells in lung homogenates of 8-10 week old wild type (*n*=3 male, 3 female) and *Cd97-/-* (*n*=5 male, 4 female) mice, assessed by flow cytometry and expressed as percentage of live CD45$^+$ cells. Differences in populations were evaluated by general linear models accounting for the effects of sex, genotype and sex:genotype interaction. Where the interaction term was significant, Tukey post hoc pairwise comparisons are shown (normal typeface); otherwise, p-values for genotype and sex main effects are shown (bold font). * p<0.05, ** p<0.01, *** p<0.001, N.S. not significant (p$\geq$0.05)

wild type female group, which were significantly reduced compared either to wild type males (total lymphocytes, CD8$^+$ T cells, B cells) or to *Cd97-/-* females (total lymphocytes, CD4$^+$ T cells). Other effects of sex included an increase in CD8$^+$ T cells in females versus males for *Cd97-/-* mice only, and a small reduction in monocyte count in females of both genotypes. The sex-genotype interactions could be influenced by artefacts of small group sizes for each genotype-sex combination, and in view of the lack of consistency of genotype effect between sexes, may not be biologically relevant. No significant effect of genotype was seen in myeloid cells. Given the reported association between *CD97* variants

and platelet count in humans (see section 3.2.1.3), the blood platelet count was also compared between *Cd97-/-* and wild type mice: there was a significant reduction in females ($p < 0.001$; data not shown), but no effect of genotype.

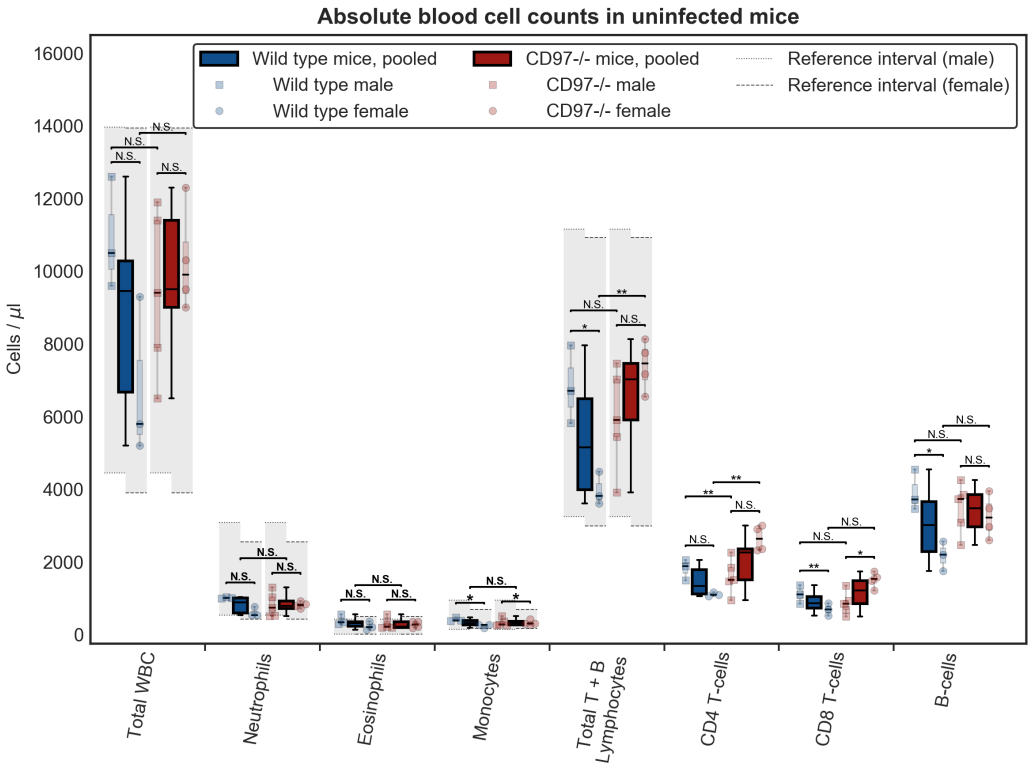

Figure 3.20 – Blood immune cell populations in unchallenged mice. A: Absolute counts of blood immune cell populations from 8-10 week old wild type (*n*=3 male, 3 female) and *Cd97-/-* (*n*=5 male, 4 female) mice, assessed by flow cytometry and haemocytometer count. Shaded boxes indicate sex-specific normal reference intervals for wild type C57BL/6J mice where known (data from Charles River Laboratories). Differences in populations were evaluated by general linear models accounting for the effects of sex, genotype and sex:genotype interaction. Where the interaction term was significant, Tukey post hoc pairwise comparisons are shown (normal typeface); otherwise, overall p-values for genotype and sex are shown (bold font). * $p<0.05$, ** $p<0.01$, *** $p<0.001$, N.S. not significant ($p \geq 0.05$)

The possible effects of sex both on CD97 and CD55 expression, and on baseline cell counts, could lead to increased noise, and a consequent decrease in power, in experiments using mice of both sexes. Challenge studies were therefore performed in female mice only.

The effect of genotype on individual T-cell subsets in the lung was not significant, but as there was an apparent divergence of effects on $CD4^+$ and $CD8^+$ subsets in a manner that was consistent between sexes, the CD4:CD8 ratio was also compared between groups (Figure 3.21). There was a significant increase in the ratio (i.e. relatively more $CD4^+$ compared to $CD8^+$ lymphocytes) in the lungs of *Cd97-/-* mice. The difference was not significant in peripheral blood.

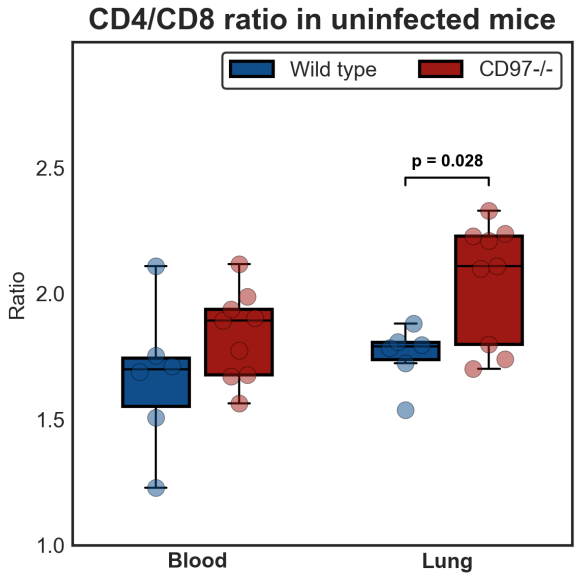

Figure 3.21 – Lung and blood CD4/CD8 ratio in unchallenged mice. $CD4^+$ and $CD8^+$ T-cell populations were measured by flow cytometry as in Figure 3.19, in 6 wild type versus 9 *Cd97-/-* mice. Analysis was by general linear models with log-transformed ratios, accounting for the effects of sex, genotype and sex:genotype interaction. Since there was no significant effect of sex, or the sex-genotype interaction, sexes have been pooled for clarity.

This mouse strain had no overt disease or major defects in immune cell population development before pathogen challenge. The circulating granulocytosis reported by other authors was not replicated here. Different knockout constructs were used in previous reports, one with deletion of exons 2 to 5 (the signal peptide and first EGF domain) and the introduction of a frame shift[197], and the other with replacement of exons 2 to 12,[208] but all three would be expected to completely abrogate receptor function. Since the difference was only seen in a subset of mice in previous reports, it is possible that it would have required a larger sample size to demonstrate this in this strain. Alternatively, the difference

could arise from differences in the genetic background or the health status of the mice. The lack of replication does however bring into question whether the altered granulocyte homeostasis, for which a mechanism has not been definitively determined, represents a true biological effect of the CD97 receptor.

The difference in T-cell ratios could represent a subtle deficit in T-cell development, differentiation or migration to the lungs. This has not been reported previously. A similar pattern has however been seen in the only other published study reporting lung cell lymphoid counts in *Cd97-/-* mice to date: 24 hours after intravenous injection of *Listeria monocytogenes*, wild type mice had similar proportions of $CD4^+$ and $CD8^+$ cells in lung tissue while *Cd97-/-* mice had a CD4:CD8 ratio of approximately 2:1; although the differences in individual T-cell subsets did not achieve statistical significance, the ratio was not analysed.[208] The same pattern was not seen in spleen, or in blood, bone marrow or spleen of unchallenged mice reported elsewhere[197], suggesting this could be a tissue-specific effect.

3.2.4 CD97 deficiency modulates disease severity after IAV challenge in a murine model of severe influenza

Having established the baseline phenotype, the next objective was to use the *Cd97-/-* strain to model the effects of CD97 on the development of severe influenza. While in most people influenza is primarily a disease of the upper respiratory tract, lower respiratory tract infection in severely affected patients can lead to diffuse alveolar damage and acute respiratory distress syndrome (ARDS), with exudation of fluid into the airway and consequent respiratory failure. Respiratory pathology is attributable both to the direct effects of the virus on infected epithelial cells, and to the effects of infiltrating inflammatory cells. Pathological lesions in severe cases in humans include widespread necrosis and desquamation of the tracheobronchial epithelium, peribronchial and interstitial mononuclear or mixed inflammatory infiltrate, interstitial oedema, small vessel thrombosis, and intraluminal oedema, exudate, haemorrhage, and sometimes hyaline membrane formation, in alveoli and small airways. Secondary bacterial pneumonia is a common complication.[243–246]

While wild mice are generally resistant to IAV infection, laboratory strains such as BALB/c and C57BL/6 are susceptible due to a loss-of-function mutation in the interferon-stimulated antiviral gene *Mx1*.[247] Pathological lesions with experimental infection in mice are initially characterised by broncho-interstitial pneumonia, with interstitial oedema and mixed mononuclear-lymphocytic inflammatory infiltrate, progressing later to diffuse alveolar damage and alveolar collapse in more severe cases. The bronchial epithelium remains largely intact until late stages of severe infection. Thus the mouse model, although not perfect due to the less prominent epithelial pathology, shares sufficient features with severe disease in humans to make it a useful model.[248] In this study, an experimental model of H1N1 challenge in wild type and *Cd97-/-* mice was used to evaluate the influence of CD97 on the development of severe disease after infection with IAV.

3.2.4.1 Clinical phenotype

To establish whether CD97 deficiency alters the gross clinical phenotype in the murine model of severe influenza, a series of challenge experiments were performed with IAV H1N1 strains A/Eng/195 and A/Cal/04/09, both closely related

to circulating strains involved in the 2009 pandemic. Mice were monitored for five to seven days after intranasal administration of either 50 plaque-forming units (pfu) A/Eng/195, or either 100 or 1000 pfu A/Cal/04/09. Doses were selected on the basis of pilot dose-finding experiments, to give either a mild to moderate phenotype in which no mouse was expected to reach the humane endpoint within seven days, or a moderate to severe phenotype (1000pfu A/Cal/04/09) in which humane endpoints were reached in five to seven days. Weight loss was used as the primary outcome measure; although clinical signs were monitored to assist with determining humane endpoints, these data were subjective and collected without blinding, and so were not used for analysis.

IAV challenge led to progressive dose-dependent weight loss in both wild type and *Cd97-/-* mice (Figure 3.22), associated with development of clinical signs including hunched posture, piloerection and increased respiratory rate and effort. Mixed model analysis of weight loss after A/Eng/195 challenge (Figure 3.22A, pooled from three independent studies) showed significant effects of time ($p < 2.2 \times 10^{-16}$), study:time interaction ($p = 6.2 \times 10^{-9}$) and genotype:time interaction ($p < 0.005$). Post hoc analysis of genotype:time effects indicated a statistically significant increase in weight loss in *Cd97-/-* mice at day 6 post challenge (estimated effect size 3.2% of starting weight, 95% C.I. 1.6 - 4.8, $p = 0.001$), while the difference at day 5 was approaching significance (estimated effect size 2.2%, 95% C.I. 0.6 - 3.8%, $p = 0.051$). For challenges with A/Cal/04/09 (Figure 3.22B and C), no effect of genotype was observed at either of the doses tested. These data provide evidence of a transient, challenge-specific exacerbation of disease severity after IAV challenge, albeit with a modest magnitude of effect.

3.2.4.2 Lung and BALF characteristics at day 7 post-infection suggest mixed effects on disease severity.

The weight loss curve for 50 pfu A/Eng/195 challenge (Figure 3.22A) suggests that after a peak difference in weights at day 6 post challenge, the phenotypes are converging by day 7, with more rapid weight loss in wild type mice over the final 24 hours. As weight loss is a relatively crude indicator of disease severity, which can reflect a number of factors including inappetence and dehydration, lung weights and bronchoalveolar lavage fluid (BALF) were assessed as indices of the overall severity of pulmonary pathology at the experimental end-point.

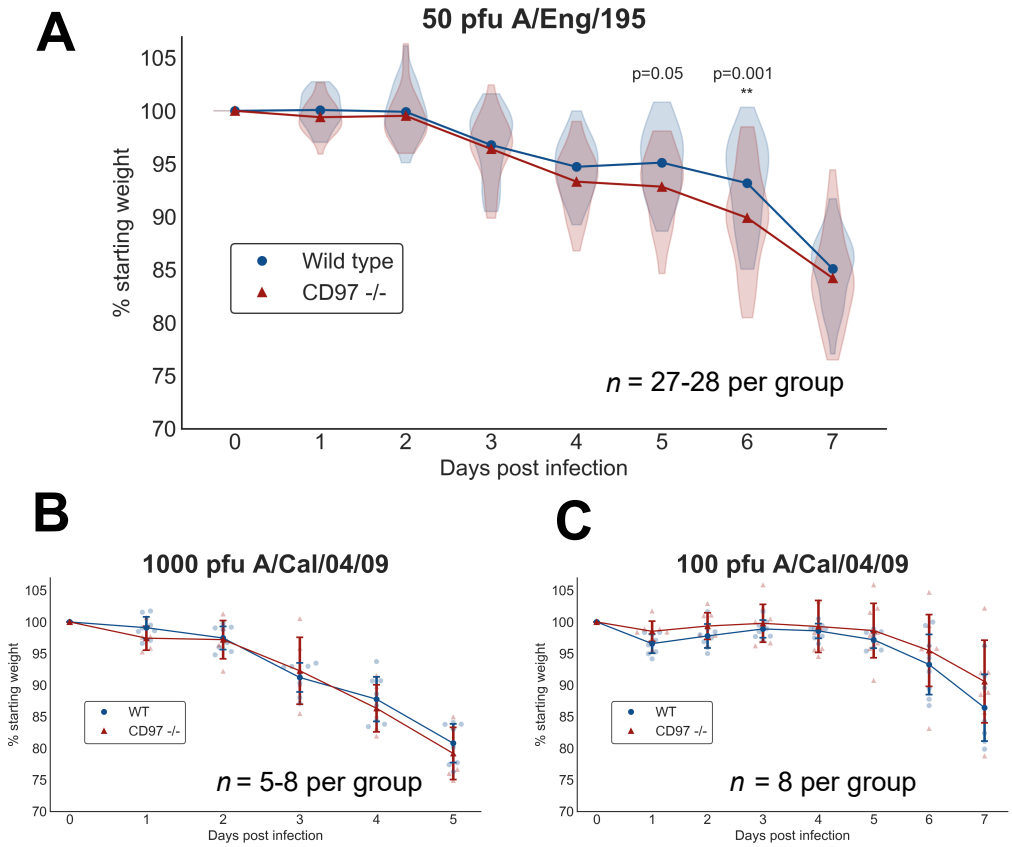

Figure 3.22 – Weight loss in *Cd97-/-* versus wild type mice after IAV challenge. A: Pooled data from three independent challenge experiments with A/Eng/195. Mean weight loss at each time is shown. Violin plots represent the Gaussian kernel density estimate of the distribution of raw data points; height of the violin plots indicates the range. B and C: Weight loss time courses for single challenge studies with 1000 or 100 pfu A/Cal/04/09. Mean ± standard deviation and individual data points are shown. P-values for genotype effects at individual time points (from post hoc tests of mixed model analysis) are displayed where reaching or approaching the threshold for statistical significance (< 0.05).

Lung weight reflects the degree of oedema in the lung tissue. Lung weights (Figure 3.23A) were consistently reduced in *Cd97-/-* mice compared to wild type mice on day 7 after IAV challenge (mean 28 mg difference, 95% C.I. 12 - 44 mg, p = 0.001). Total BALF cell counts (Figure 3.23B), which reflect the severity of the intraluminal inflammatory exudate, were similarly reduced in *Cd97-/-* mice (geometric mean 45% decrease, 95% C.I. 27 - 58%, p = 0.0001). A small reduction in BALF total protein concentration provided further evidence for reduced exudation across the alveolar membranes (p = 0.04, Figure 3.23C). Coupled

with the weight loss data, these results suggest that the role of CD97 in influenza, rather than comprising a simple protective effect, entails a more complex effect on the host response: initial attenuation of disease severity in the presence of CD97 is not maintained, and there is some evidence of exacerbation of pulmonary pathology later in the course of disease.

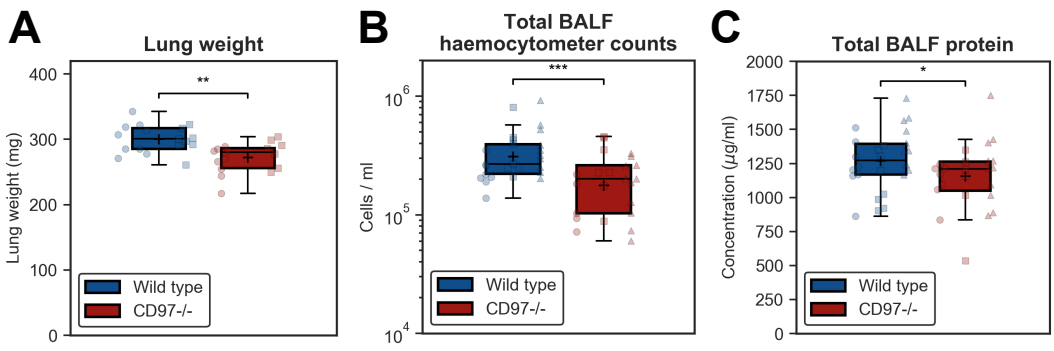

Figure 3.23 – Lung weight and BALF characteristics 7 days after A/Eng/195 challenge. Mice were challenged with 50 pfu A/Eng/195, and lung tissue and BALF were harvested after 7 days. A: Total weight of right lung lobes (data from 2 independent experiments). B and C: Total BALF cell haemocytometer counts (B) and total protein measured by BCA colourimetric analysis (C), from 3 independent experiments, total n = 23-26 per group. '+' denotes the mean (A,C) or geometric mean (B). Data points from a single experimental replicate are assigned a common symbol. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, from general linear model analysis.

3.2.4.3 Histopathology

To investigate whether the observed differences in disease severity were caused by differences in the type or extent of pulmonary lesions, the overall pattern of histopathology was first evaluated in lung tissue harvested seven days after A/Eng/195 challenge (Figure 3.24). Lung sections showed heterogeneous but widely distributed lesions across all lung lobes for both genotypes. Lesions were consistent with pathology previously described for the mouse-adapted A/PR/8/34 IAV strain[248], and consisted principally of varying degrees of alveolar infiltrate, interstitial thickening and infiltration (Figure 3.24B/E), lymphoid cell-rich peribronchial and perivascular leukocyte accumulations (Figure 3.24A/D),

and localised haemorrhage (Figure 3.24C/F). Necrosis of the bronchial epithelium, only expected in late stages of disease approaching the humane endpoint, was observed occasionally, but hyaline membranes or fibrosis were not detected. The findings were similar between genotypes, and histopathology scores, assessed using a previously described scoring system[155], were not significantly different (wild type median 5.5, interquartile range (IQR) 1.0; *Cd97-/-* median 7, IQR 1.5; p = 0.15 by Wilcoxon rank-sum test).

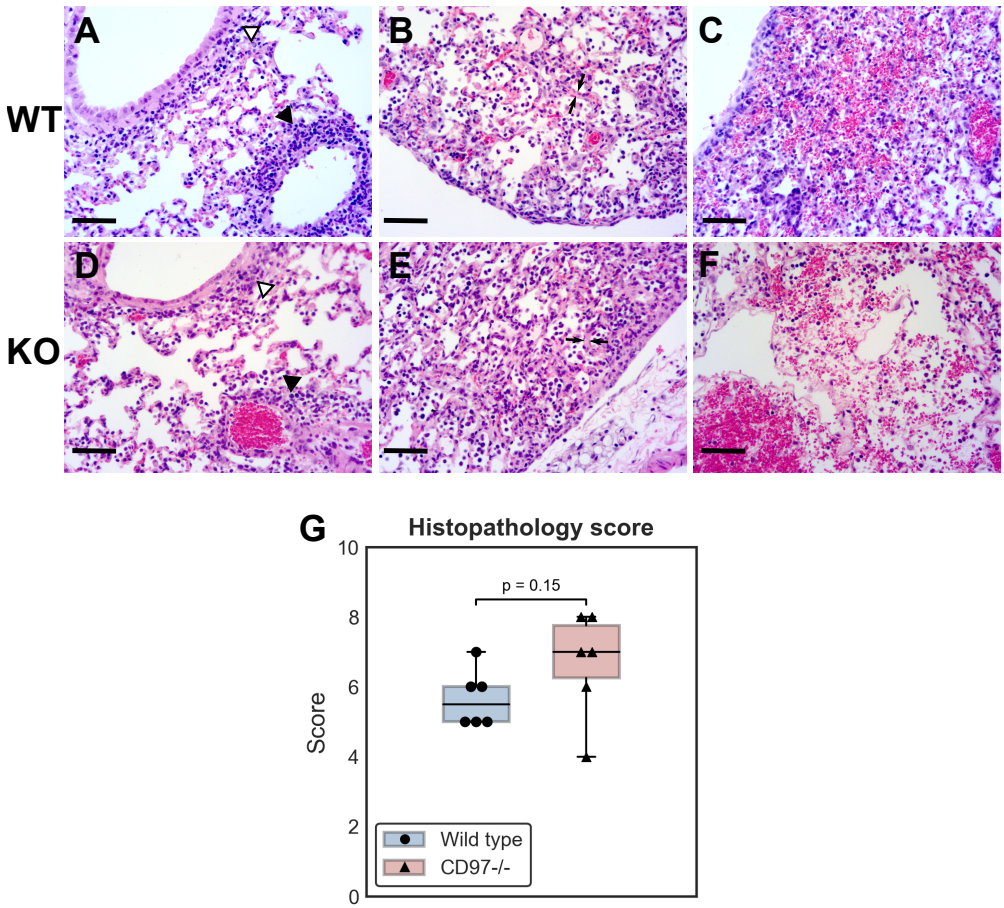

Figure 3.24 – Lung histopathology in *Cd97-/-* versus wild type mice after IAV challenge. Haematoxylin / eosin stained sections of lung from wild type (A-C) and *Cd97-/-* (D-F) mice 7 days after challenge with 50 pfu A/Eng/195. Both genotypes show similar pathological changes including peribronchial and perivascular infiltrate (A and D, open and closed arrowheads respectively), interstitial thickening and infiltrate (B and E; arrows show examples of alveolar wall thickening), and regional haemorrhage (C and F). Scale bars indicate 100µm. G: Histopathology scores, compared by Wilcoxon rank-sum test (*n*=6 per group).

3.2.5 Viral loads are increased after IAV challenge in CD97-deficient mice.

3.2.5.1 Viral titres in lung tissue and broncho-alveolar lavage fluid after IAV challenge

Exacerbation of disease severity with CD97 deficiency could arise from impaired viral clearance or increased tissue damage from a dysregulated host inflammatory response. To distinguish between these possibilities, viral titres were measured in lung homogenates and BALF after IAV challenge.

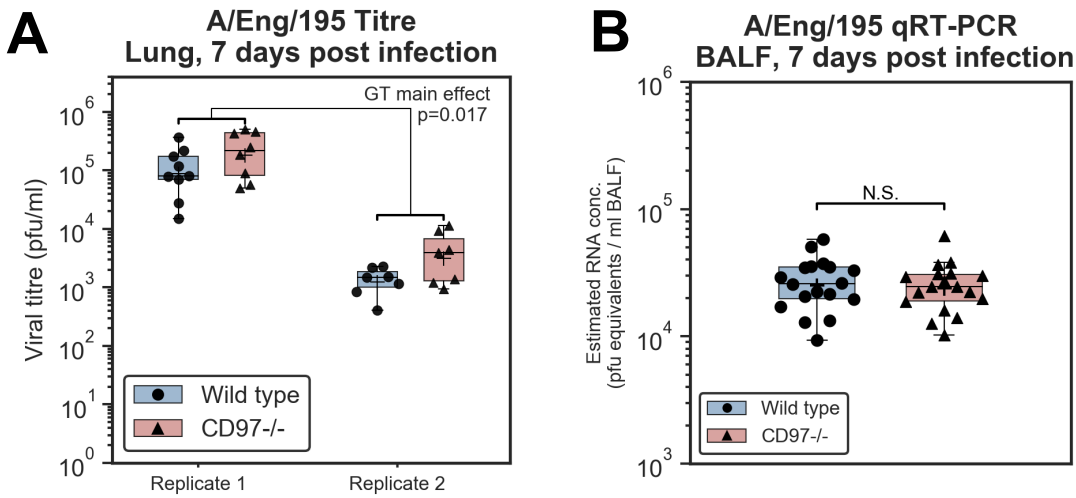

Figure 3.25 – Effect of CD97 on viral titres 7 days after IAV challenge. Lung and BALF were harvested 7 days after challenge with 50 pfu A/Eng/195. Effects of genotype have been analysed by general linear models, adjusting for batch effect between replicates. A: Viral titre, by plaque assay, in lung homogenates (all right lobes); *n*=7-9 per group per replicate. B: qRT-PCR estimate of viral RNA concentration in BALF 7 days after A/Eng/195 challenge, based on the geometric mean of values for Segments 2 and 7, compared to a standard curve of diluted RNA from viral stock of known titre. As there was no significant batch effect between replicates, 2 replicates have been pooled for clarity. *n*=8 per group. '+' denotes the geometric mean. N.S.: not significant ($p \geq 0.05$).

As modulation of the weight loss phenotype was seen only in the later stages of our challenge models, viral loads in lung and BALF were evaluated at 7 days after A/Eng/195 challenge (Figure 3.25A). Lung titres showed a strong batch

effect between experimental replicates, likely to be due to small differences in sample timing at a time when titres are expected to be declining rapidly, possibly compounded by differences in sample handling and lack of robustness of plaque formation of this strain *in vitro*. There was a small but consistent increase in titres in *Cd97-/-* mice that retained statistical significance after adjusting for this batch effect (p = 0.017 for genotype main effect in a general linear model, estimated effect size 2.3-fold change, 95% C.I. 1.2-4.4 fold change). This could reflect enhanced viral replication, or delayed viral clearance by the immune system. Viral RNA quantification in BALF by qRT-PCR (Figure 3.25B) did not reflect this increase, and was not different between genotypes (p = 0.79 by general linear model). This will however include RNA from non-viable virus, and may be slower to reflect changes in viral replication and clearance. In the more severe 1000 pfu A/Cal/04/09 challenge model (Figure 3.26A), the increase in lung titre, measured by plaque assay, did not reach statistical significance (p = 0.15 by *t*-test, 95% C.I. 0.96-1.29 fold change in *Cd97-/-* mice).

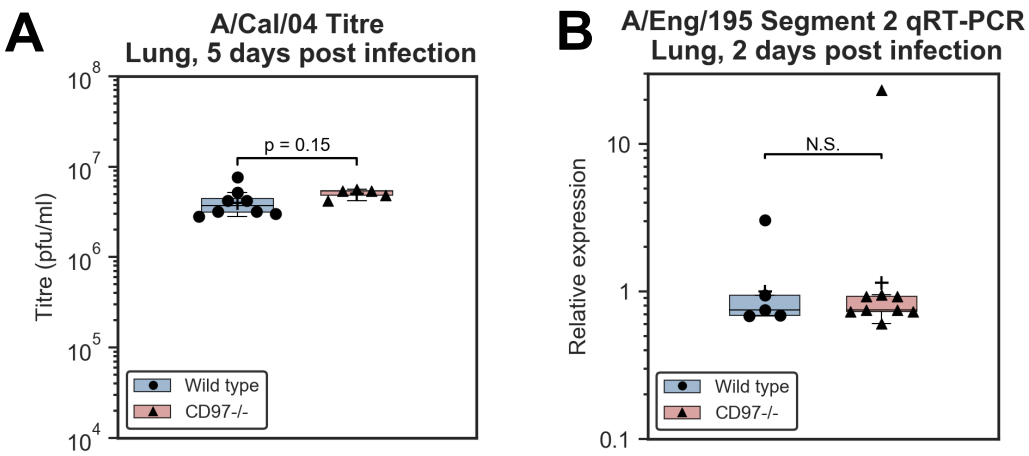

Figure 3.26 – Effect of CD97 on viral titres in other IAV challenge models. A: Viral titre, by plaque assay, in lung (right middle lobe) 5 days after challenge with 1000 pfu A/Cal/04/09, *n*=5-8 per group. B: Relative viral load in left lung lobe 2 days after challenge with 850 pfu A/Eng/195, assessed by qRT-PCR for Segment 2 (PB1), *n*=5-9 per group. Raw CT values have been normalised to the mean of 3 endogenous controls (*Gapdh*, *Hprt1* and *B2m*), and relative expression values ($2^{-\Delta CT}$) subsequently normalised to the geometric mean for wild type mice. Statistical analysis was by Student's t-test on ΔCT or log-transformed titre values. '+' denotes the geometric mean. N.S.: not significant ($p \geq 0.05$).

To assess whether CD97 deficiency influences early viral replication in the host, viral RNA was measured in lung samples two days after high-dose A/Eng/195 challenge by qRT-PCR (Figure 3.26B). Viral RNA for segment 2 was detectable in the left lung (CT values lower than no-reverse-transcriptase or no-template controls by $\geq$5) in all animals, with substantially higher expression in a single animal in each group. Segment 7 RNA was undetectable or had high CT values ($\geq$37) in most animals, with multiple peaks on melt curve analysis; the assay was thus considered unreliable and not used for analysis. There was no difference in viral RNA between genotypes (Wilcoxon rank-sum test, p = 0.80). Although localisation of viral replication and pathology within regions of the lung at this early time point could render the tissue sample non-representative of the whole lung in some individuals, these data do not provide support for an early effect on viral replication or restriction. This is consistent with the lack of difference in weight loss seen in the early stages of the A/Eng/195 challenge model.

3.2.5.2 *In vitro* models show no effect of CD97 deficiency on viral replication or restriction.

Given the complexity of the *in vivo* system, and the varied biological mechanisms which could affect viral titres, a simpler *in vitro* system was used to clarify whether CD97 influences intrinsic cellular resistance to IAV replication. Genome-wide RNA interference and CRISPR screens using an epithelial cell line (A549 cells) have not identified CD97 as a host dependency factor for IAV entry and replication,[67,69,249,250] suggesting that CD97 deficiency will have no effect on IAV replication in isolated cultured cells. However, in addition to potential physiological differences between immortalised cell lines and primary epithelial cells, CD97 is expressed only at relatively low levels in A549 cells compared to cells of immune origin, with transcript levels in the Fantom5 database, for example, lower than in almost all immune cell groups and 10-fold or more lower than in most myeloid cells[50]; previous models may thus have lacked sensitivity.

Although IAV replication occurs primarily in respiratory epithelial cells, pulmonary macrophages are key players in the initial response to IAV infection, and are involved in sensing the pathogen and in triggering and regulating the resulting immune response.[251,252] In some cases, productive infection of macrophages

can occur, depending on macrophage subtype and viral strain, whereas in other cases infection is abortive, due to barriers including inefficient viral entry, failure to escape from endosomes, or defective assembly and budding of virions the cell membrane.[251] Both productive and abortive infections will result in release of a range of cytokines and other mediators that will direct the ensuing inflammatory response. An *in vitro* model was therefore used to investigate whether CD97 deficiency would modulate viral restriction in murine macrophages.

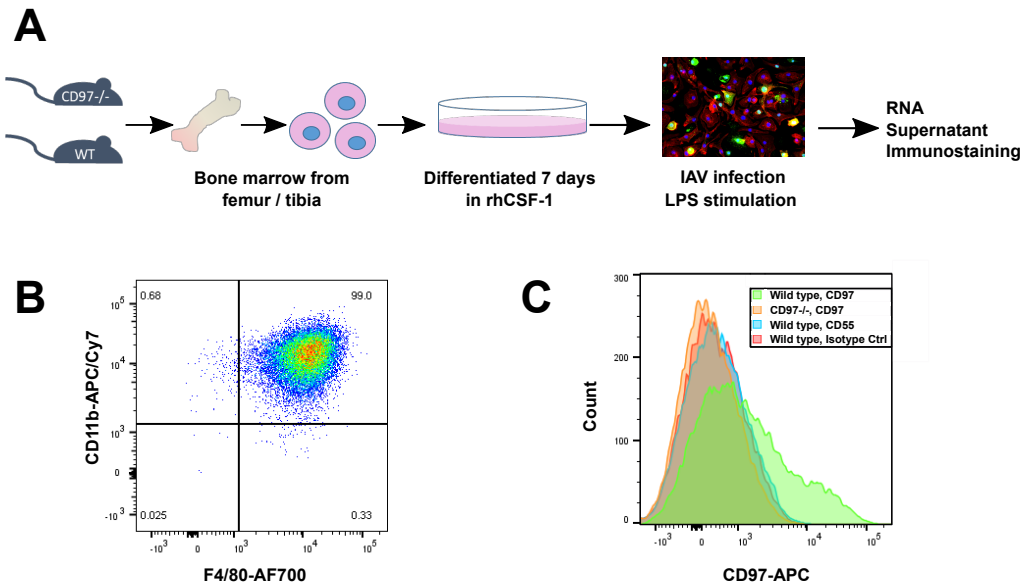

Figure 3.27 – Bone marrow-derived macrophages as an *in vitro* model of macrophage function. A: Workflow for production and stimulation of murine bone-marrow derived macrophages. B: Characterisation of the BMDM population by staining for myeloid marker CD11b (APC-Cy7-conjugated antibody) and the pan-macrophage marker F4/80 (Alexa Fluor 700). The displayed plot has been gated on live CD45$^+$ cells. C: Histograms of CD97 and CD55 expression versus isotype control (all APC-conjugated antibodies) in wild type BMDMs, and confirmation of lack of CD97 expression in *Cd97-/-* BMDMs. The plot has been gated on CD11b$^+$ F4/80$^+$ macrophages.

Bone marrow-derived macrophages (BMDMs) were produced from wild type and *Cd97-/-* mice by differentiation of precursor cells from bone marrow using recombinant human colony stimulating factor 1 (rhCSF1) for seven days (Figure 3.27A). Analysis by flow cytometry confirmed that the resulting cells consisted of a highly pure population of CD11b$^+$ F4/80$^+$ macrophages (Figure 3.27B), which expressed CD97 (albeit with a broad range of expression across the cell population) but did not express detectable levels of CD55 (Figure 3.27C).

Macrophages were infected with IAV strains A/Cal/04/09 or A/WSN/33 at a multiplicity of infection (MOI) of five, and viral replication or restriction was assessed by immunofluorescence (IF) for viral nucleoprotein (NP) in the cells and by measuring supernatant viral titres.

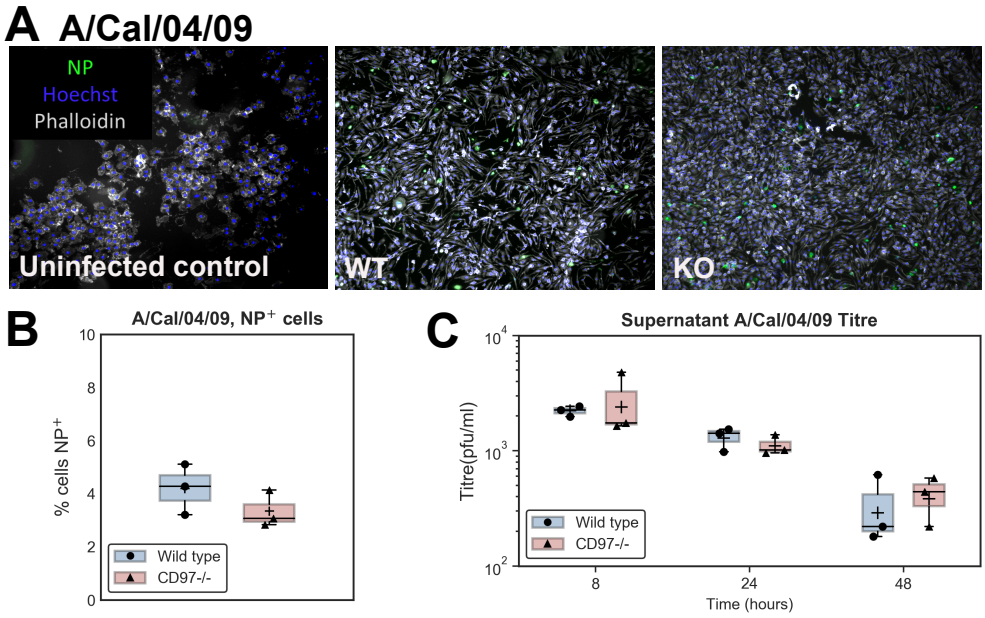

Figure 3.28 – CD97 has no effect on abortive A/Cal/04/09 infection in cultured macrophages. BMDMs from wild-type and *Cd97-/-* mice (*n*=3 per group) were challenged with A/Cal/04/09 at MOI 5, and virus was quantified by nuclear NP staining on IF or by titration in culture supernatants by plaque assay. A: Low-power (40X) IF images of BMDMs 24h after infection with A/Cal/04/09, or uninfected control cells (100X). NP-staining (AF488) is shown in green, Hoechst nuclear staining in blue, and cellular actin (phalloidin-Texas Red) in grey. B: Boxplot of the percentage of macrophages with positive nuclear staining for NP. C: Supernatant A/Cal/04/09 titres decline rapidly with time. '+' denotes the mean (B) or geometric mean (C). Groups have been compared with Student's *t*-test (B) or with linear mixed models to account for repeated measures over time (C).

The 2009 pandemic H1N1 strain A/Cal/04/09 resulted in abortive infection in BMDMs. By 24 hours after infection, less than 6% of macrophages contained detectable viral nucleoprotein (Figure 3.28A and B), and supernatant titres decreased rapidly over the first 48 hours (time effect p = 0.001 in mixed models; Figure 3.28C), suggesting detection of residual input virus only rather than viral replication. The cultured macrophages were, in contrast, permissive to infection with the alternative H1N1 strain A/WSN/33 at the same MOI, with up to 40% of cells containing detectable viral nucleoprotein (with staining overlying the nu-

cleus in most cells) by 24 hours post-infection (Figure 3.29A and B), and titres more than 10-fold higher at 24 hours than those for A/Cal/04/09 (Figure 3.29C). While active viral replication cannot be proven from this single time point, our data are consistent with previous reports of productive infection with A/WSN/33, but not A/Cal/04/09, in the RAW264.7 murine macrophage cell line: this difference has been linked at least in part to strain differences in neuraminidase (NA).[253,254] There was no difference between genotypes in either the percentage of NP-positive cells or in titres at any time for either infection model, indicating that CD97 is not required for the processes of viral entry, replication, or restriction in macrophages.

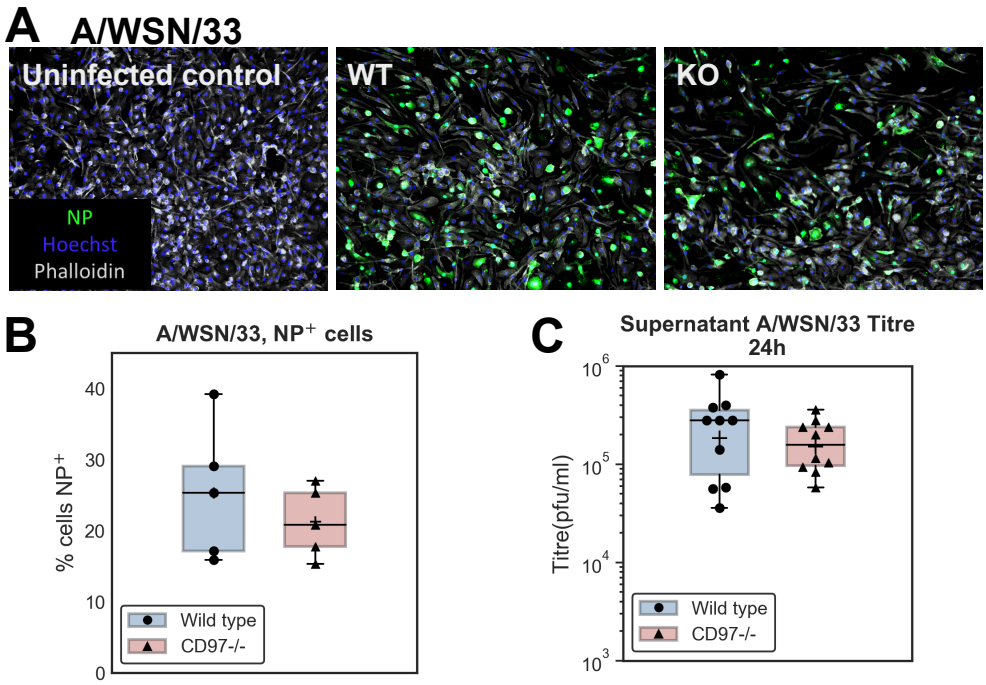

Figure 3.29 – CD97 has no effect on A/WSN/33 infection in cultured macrophages. BMDMs from wild-type and *Cd97-/-* mice were challenged with A/WSN/33 at MOI 5, and virus was quantified by nuclear NP staining on IF or titrated in culture supernatants by plaque assay. A: IF images (100X magnification) of BMDMs 24 hours after A/WSN/33 challenge. NP staining (AF488) is shown in green, Hoechst nuclear staining in blue, and cellular actin (phalloidin-Texas Red) in grey. B: Percentage of macrophages with positive nuclear staining for NP (n=5 per group). C: Supernatant A/WSN/33 titres after 24 hours (n=10 per group). '+' denotes the mean (B) or geometric mean (C). Groups have been compared with Student's t-test with log-transformation where appropriate.

Since the type I interferon response is an integral part of intrinsic cellular resistance to IAV infection, IFNα production was also compared between *Cd97-/-* and wild type BMDMs after A/WSN/33 infection (Figure 3.30). Supernatant IFNα concentrations, quantified by enzyme-linked immunosorbent assay (ELISA) were below the limit of detection in mock-treated macrophages and in the majority of samples after 4 hours of IAV challenge (data not shown), but rose rapidly between 4 and 24 hours. There was no significant difference between genotypes at 24 hours post challenge (Student's t-test for independent samples, p = 0.12), indicating that CD97 does not affect the primary type I interferon response in BMDMs.

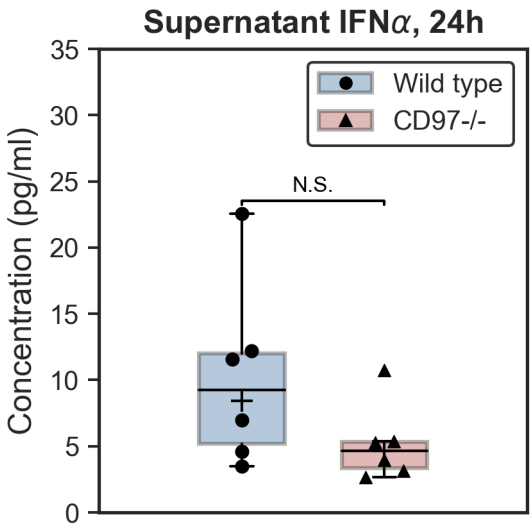

Figure 3.30 – CD97 has no effect on macrophage IFNα production after A/WSN/33 infection. IFNα concentration, measured by ELISA, in supernatants from BMDMs challenged with A/WSN/33 as in Figure 3.29 (*n*=6 per group). '+' denotes the geometric mean. Groups have been compared with Student's *t*-test with log-transformation.

Taken together, the *in vivo* and *in vitro* results indicate that CD97 does not influence initial viral replication or restriction, but may promote efficient clearance of IAV. The *in vivo* evidence of increased viral loads in CD97-deficient mice is to a certain extent equivocal, and depends both on challenge dose and sample type, but the effect could be underestimated by the analyses presented here as the time of sampling was 24 hours after the maximal difference in weight loss.

Modulation of immune functions necessary for clearance of infected cells would be consistent with the time scale of effect and with the expression profile and previously reported functions of this receptor. I thus proceeded to investigate whether the immune response to IAV challenge was perturbed in *Cd97-/-* mice.

3.2.6 CD97 deficiency modulates pulmonary immune cell infiltration in response to IAV challenge

Immune cell migration to the lung following IAV infection is necessary for clearance of infected cells, but can also mediate or exacerbate tissue damage. Given the prior evidence of immune dysregulation seen in the baseline increase in the pulmonary CD4:CD8 ratio observed in unchallenged mice (Figure 3.21), and in the circulating and tissue neutrophilia previously reported in *Cd97-/-* mice following *Listeria monocytogenes* challenge[208], the immune cell response to IAV infection in *Cd97-/-* and wild type mice was assessed by flow cytometry.

3.2.6.1 Immune cell populations in lung and peripheral tissues after IAV challenge

Bronchoalveolar lavage fluid and blood were collected from *Cd97-/-* and wild type mice seven days after challenge with 50 pfu A/Eng/195. Mock-infected control wild type mice (data not shown) had a low cellularity in lavage fluid, with total counts between 5,000 and 12,000 cells/ml, with immune cells consisting primarily of alveolar macrophages (32 - 95%) and lower proportions of lymphocytes (3 - 43%) and neutrophils (0.1 - 10%). Mice infected with IAV had a large increase in BALF cellularity, which was more marked in wild type mice (see Figure 3.23B). Consistent with the expected pulmonary immune response to viral challenge, predominant inflammatory cell types included neutrophils, alveolar macrophages, and $CD8^+$ T lymphocytes. Statistically significant decreases in absolute counts in *Cd97-/-* mice were seen for several myeloid cell populations, including alveolar macrophages, but these were mostly in proportion to the decrease in total cell count: a small decrease in relative abundance was seen for eosinophils only (Figure 3.31). No absolute or relative difference was seen for neutrophils. For lymphoid cells, there was a decrease in relative and absolute

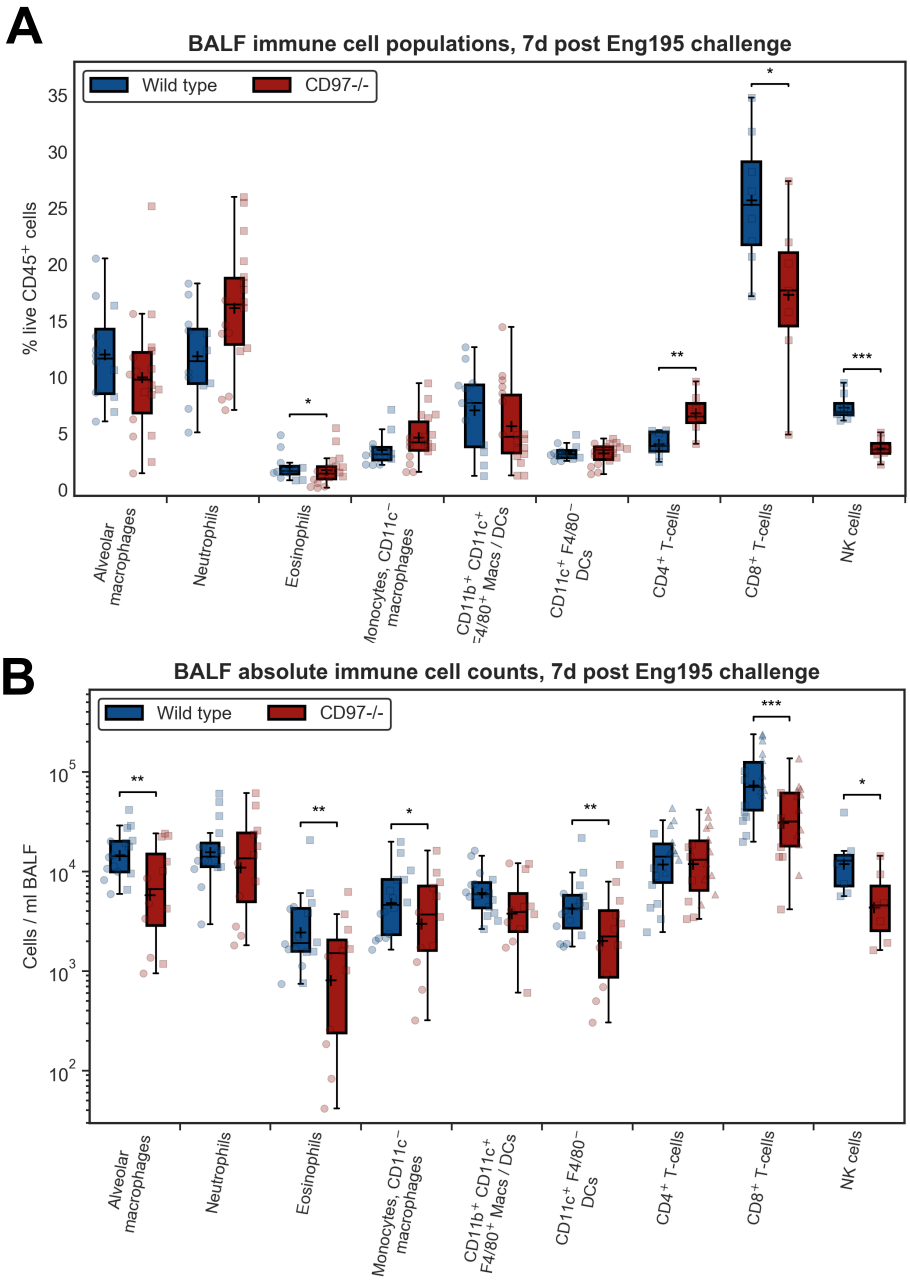

Figure 3.31 – Immune cell populations in BALF 7 days after A/Eng/195 challenge. Mice were challenged with 50 pfu A/Eng/195. Relative immune cell populations, as percentage of live CD45$^+$ cells (A), and absolute cell counts (B) in BALF were determined by flow cytometry after 7 days. Data are from 1-2 independent studies, total n = 7-18 per group. '+' denotes the mean (A) or geometric mean (B). Data points from a single experiment are assigned a common symbol. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, by t-test (single experiment) or general linear model to adjust for batch effects of experimental replicates.

counts of CD8⁺ T cells and NK cells in *Cd97-/-* mice, and a relative increase in CD4⁺ T cells without a change in absolute numbers.

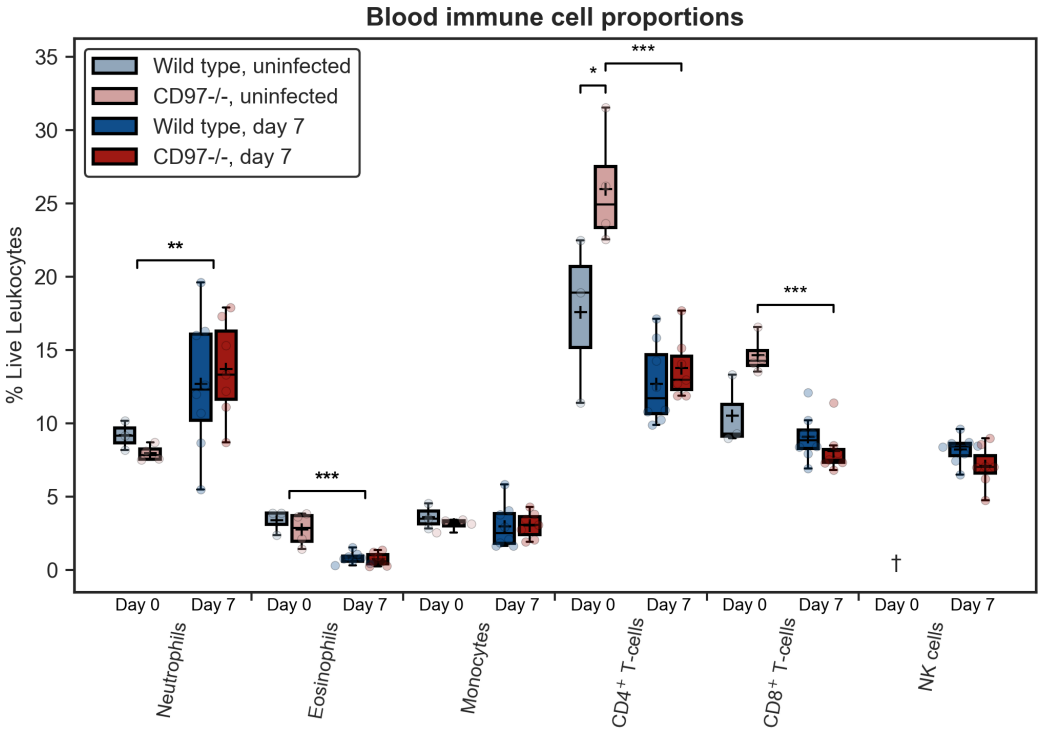

Figure 3.32 – Circulating immune cell populations 7 days after A/Eng/195 challenge. Mice (7-8 per group) were challenged with 50 pfu A/Eng/195 and blood immune cell populations were determined by flow cytometry after 7 days. Populations are also compared to those in unchallenged mice (*n*=3-4 per group, see Fig 3.20). '+' denotes the mean. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, by *t*-test or general linear models. Where genotype-time interaction was significant, pairwise Tukey post hoc comparisons are shown; otherwise main effects for time or genotype are shown (bold font). † no data.

The circulating immune cell response to infection (Figure 3.32) was characterised by neutrophilia, eosinopaenia and reduction in T-cell counts, which could have resulted either directly from inflammatory stimulation or from glucocorticoid effects. A pronounced genotype-time interaction was evident for the T-cell response, with significant reductions in both CD4⁺ ($p = 0.0001$) and CD8⁺ lymphocytes ($p = 0.0001$) for *Cd97-/-* mice only; the circulating proportions of these cells were similar by day 7. This apparent difference could be an artifact of small sample size at baseline, but in combination with the T-cell effects seen in

the lungs this adds additional support to the possibility of T-cell dysregulation in the *Cd97-/-* mice. No other significant effects of genotype on peripheral blood immune cell proportions were observed.

In view of the change in CD4:CD8 ratio observed at baseline, the differences in relative and absolute T-cell proportions in the BALF and the genotype-time interaction for T cells in blood, CD4:CD8 ratios were next compared in BALF, blood and splenocytes of IAV-challenged mice. $CD4^+$ T cells were the dominant population (ratio > 1) in peripheral tissues (blood, spleen), while $CD8^+$ cells were more abundant in BALF (ratio < 1), consistent with specific migration of effector cells to the site of infection, in both groups of mice. This ratio is expected to be at its lowest around days five to eight following sub-lethal IAV infection in other challenge models.[255] The ratio was significantly increased in all three tissues in *Cd97-/-* mice, indicating a response less skewed towards $CD8^+$ cells (or conversely more skewed towards $CD4^+$ cells). This was most marked in BALF (2.4-fold change, 95% C.I. 1.9 - 3.1-fold, $p = 4 \times 10^{-8}$), with more modest differences in blood (1.2-fold change, 95% C.I. 1.1 - 1.5-fold, $p = 0.004$) and spleen (1.1-fold change, 95% C.I. 1.02 - 1.2-fold, $p = 0.02$).

These data indicate an imbalance in the T-cell response in *Cd97-/-* mice that is present at baseline and in the periphery, but is more marked in the lung in the presence of an inflammatory stimulus that is triggering a Th1 / $CD8^+$ cytotoxic T-cell response. The absolute cell counts suggest that this is due to a failure to adequately mobilise $CD8^+$ T cells rather than an excessive $CD4^+$ T-cell response. Reductions in the pulmonary natural killer cell or myeloid cell responses could also be contributing to the phenotype. Since alveolar macrophages are depleted by IAV infection, the reduction in *Cd97-/-* mice (albeit only apparent in absolute counts) could be a result of, rather than a cause of, recent increased disease severity.[256] There was, however, no evidence of the exaggerated neutrophil influx that has been reported in other disease models.[208]

3.2.6.2 Relative proportions of $CD4^+$ T-cell subsets are maintained in the absence of CD97

To investigate whether the imbalance in $CD4^+$ and $CD8^+$ T cells was being driven by selective expansion or depletion of specific subpopulations, intracel-

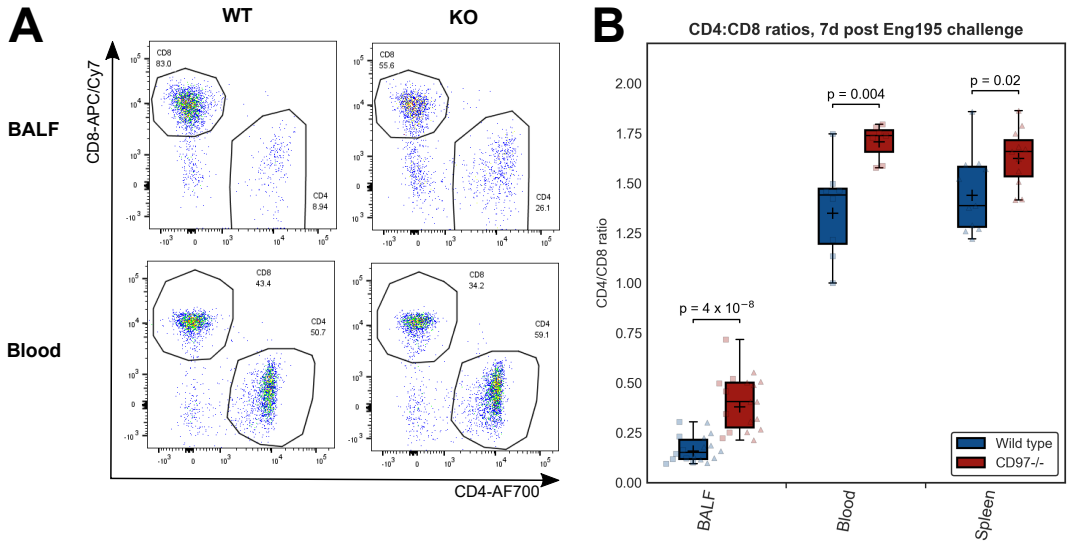

Figure 3.33 – Pulmonary and peripheral CD4/CD8 ratios are increased in *Cd97-/-* mice after IAV challenge. Mice (7-18 per genotype) were challenged with 50 pfu A/Eng/195 and immune cell populations in were determined by flow cytometry after 7 days. A: Representative pseudocolour plots of $CD4^+$ and $CD8^+$ T-cell populations in BALF and blood of *Cd97-/-* versus wild type mice. B: CD4/CD8 ratios are increased in *Cd97-/-* mice in BALF, spleen and blood. '+' denotes the geometric mean. BALF data were pooled from two independent experiments (denoted by symbols), and groups compared by general linear models accounting for study batch effect. Log-transformed data for spleen were analysed by Student's *t*-test, and for blood by Wilcoxon rank-sum test.

lular cytokine staining, in combination with surface marker staining, was used to further characterise the T-cell populations. BALF and splenocytes were harvested seven days after A/Eng/195 infection as above. $CD4^+$ cell subsets were defined as in Table 3.3.

As expected, the predominant T-cell response in BALF was a T helper type 1 (Th1) / $CD8^+$ cytotoxic T-cell response, characterised by high IFNγ expression in both $CD4^+$ and $CD8^+$ cells (Figure 3.34). The anti-inflammatory cytokine IL-10 was also detected in subsets of both cell types, often co-expressed with IFNγ in activated $CD8^+$ cells, which also frequently expressed the activation marker CD25 (IL2 receptor). Th2 cells, which are characterised by IL-4 expression and are more typical of allergic or anti-parasitic responses, and $CD4^+$ CCR6+ T cells, which will include Th17 cells (which promote neutrophilic inflammation), were only rarely detected in BALF of IAV-infected mice.

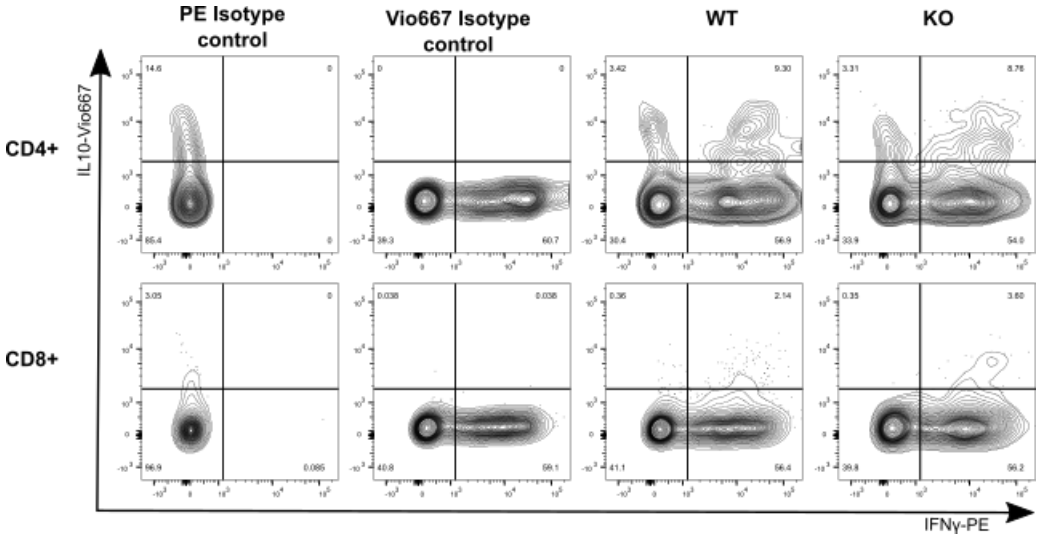

Figure 3.34 – Intracellular cytokine staining in BALF T-cell subsets after A/Eng/195 infection. BALF subpopulations of CD4$^+$ and CD8$^+$ T cells were quantified in *Cd97-/-* and wild type mice by intracellular cytokine staining, 7 days after challenge with 50 pfu A/Eng/195. Cells were cultured with brefeldin A, PMA and ionomycin for 4 hours prior to staining. Flow cytometry contour plots of IFNγ and IL-10 production are shown for representative wild type and *Cd97-/-* mice, stained with anti-cytokine antibodies or isotype controls (shown for a wild type mouse). Isotype controls confirm lack of non-specific antibody binding.

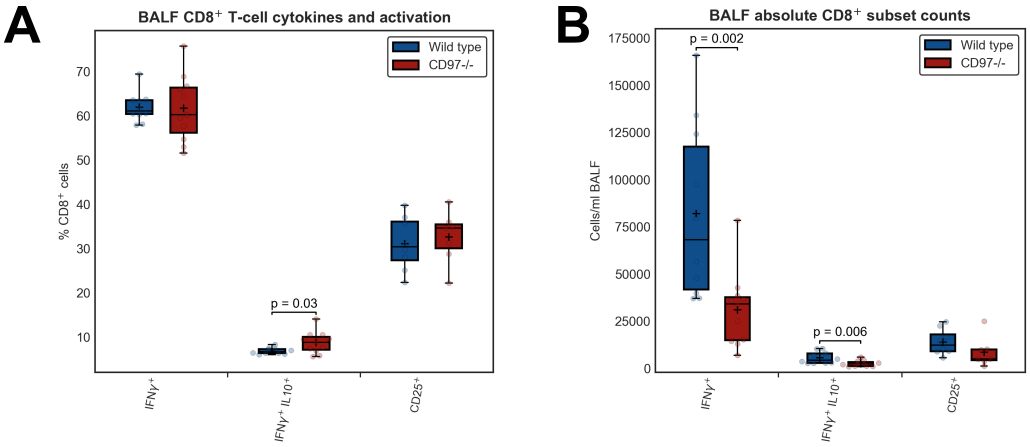

Figure 3.35 – CD8$^+$ T-cell phenotype in BALF after A/Eng/195 infection. Relative proportions (A) and absolute numbers (B) of CD8$^+$ T cells expressing IFNγ and IL-10 (n=10-11 per group) or the activation marker CD25 (n=7-8 per group). '+' denotes the mean. Genotypes have been compared by Student's or Welch's t-test for independent samples.

Cell type	Markers
T_h1	$CD4^+$ $IFN\gamma^+$
T_h2	$CD4^+$ $IFN\gamma^-$ $IL4^+$
Activated $CD4^+$ T cells and regulatory T cells[257,258]	$CD4^+$ $CD25^+$ $IL10^{\pm}$
Other IL10-secreting T cells	$CD4^+$ $CD25^-$ $IL10^+$
Presumed T_h17	$CD4^+$ $CCR6^+$

Table 3.3 – $CD4^+$ subpopulations, defined by cell surface markers and cytokine production.

Cytokine expression patterns and relative proportions of T-cell subsets in BALF were similar between genotypes in both $CD4^+$ and $CD8^+$ cells. Although the absolute number of IFNγ-expressing $CD8^+$ cells was significantly lower in *Cd97-/-* mice, as expected from the lower overall $CD8^+$ T-cell count, the relative proportion of $CD8^+$ cells expressing IFNγ was not different between genotypes (Figure 3.35). Similarly, the proportion expressing CD25 was not different in *Cd97-/-* mice, suggesting that although the number reaching the airway is lower, the activation state of those reaching the site of inflammation is similar in the absence of CD97. Although a slightly higher proportion of $CD8^+$ cells co-express IL-10 in *Cd97-/-* mice, the absolute numbers of these cells are still lower, and so it is uncertain if this difference could be biologically relevant in dampening the inflammatory response.

There was no difference in proportional counts (Figure 3.36), when expressed as a percentage of total $CD4^+$ cells, or in absolute counts (data not shown) between *Cd97-/-* and wild type mice for any $CD4^+$ subset examined. These included Th1 cells, and subsets expressing CD25 and/or IL-10 (which will include regulatory T cells). Th2 and Th17 cells were not present in sufficient quantities for statistical comparison. Although additional markers such as FoxP3 were not used to define regulatory T-cells with greater resolution, there was no evidence of loss of differentiation and proliferation of specific IL-10-secreting subsets, previously reported to depend on the CD97-CD55 interaction in human T cells.[213] Maintenance of relative proportions of $CD4^+$ T cells in the absence of CD97 is consistent with a defect in $CD8^+$ T-cell mobilisation as the driving force of the change in CD4:CD8 ratio.

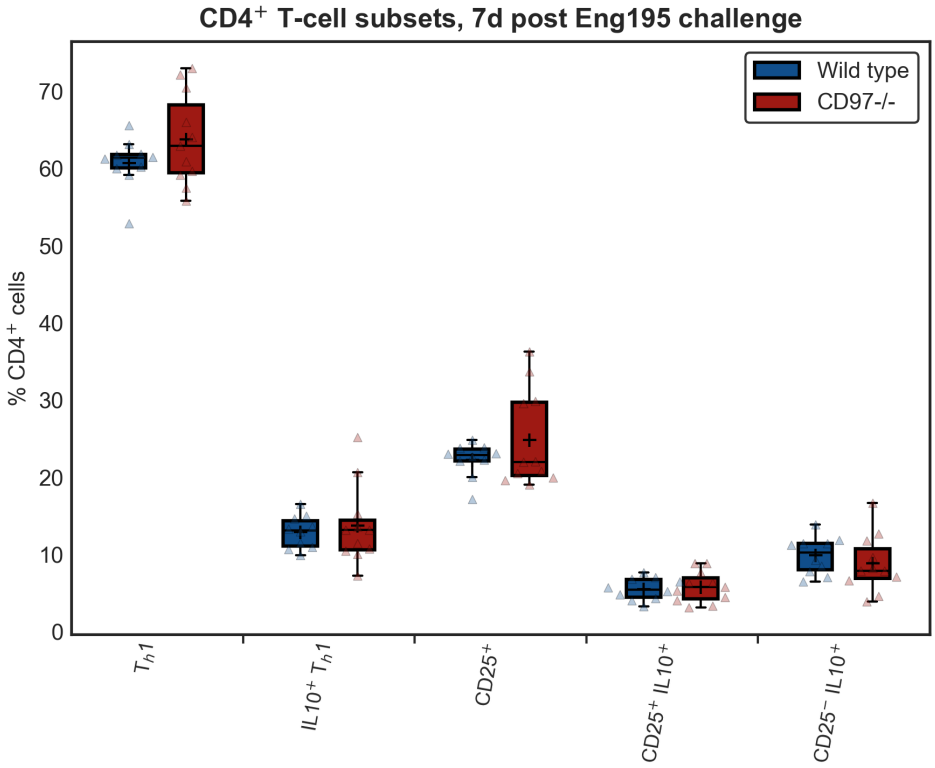

Figure 3.36 – Subsets of CD4⁺ T cells in BALF after A/Eng/195 infection. Cells were harvested and cultured as described above, *n*=10-11 per group. Relative cell populations have been compared by Student's or Welch's t-test, or by Wilcoxon rank-sum test if violating assumptions of normality (Th1⁺IL10⁺ and CD25⁺ subsets). '+' denotes the mean. T-cell subsets shown are not mutually exclusive.

To further investigate whether CD97 deficiency could affect function, rather than number, of Th1 or CD8⁺ T cells, the fluorescence intensity of IFNγ staining after four hours of culture was also compared for BALF leukocytes and splenocytes from mice infected as above. There was no significant difference between genotypes for either Th1 or CD8⁺ T cells in BALF, but there was a small but significant reduction in IFNγ production by CD8⁺ cells from the spleen (Figure 3.37), suggesting a minor reduction in systemic activation of the cytotoxic T-cell response.

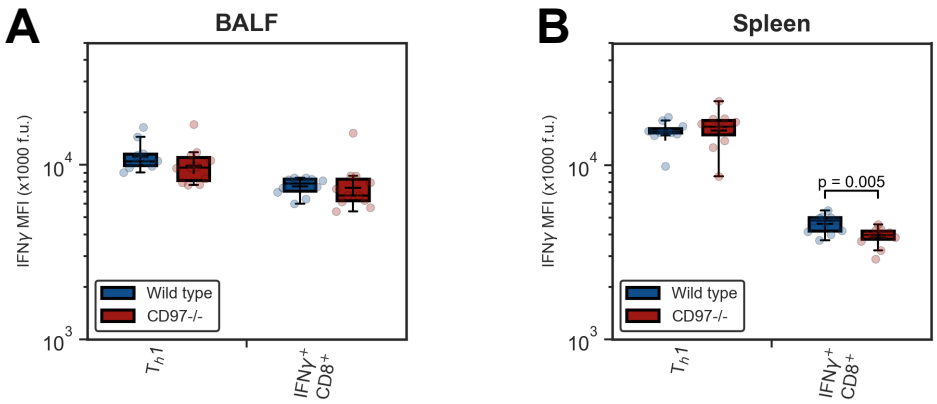

Figure 3.37 – T-cell IFNγ production in BALF and spleen after A/Eng/195 infection. Median fluorescence intensity (MFI), in arbitrary fluorescence units, in T cells from BALF (A) or spleen homogenates (B). Cells were harvested and cultured as described above, *n*=10-11 per group. Comparisons between genotypes were by Wilcoxon rank-sum test. '+' denotes geometric mean.

These findings indicate an inefficient cytotoxic T-cell response in *Cd97-/-* mice after challenge with A/Eng/195, which could be contributing to delayed viral clearance, with no evidence of change in the balance of Th1 or regulatory T-cells driving or limiting this response.

3.2.6.3 T-cell changes are consistent between challenge models.

Changes in immune cell populations could be a cause or a downstream effect of a change in disease severity. If the change in T-cell ratio was a direct result of CD97 deficiency, rather than representing a secondary reaction to altered disease severity, a consistently observable effect would be expected (albeit at varying magnitude) across IAV challenge models. Changes in immune cell populations observed in the A/Eng/195 challenge model were therefore compared with those seen with the alternative H1N1 strain A/Cal/04/09, as well as with differences seen in unchallenged mice or in tissues harvested at early time points after high-dose A/Eng/195 challenge.

Effect sizes for the change in CD4:CD8 T-cell ratio in the BALF *Cd97-/-* mice in experimental replicates of the seven-day A/Eng/195 challenge model, as described above, were highly consistent (Figure 3.38A). The effect size in the

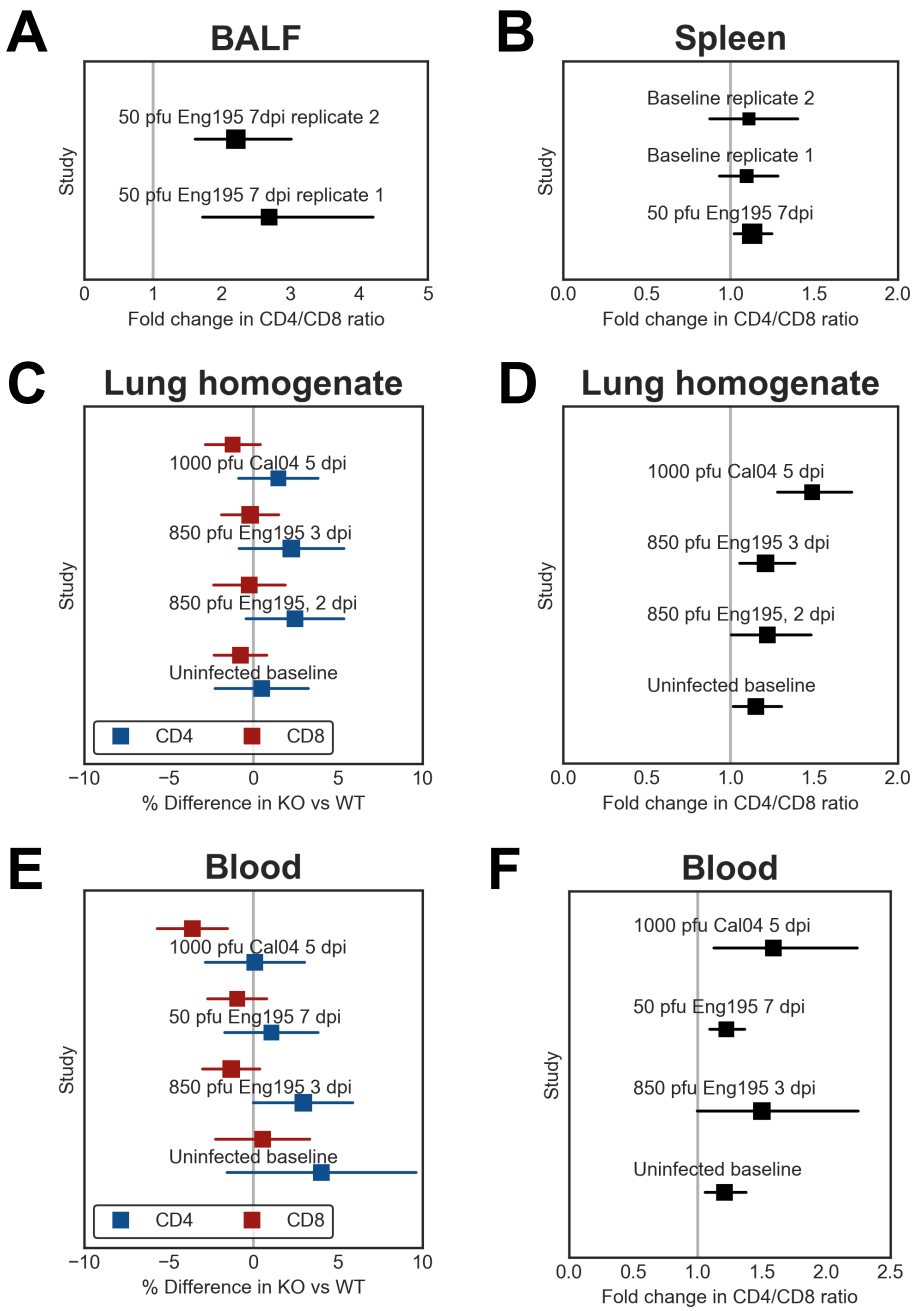

Figure 3.38 – Consistency of effects of CD97 deficiency on T-cell populations and ratios across challenge models. Forest plots for BALF (A), spleen (B), lung homogenate (C and D) and blood (E and F) show effect size estimate (square), either as fold change in ratio in *Cd97-/-* versus wild type mice, or as difference in proportional cell population (as percentage of live $CD45^+$ cells), with 95% C.I. (horizontal line). Area of squares is proportional to *n* (6 - 22 animals total per experiment). Baseline and day 7 data are as in Figures 3.19, 3.20, 3.31, 3.33 and 3.32. dpi: days post infection.

spleen was similar to that observed in uninfected mice, although the latter did not achieve statistical significance (Figure 3.38B). In lung homogenates and peripheral blood, effects of genotype on the populations of $CD4^+$ and $CD8^+$ T cells individually were small and only rarely achieved statistical significance (Figure 3.38C and E); however, the direction of change of the ratio was always the same, and was statistically significant in almost all cases (Figure 3.38D and F). Although no other tissue showed as great an effect size as in BALF of infected mice (in which the $CD8^+$ cells are expected to comprise the greatest proportion under normal conditions), an increased CD4:CD8 ratio was seen both in the lungs and in the periphery, and was present even for the severe A/Cal/04/09 challenge for which no difference in weight loss could be detected.

Changes in other immune cell types were less consistent. Neutrophils tended to increase in lung and BALF with greater disease severity, and were significantly higher in BALF of *Cd97-/-* mice in two of three experiments (Figure 3.39A), and in lung homogenate and blood in only one of three challenge studies each (Figure 3.39B and C). Dysregulated granulocytosis due to CD97 deficiency has been reported by other authors and cannot be excluded completely on the basis of these data.[197,208] However, if the neutrophilia were a direct result of this effect, and if that neutrophilia were driving the observed change in disease severity, greater consistency would be expected, at least in the A/Eng/195 model.

Eosinophil counts (Figure 3.39D-F) tended to decrease with more advanced disease, especially in lung homogenate and blood, but the decreased count seen in BALF after A/Eng/195 challenge was only seen in one experimental replicate and not replicated in any other tissue or challenge model. It is thus unlikely to be relevant to the role of CD97 in the host response. There was similarly little consistency in changes in alveolar macrophage populations in BALF (Figure 3.40A), although there was a trend towards increased counts in *Cd97-/-* mice in lung homogenates, at least at early time points (Figure 3.40B), which could reflect developmental effects or reduced depletion after IAV challenge. Although little change was seen in circulating natural killer (NK) cells, a decrease was observed in lung or BALF in two of three experiments (both with A/Eng/195 challenge) for which data was available (Figure 3.40C and D): while not completely consistent this is also a plausible contributory factor to the observed phenotype.

The change in T-cell ratios, although subtle in some cases, is the most con-

sistent of the observed changes in cell population, across tissues, challenge models and time points. This makes it highly likely to be a direct consequence of CD97 deficiency rather than a secondary response to changes in disease severity. It may not be the only relevant change to the host response in this model, and dysregulation of neutrophils or NK cells could also contribute to the immune phenotype, but this altered T-cell response is mechanistically plausible as a modulator of influenza severity.

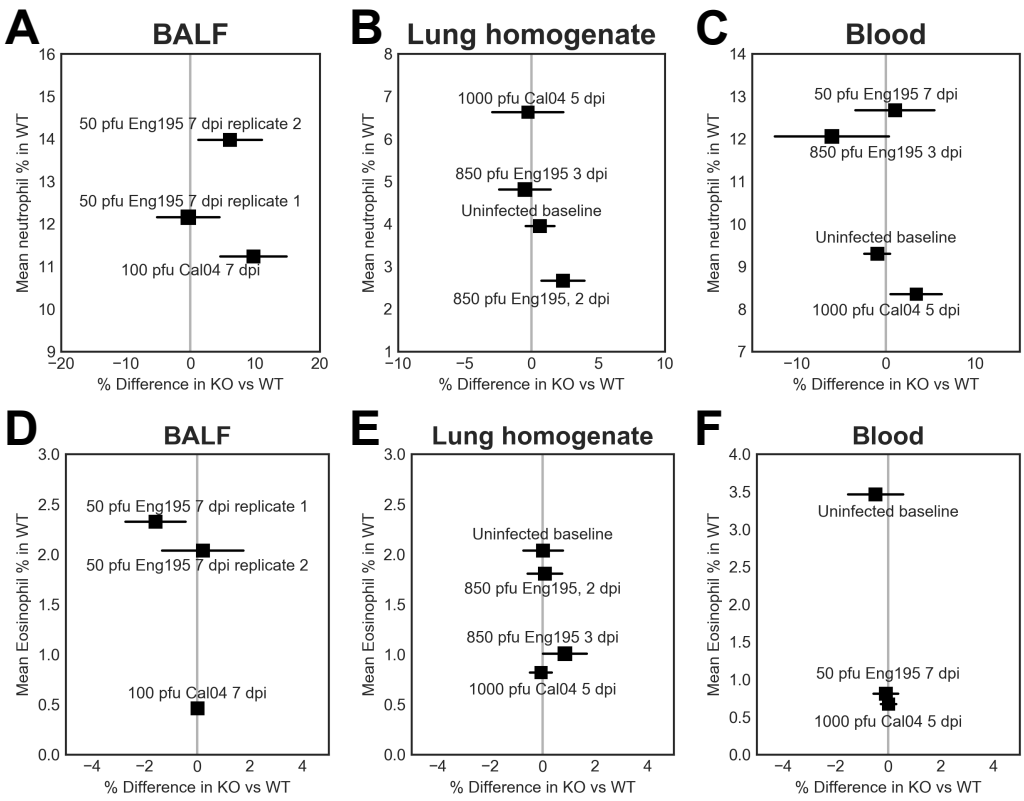

Figure 3.39 – Consistency of effects of CD97 deficiency on granulocyte populations across challenge models. Differences in neutrophil (A-C) and eosinophil (D-F) populations were evaluated across challenge models with different IAV strains, doses and time points. Forest plots for BALF (A and D), lung homogenate (B and E) and blood (C and F) show effect size estimate (square) for the difference in proportional cell population (as percentage of live CD45$^+$ cells), with 95% C.I. (horizontal line). Area of squares is proportional to *n* (13 - 19 animals total per experiment). The y-axis shows the mean cell proportion in wild type mice. Baseline and day 7 data are as in Figures 3.19, 3.20, 3.31 and 3.32. dpi: days post infection.

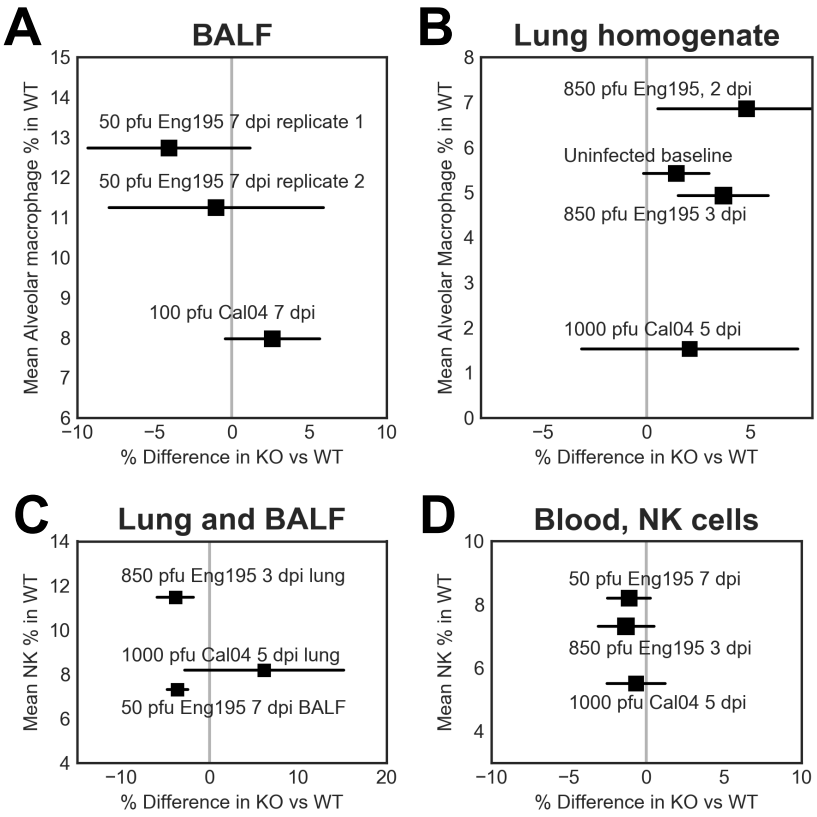

Figure 3.40 – Consistency of effects of CD97 deficiency on alveolar macrophage and NK-cell populations across challenge models. Forest plots for differences in alveolar macrophage (A and B) and NK-cell (C and D) populations show effect size estimate (square) for the difference in proportional cell population (as percentage of live CD45$^+$ cells), with 95% C.I. (horizontal line). Area of squares is proportional to *n* (13 - 19 animals total per experiment). The y-axis shows the mean cell proportion in wild type mice. Baseline and day 7 data are as in Figures 3.19, 3.20, 3.31 and 3.32. dpi: days post infection.

3.2.7 Lack of evidence of effects of CD97 deficiency on complement activation

CD55 (DAF), the major ligand for CD97, is a key regulator of complement activation (Figure 3.41), reducing activation and amplification of the classical, lectin-mediated and alternative complement pathways via degradation of the C3 convertases at the cell surface. Inhibition of CD55 function or reduction in expression would be expected to enhance complement activation, leading to increased production of activated complement components such as the pro-

inflammatory mediators C3a and C5a (anaphylotoxins), although paradoxically, reduced C3a concentrations are found in *Cd55-/-* mice in the later stages of IAV challenge models.[149] These soluble mediators are chemotactic for myeloid cells and enhance the secretion of pro-inflammatory cytokines such as TNFα and IL-6, and could exacerbate influenza severity.[259] Complement has both protective and harmful effects in influenza, promoting viral clearance but also exacerbating some aspects of immunopathology.[131–133,155,259] Although CD97 is not expected to affect CD55 function directly, as it binds the opposite face of the receptor from the C3 convertase binding site, reduced expression, as observed in granulocytes of uninfected *Cd97-/-* mice (Figure 3.16), could influence complement activation and disease severity.

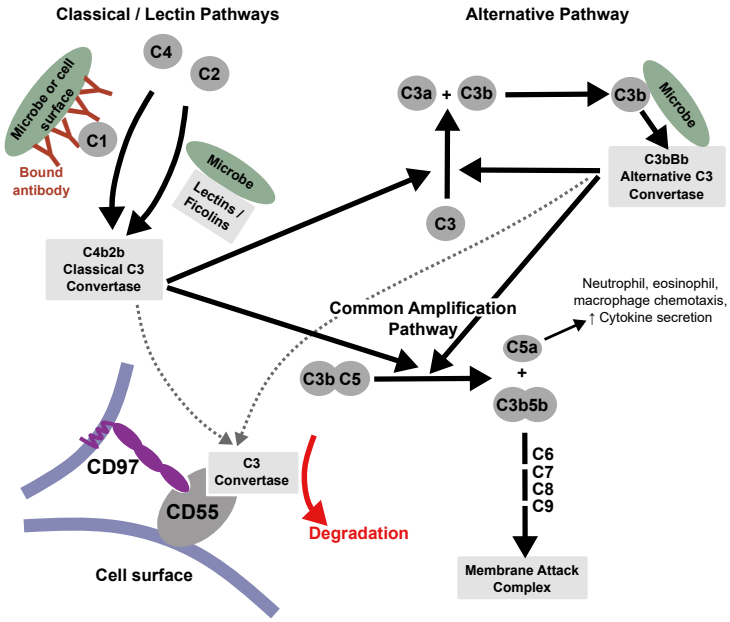

Figure 3.41 – Simplified schematic of complement pathways and the role of CD55 in complement regulation. Cell-bound CD55 promotes degradation of both the classical and alternative C3 convertases, reducing activation of both arms of the cascade.

To assess first whether the reduction in neutrophil CD55 expression in unchallenged mice is also present in IAV-infected mice, neutrophil CD55 expression was measured by flow cytometry. This showed no significant difference between genotypes in peripheral blood seven days after A/Eng/195 challenge (Figure 3.42A), although there was a small reduction approaching statistical significance after 1000 pfu A/Cal/04/09 challenge ($p = 0.06$; Figure 3.42B). No such

reduction was seen in BALF with either virus strain (Figure 3.42C and D): for A/Eng/195 there was a significant study:genotype interaction (p = 0.01) with a significant *increase* in *Cd97-/-* mice in one replicate only (p = 0.004 on Tukey's post hoc comparison). In whole lung homogenates after 1000 pfu A/Cal/04/09 challenge (Figure 3.42E), neutrophil expression showed a similar pattern to that in peripheral blood neutrophils (p = 0.06). The lack of reduction in BALF suggests that, if CD97 deficiency does reduce CD55 expression, this effect is outweighed by the effects of neutrophil activation on CD55 expression at the site of inflammation.[242]

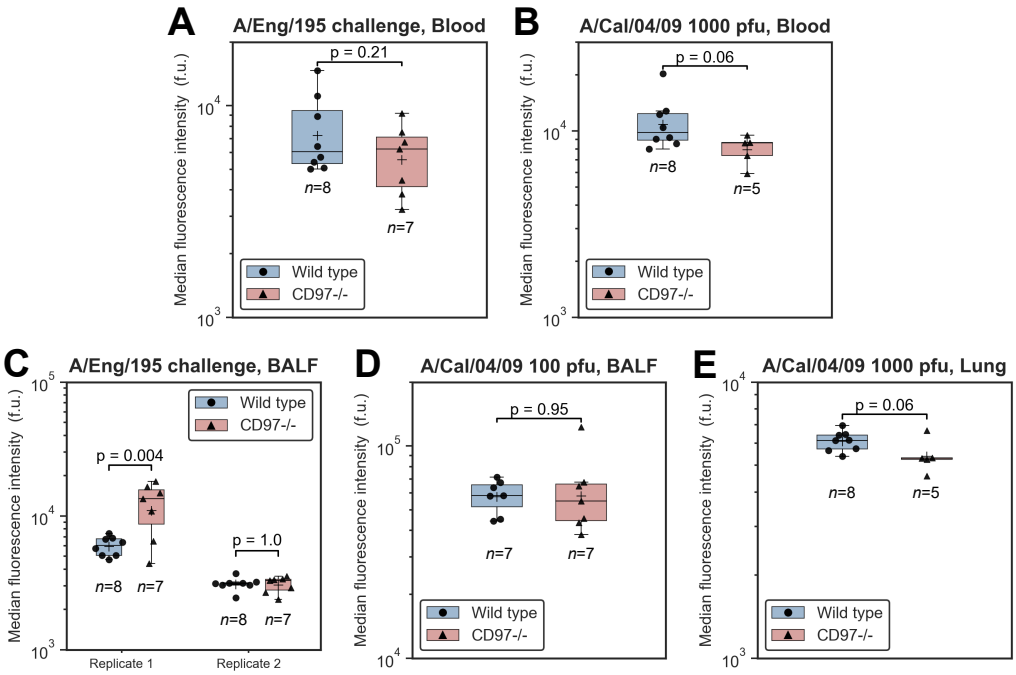

Figure 3.42 – Neutrophil CD55 expression after IAV infection. CD55 expression on neutrophils was evaluated by flow cytometry 7 days after challenge with 50 pfu A/Eng/195 (A and C), 5 days after 1000 pfu A/Cal/04/09 challenge (B, E), or 7 days after 100 pfu A/Cal/04/09 challenge (D), in peripheral blood (A and B), BALF (C and D) or lung homogenate. Median fluorescence intensity (in arbitrary fluorescence units) was compared between genotypes by Student's t-test on log-transformed data (A, B, D, E), or by general linear model to control for study batch effects, with Tukey post hoc pairwise comparisons (C). The '+' symbol denotes geometric mean. Note that fluorescence intensities are not directly comparable between sample types due to differences in antibody panels and acquisition voltages used.

The lack of clear differences in CD55 expression in neutrophils in the BALF of infected mice suggests that CD97 deficiency is unlikely to have a substantial effect on pulmonary complement activation via modulation of granulocyte CD55 activity during IAV infection. Although granulocytes were the main immune cell population in which a baseline difference in CD55 expression was seen, it is of course possible that complement activity could still be affected by differences in expression in other cell types, or via CD55-independent mechanisms. To investigate this possibility, concentrations of C5a, a major product of complement activation, were measured by ELISA in BALF, plasma and lung homogenate after A/Eng/195 or A/Cal/04/09 infection (Figure 3.43). No significant differences were observed between genotypes. These data do not provide support for dysregulation of systemic or pulmonary complement activation, at least in the advanced stages of disease examined here, as a major effect of CD97 deficiency on influenza pathophysiology.

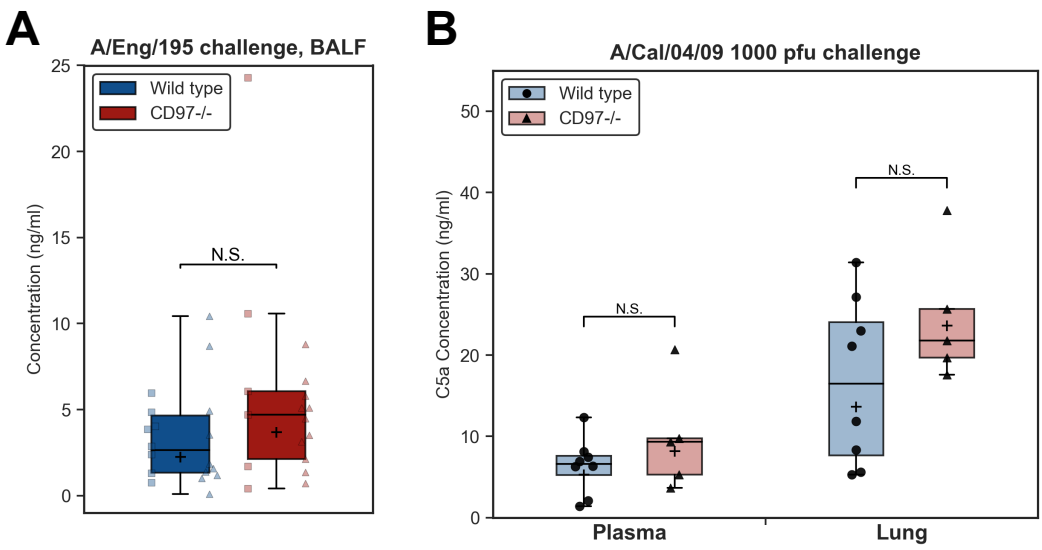

Figure 3.43 – C5a in plasma and lung after IAV infection. Concentration of activated C5 was determined by ELISA. A: C5a in BALF, 7 days after challenge with 50 pfu A/Eng/195. Data are pooled from two experimental replicates (denoted by symbols), total *n*=17-18 per genotype. B: C5a in lung homogenate and plasma, 5 days after challenge with 1000 pfu A/Cal/04/09 (*n*=5-8 per group). The '+' symbol denotes geometric mean. Differences between genotypes have been analysed by general linear model (A) Student's t-test (B) or Welch's t-test (C) on log-transformed data.

3.2.8 *Cd97-/-* mice have an exaggerated pulmonary IFNγ response to IAV infection

CD97 has been reported to suppress NF-κB activation and TNFα secretion in response to LPS in human macrophages, by a mechanism that involves enhanced expression of the inhibitory transcription factor PPAR-γ.[214] NF-κB is also activated in response to IAV infection, in part via activation of pattern recognition receptor such as TLR7 and RIG-I, leading to the production of pro-inflammatory cytokines that will play a key role in orchestrating the ensuing immune response.[119] Modulation of cytokine responses could thus be a plausible mechanism of action of modulation of disease severity by CD97.

First, cells derived from wild type and *Cd97-/-* mice were used to investigate whether previously reported effects in human monocyte-derived macrophages[214] could be replicated in murine bone marrow-derived macrophages (BMDMs). Bone marrow-derived macrophages were differentiated as described above (see Figure 3.27) and stimulated with 100 ng/µl LPS, and cytokine expression was evaluated by qRT-PCR after 4 and 24 hours (Figure 3.44). Expression of pro-inflammatory cytokine genes *Tnf*, *Il6* and *Il12b* (encoding IL12-p40, a subunit of both T-cell stimulatory factor IL-12 and IL-23) was highest at the four-hour time point ($p = 2 \times 10^{9}$ for time effect on mixed linear models), but there was no effect of genotype on expression of any of the cytokines measured.

Cytokine responses were next evaluated in BMDMs after infection with IAV strain A/WSN/33 (Figure 3.45). Expression of *Tnf*, *Il6* and interferon-responsive cytokine Cxcl10 increased between 4 and 24 hours post infection, consistent with the later and more sustained response with IAV compared to LPS in human monocyte-derived macrophages.[252] Expression of the anti-inflammatory cytokine gene *Il10* was not consistently detectable within the limits of detection of the assay. As with LPS, there was no significant difference in expression of pro-inflammatory cytokines between genotypes.

CD97-mediated modulation of cytokine expression could not be replicated in this model. This could be due to species differences, or differences between macrophage subtypes. These results do, however, suggest that it is unlikely that the observed differences in immune phenotype in *Cd97-/-* mice are mediated by differences in macrophage cytokine expression in the early innate im-

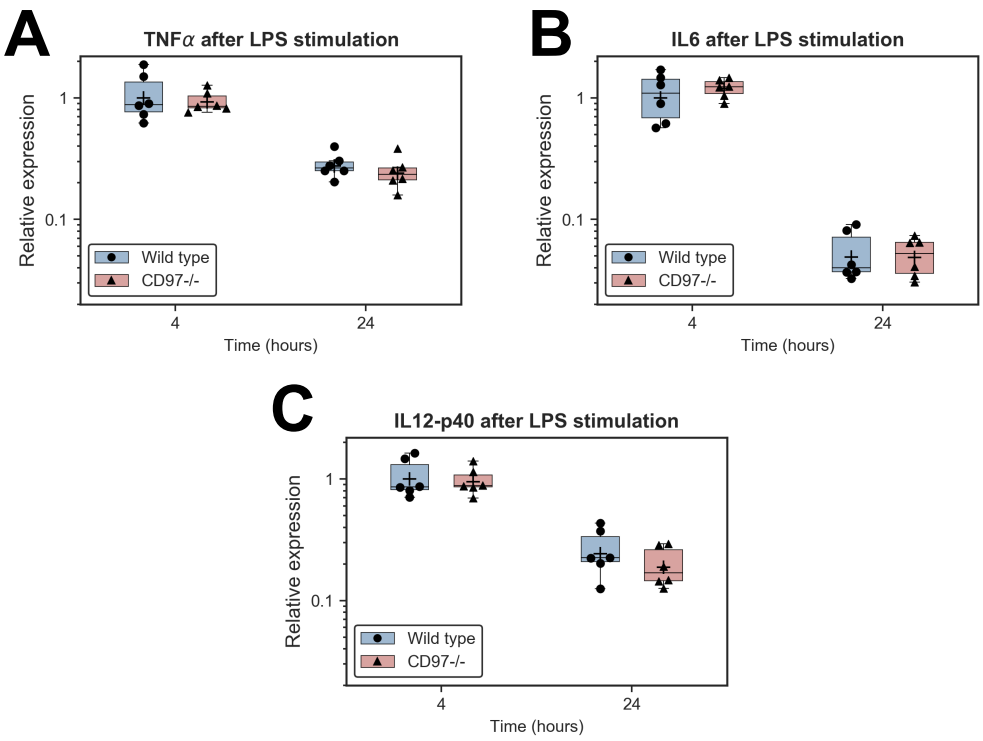

Figure 3.44 – BMDM cytokine expression after LPS stimulation. BMDMs from wild type and *Cd97-/-* mice (n = 6 per group) were stimulated with 100 ng/µl LPS, and cytokine gene expression was evaluated by qRT-PCR after 4 and 24 hours. CT values have been normalised to the mean of 2 endogenous controls (*Hprt1* and *B2m*), and relative expression normalised to the mean for wild type mice at 4 hours ($2^{-\Delta\Delta CT}$). ΔCT values were compared between genotypes by mixed linear models. Statistical significance of differences between time points is not shown.

mune response. Cytokine secretion *in vivo* is, however, not solely determined by innate macrophage responses: cytokines can be secreted by a wide range of cells including lymphoid, epithelial and endothelial cells, and regulation of cytokine responses involves multiple interactions between these cell types, which cannot be fully modelled *in vitro*. The overall pulmonary cytokine response was therefore assessed by measuring cytokine concentrations in BALF seven days after A/Eng/195 challenge.

There was no significant difference in BALF concentration of TNFα or IFNα between wild type and *Cd97-/-* mice (Figure 3.46A/B). Interferon-γ concentration (Figure 3.46C), however, was consistently and significantly increased in the

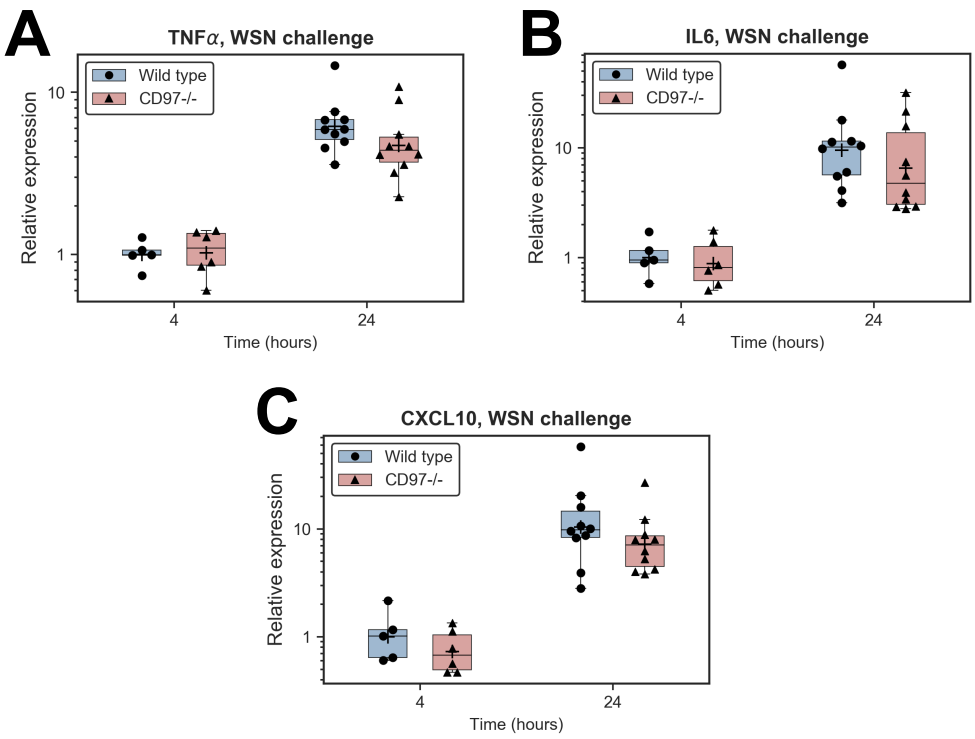

Figure 3.45 – BMDM cytokine expression after IAV infection. BMDMs from wild type and *Cd97-/-* mice (n = 5-10 per group) were infected with IAV strain A/WSN/33 at MOI 5, and cytokine gene expression was evaluated by qRT-PCR after 4 and 24 hours. Raw CT values have been normalised to the mean of 2 endogenous controls (*Gapdh* and *B2m*), and ΔCT values normalised to the mean for wild type mice at 4 hours. ΔCT values were compared between genotypes by general linear models. Statistical significance of differences between time points is not shown.

BALF of *Cd97-/-* mice (1.5-fold change, 95% C.I. 1.2 - 1.9-fold, p = 0.0009), despite a small reduction in overall BALF protein concentration (see Figure 3.23C).

Interferon-α is secreted by many cell types in response to viral infection. The lack of difference in airway concentrations of this cytokine is consistent with our earlier observations that viral restriction and IFNα production are not different between *Cd97-/-* and wild type mice *in vitro* (section 3.2.5.2). Lack of difference in TNFα, which is secreted predominantly (but not exclusively) by macrophages, is also consistent with *in vitro* models. Interferon-γ, in contrast, is mainly secreted by innate and adaptive lymphoid cells including natural killer cells, natural killer T cells, gamma-delta T cells, Th1 cells and cytotoxic $CD8^+$ lymphocytes.

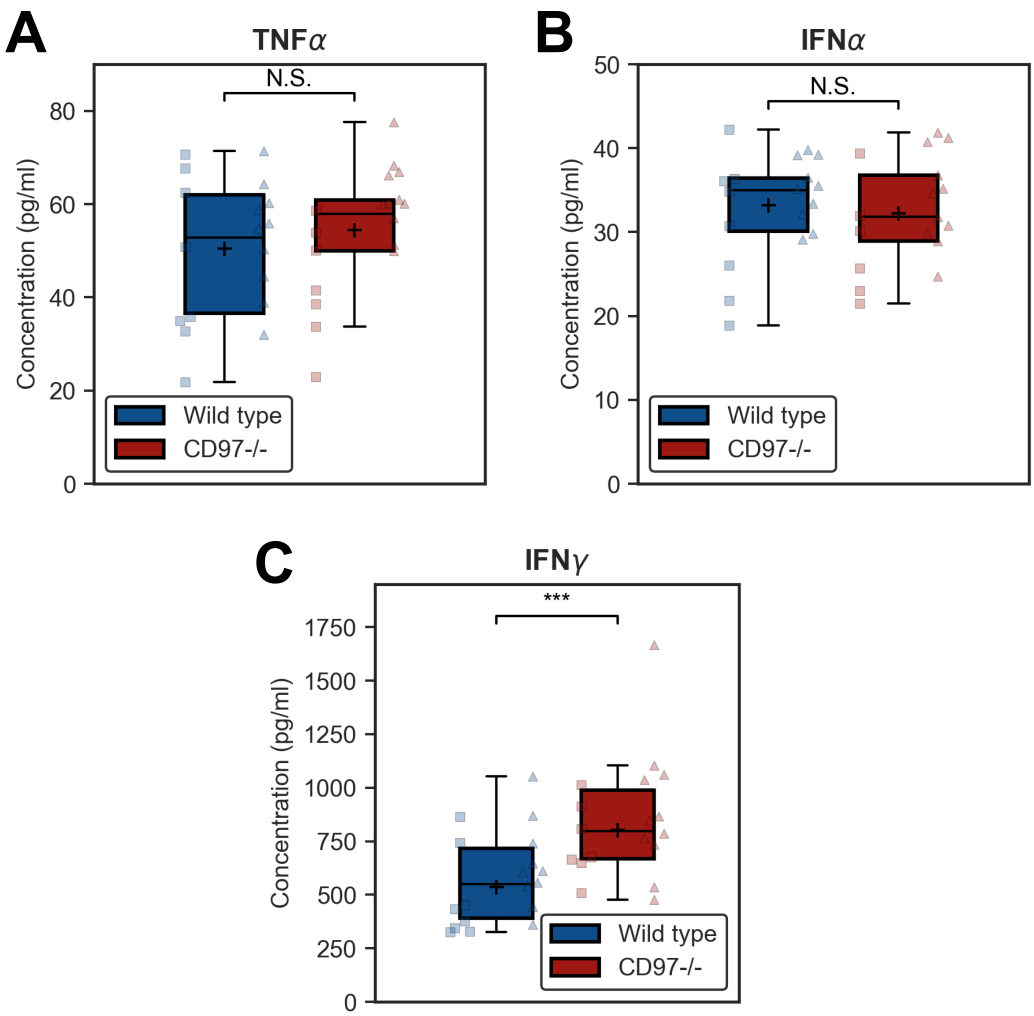

Figure 3.46 – BALF cytokines after IAV challenge. Mice (7-18 per genotype per experimental replicate) were challenged with 50 pfu A/Eng/195 and culled after 7 days. Concentrations of TNFα (A), IFNα (B), and IFNγ (C) in BALF were determined by ELISA. The '+' symbol shows the mean (A, B), or geometric mean where log-transformation was necessary for analysis (C). BALF data were pooled from two independent experiments (denoted by symbols), and groups compared by general linear model accounting for study batch effect. *** $p<0.001$

The selective increase in IFNγ in the BALF of *Cd97-/-* mice could either represent an appropriate response to the increase in viral titre (section 3.2.5.1), or represent dysfunction of this branch of the immune response, consistent with observations of an imbalance in the T-cell and possibly natural killer cell populations.

The source of the increased IFNγ in this context is not clear. $CD8^+$ T cells are a major source of this cytokine in BALF after IAV infection.[260] The natural killer cell and $CD8^+$ populations were reduced in the BALF of *Cd97-/-* mice, and although the relative population of Th1 cells were increased, no difference in absolute numbers was detected. No increase in IFNγ production by Th1 or $CD8^+$ T cells was detected by intracellular cytokine staining, but as culture and stimulation is required to measure cytokines by this technique, it is more an assessment of the cells' capacity for secretion rather than a measurement of actual *in vivo* secretion. Increased secretion from innate lymphoid populations such as natural killer T cells, which were not specifically assessed, is a possibility. Alternatively, there could be a difference in behaviour between luminal and interstitial T cells: the latter have been shown to have greater secretory activity in influenza.[260] Increased secretion could result from an alteration in the balance of stimulatory (e.g. IL-12, IL-18 and IL-2) and inhibitory (e.g. IL-6, TGFβ) cytokines, or in T cells could result from enhanced interaction with antigen-presenting cells.[261]

3.3 Discussion

Our results show that while the genetic association between a single variant in *CD97* and influenza severity in humans is not as robust as some other associations for influenza, this receptor does modulate the host response to infection with influenza A virus in a manner which could plausibly influence the risk of developing critical illness. In a mouse model, CD97 deficiency was associated with a modest change in clinical phenotype which was highly context-dependent, and may have involved elements of both increased and decreased severity for different aspects of the response. The primary cellular phenotype however, an imbalance in the T-cell response with an apparent reduced efficiency of the cytotoxic $CD8^+$ T-cell response in the lung, although subtle, was highly consistent, and was accompanied by delayed viral clearance, increased IFNγ production, and variable other cellular deficits such as reduced natural killer cell infiltration.

3.3.1 Is modulation of T-cell homeostasis responsible for the observed phenotype?

After the initial innate immune response to IAV infection, mediated primarily by the type I interferon and natural killer / natural killer T-cell responses, T-cell immunity is one of the principal means of eliminating the virus in a naïve host, or in subsequent re-infection if the humoral response has been insufficient for virus neutralisation. At a basic level, this involves proliferation and activation of type I $CD4^+$ T helper (Th1) cells and cytotoxic $CD8^+$ T cells, with Th1 help, which migrate to the lungs where cytotoxic T cells are instrumental in clearing infected epithelial cells.[262] The full spectrum of the T-cell response is however more complex, and other sub-populations have specific roles in activation or regulation of the immune response. For example, natural killer T cells and mucosal-associated invariant T cells have innate-like protective functions, with direct cytokine-mediated activation earlier in the course of the disease than the adaptive T-cell response.[134,135] Similarly, gamma-delta T cells (which can respond to antigens without the need for presentation on the major histocompatibility complex) are part of the innate-like response, responding with both direct and indirect antiviral effects via cytolysis of infected cells and secretion of cytokines such as IFNγ respectively, and can also help to direct the adaptive response.[136] Regulatory T cells limit aspects of the innate and adaptive immune response such as monocyte infiltration, activation of $CD8^+$ T cells and migration of $CD4^+$ T cells, potentially preventing the damaging effects of excessive immune activation, but also seem to be necessary for efficient viral clearance in some murine models.[263–266] T follicular helper cells are necessary for an efficient B-cell antibody response, while Th17 cells secrete IL-17, a cytokine that promotes neutrophil responses and efficient clearance of bacterial infection, but which can be detrimental in influenza.[267,268]

Whether the T-cell response, or aspects thereof, is indispensable for recovery from IAV infection in murine models depends on the subpopulations of T cells assessed, the infection model used, and the competence of the B-cell response. *Rag1-/-* or *Rag2-/-* mice, which lack both B cells and T cells, show similar weight loss to wild type mice in early stages but fail to recover after H1N1 A/PR/8/34 infection at a dose that is sub-lethal in wild type mice, and treatment with natural IgM from naïve mice delays but does not prevent death, suggesting that recov-

ery requires T-cell or specific B-cell responses.[131,269] In SCID mice, which lack T cells, B cells and NK cells, virus-neutralising IgG treatment has been reported to rescue an otherwise lethal phenotype after low-dose A/PR/8/34 infection.[270] Mice depleted of $CD4^+$ T cells (using monoclonal antibodies) can still clear the A/PR/8/34-derived strain A/X-31(H3N2), provided that the perforin or Fas pathways for $CD8^+$ T-cell cytotoxicity are intact, while $CD8^+$ T-cell depletion cause a large increase in susceptibility to this strain only in mice lacking mature B cells.[271,272] Pan-T-cell deficient *Foxn1^{nu}* (nude) mice can recover from low-dose A/PR/8/34 infection, most likely due to a T-cell-independent B-cell response.[273]. Mice with a specific deficiency in mucosal-associated invariant T cells show enhanced susceptibility to A/PR/8/34 but not to the less virulent H3N2 strain A/X-31.[135] These results indicate that there is a degree of redundancy in the immune response to IAV, and T cells are not required for viral clearance in all circumstances.

If even a complete absence of T-cell immunity does not preclude recovery, the impact of a partial T-cell deficiency or dysfunction is likely to have effects that are more subtle and highly dependent on context such as challenge strain, dose, and other aspects of host immunocompetence. It is thus entirely plausible that a partial reduction in the efficiency of the $CD8^+$ response, as observed in the *Cd97-/-* mice, could be sufficient to cause a measurable increase in disease severity after challenge with one IAV strain (A/Eng/195) but not in another 2009 pandemic-associated H1N1 strain (A/Cal/04/09). Cytotoxic T lymphocytes can moreover contribute to lung immunopathology after IAV infection, and in some circumstances (especially at higher viral challenge doses) can exacerbate disease severity and mortality in mouse models.[274] Attenuation of such immunopathology in *Cd97-/-* mice could explain the later convergence in the weight loss phenotype and reduction in other measures of lung injury severity by day seven in the A/Eng/195 model.

The observation of altered T-cell ratios in *Cd97-/-* mice is not in itself sufficient to demonstrate a causative link between CD97 deficiency, the T-cell response and disease severity. A number of additional observations, however, suggest that these changes are neither secondary nor incidental. While a relative reduction in the $CD8^+$ cytotoxic T-cell response would be expected with a less severe viral challenge, the consistency of this observation across models and time points, especially in the face of similar or greater viral titres, indicates that

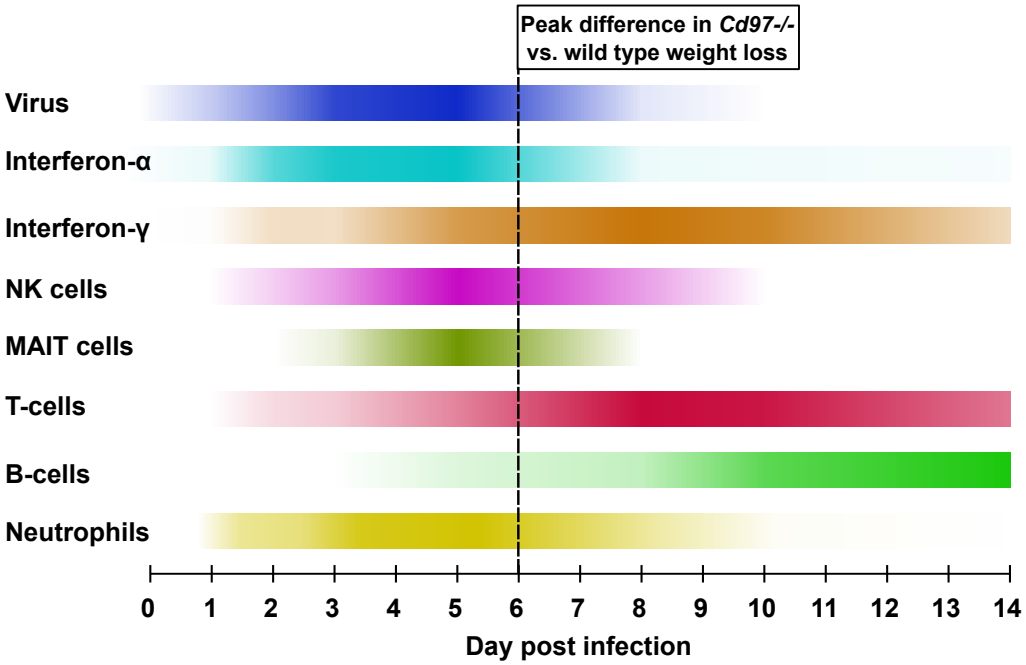

Figure 3.47 – Schematic of the time course of the pulmonary immune response in mice after sub-lethal IAV infection. The approximate time course of the interferon and cellular immune responses in lung following sub-lethal infection with IAV strains A/PR/8/34 or A/X-31 is shown, with bar opacity reflecting the degree of the response between baseline (uninfected) and maximal levels (not to scale). These time courses are based on previously published transcriptomic and flow cytometry data, in models of comparable severity to those used in our studies.[135,255,275]

while the magnitude of effect may vary with the strength of the immune stimulus, the effect is likely to be a primary effect of CD97 deficiency rather than a physiological response to altered disease severity. Increased viral titres are also consistent with the hypothesis that reduced efficiency of the $CD8^+$ cytotoxic T-cell response is sufficient to impair viral clearance, as expected with the known role of cytotoxic T cells in the immune response to IAV. Finally, although the change in weight loss is transient and context-dependent, the timing of the effect in the A/Eng/195 model is consistent with a T-cell effect: the peak difference in weights coincides with the expected timing (based on published data from similar models) of the most rapid rise in the T-cell response and corresponding rapid fall in viral titre, around the time of the nadir in the CD4:CD8 ratio.[255] At this time, the type I interferon, natural killer cell and MAIT responses are expected to be declining, while the B-cell response peaks later following infection (Figure 3.47).

No difference in phenotype was seen at these earlier or later, innate immune or B-cell-dominant phases. Together these results indicate that modulation of the T-cell response by CD97 deficiency is likely to modulate disease severity by reducing the efficiency of viral clearance, although abrogation of the phenotype in T-cell deficient *Cd97-/-* mice would be required to confirm this.

It is less clear whether the observed increase in IFNγ at day seven post infection could have a causative role in the association between CD97 deficiency and disease severity. While IFNγ has important antiviral actions, both direct for some viruses (for example by inhibiting viral gene expression[276]) and indirect (for example, by enhancing antigen presentation in macrophages and promoting immune cell migration), its activity can be detrimental in some cases, and can exacerbate inflammatory lung injury. In mice lacking the IFNγ receptor, lung viral titres are counterintuitively lower, and virus spread within the lungs more restricted, after infection with IAV strain A/WSN/33, corresponding to reduced clinical signs and inflammatory cytokine production but more severe neutrophil infiltration and lung histopathology scores.[277] Inhibition of protective type 2 innate lymphoid cell responses contributes to the detrimental effects of IFNγ.[278] The increased BALF IFNγ in *Cd97-/-* mice could therefore be contributing to impaired viral clearance and an excessive inflammatory response, but could conversely be an entirely appropriate secondary response to the increased residual viral load rather than being a direct consequence of CD97 deficiency. Reduced NK-cell infiltration could conceptually be a cause of impaired viral clearance, but while the impact of NK cells on A/Eng/195 infection has not been evaluated specifically, NK-cell depletion has been shown to have little effect on weight loss after infection with other IAV strains.[149]

3.3.2 Mechanisms of the effect of *Cd97-/-* on the T-cell response to influenza

A number of biological processes could affect the CD4:CD8 balance in the T-cell response. Differentiation of immature $CD4^+$ $CD8^+$ double-positive T lymphocytes into single-positive cells occurs in the positive selection process in the thymus, where a strong interaction with MHC class I leads to downregulation of CD4 expression to yield a $CD8^+$ single-positive cell, or conversely a strong inter-

action with MHC class II will produce a $CD4^+$ lymphocyte.[279] Positively-selected T cells must also survive negative selection, whereby cells reacting to self antigens are removed (or occasionally differentiate into regulatory T cells) before release from the thymus. There is precedent for a role of adhesion G-protein-coupled receptors in regulation of T-cell development in mouse models, as the closely-related receptor F4/80 (ADGRE1) is required for efficient generation of $CD8^+$ regulatory T cells.[21] CD97 is unlikely to be involved at this early developmental stage given the lack of difference seen in lymphoid tissue (spleen) of uninfected mice, although a small difference was seen in peripheral blood. Furthermore, although CD97, which is absent in earlier double-negative precursor cells, starts to be expressed in these double-positive cells, CD55 is not expressed in the thymic epithelial cells with which they interact (see Figure 3.9), so CD97 could only be involved in this process if an alternative receptor-ligand interaction were involved.

During development of the adaptive immune response, differentiated naïve T cells are activated when their T-cell receptor recognises the target antigen presented via MHC on antigen-presenting cells such as dendritic cells, primarily in local draining lymph nodes but also locally within the lung. This is not sufficient in itself for full activation, and additional costimulatory signals are required, such as the interaction between CD28 on T cells and CD80 or CD86 on antigen-presenting cells. This combination of signals leads to clonal expansion and acquisition of effector functions, such as expression of effector cytokines and cytotoxic molecules.[262] This is one of the most likely steps in the T-cell response in which CD97-CD55 signalling could be implicated.

There is evidence in human T cells that CD97-CD55 signalling acts as an alternative costimulatory signal. Stimulation with plate-bound anti-CD55 antibody or CD97-Fc fusion protein, together with anti-CD3 antibody as the primary stimulus, results in enhanced proliferation and cytokine secretion in naive $CD4^+$ T cells (but not in committed Th1 effector cells), with a similar magnitude of effect to CD28 costimulation but a relatively greater stimulation of IL-10 release.[148] These interactions can also be bidirectional. Human monocytes upregulate CD55 in response to IFNγ stimulation (contrasting with downregulation on differentiation to dendritic cells induced by IL4/GM-CSF), and antibody pre-treatment of either CD55 on these antigen-presenting cells, or of the EGF domains of CD97 on T cells, inhibits proliferation and IFNγ secretion on co-culture.[211] It

is not yet known if this signalling mechanism also operates in murine T cells. Both $CD4^+$ and $CD8^+$ subsets express similar levels of CD97 and CD55, but if this signal has a greater effect in $CD8^+$ than $CD4^+$ T-cells (for example due to differences in downstream signalling pathways or differences in expression of alternative costimulatory molecules with potential for functional redundancy), a relative reduction in $CD8^+$ T-cell proliferation and consequent increase in the CD4:CD8 ratio would be predicted with CD97 deficiency.

The role of complement in T-cell responses has received some attention in the context of CD55-mediated effects. $CD4^+$ and $CD8^+$ T cells in *Cd55-/-* mice show complement-dependent hyper-responsiveness to antigen restimulation, showing enhanced proliferation and IFNγ secretion, but lower secretion of IL-10.[147,280] Complement component C3 has also been shown to be required for optimal $CD8^+$ T-cell expansion in response to lymphocytic choriomeningitis virus infection in mice.[281] The CD55-CD97 interaction, however, involves the opposite surface of the CD55 molecule to the complement binding region, and does not affect CD55-mediated complement inhibition, so effects of CD55 on T-cell function include both complement-mediated and complement-independent mechanisms, which seem to have opposing effects.[148,211] In our murine influenza model, the increase in BALF IFNγ in *Cd97-/-* mice is contrary to expectations with reduced T-cell costimulation; although no evidence of a global increase in complement activation (as measured by C5a concentration) was found in BALF or plasma, an increase in local complement signalling could have resulted in the observed increase. There was no evidence here of a reduction in the proportion of either $CD4^+$ or $CD8^+$ T cells producing IL-10, as would have been expected from the loss of CD97 costimulation based on data in humans, although as for IFNγ, interstitial lymphocytes may behave differently from those in the airway due to closer contact with antigen-presenting cells.

In this mouse model, while CD97 expression was widespread among immune cells, little CD55 was expressed in monocytes, macrophages and dendritic cells (Figures 3.7, 3.9 and 3.16). Although upregulation of CD55 expression on stimulation is a possibility, this expression pattern suggests that CD97-CD55 interactions between conventional antigen-presenting cells and T cells are more likely to involve CD97 on the antigen presenting cell and CD55 on the T cell, rather than the reverse direction of signalling. CD55 was expressed in neutrophils, however, which can present IAV antigen to $CD8^+$ T cells in the lung, enhanc-

ing cytokine secretion, so bidirectional signalling is possible in this scenario.[282] Although CD55 is a membrane-anchored protein with no intracellular signalling domain, CD97 engagement of CD55 on T cells (as well as the converse interaction between CD55 on antigen presenting cells and CD97 on T cells) could trigger signal transduction via association with Src family protein tyrosine kinases such as Lck.[283] Downstream signalling events in CD55-costimulated lymphocytes have yet to be investigated. A possible model of CD97 and CD55 function in T-cell activation is shown in Figure 3.48.

Impaired migration of $CD8^+$ T cells to the lung or impaired cell survival could also conceptually explain the cellular phenotype observed. Selective migration of activated versus naïve $CD4^+$ and $CD8^+$ T cells to the lungs is mediated in part by selective expression of the integrin CD11a and chemokine receptors such as CCR4 and IFN-γ-induced CXCR3 (the receptor for CXCL10).[262] CD97 could influence cell migration and localisation via adhesive interactions with CD55 or other ligands, although evidence for this effect is lacking in murine knockout models (at least for granulocytes), and it is unclear whether the binding site for CD90, one of the implicated ligands, is present in murine CD97.[184,193] The rho-associated kinase pathway, one of the pathways implicated in CD97 signalling in other cell types[230], has been implicated in $CD8^+$ T-cell migration in inflamed lung[284], so if this pathway is activated differently in $CD8^+$ versus $CD4^+$ cells, a selective migration deficit would be possible. However, an effect on specific migration seems less likely than an effect on activation given that a change in the CD4:CD8 ratio (albeit of a lower magnitude) was also seen in peripheral blood of *Cd97-/-* mice. Reduced T-cell survival in *Cd97-/-* mice, for example due to enhanced apoptosis in response to stimuli such as TNFα, could be a plausible alternative explanation for the phenotype. Although it has not yet been documented in T cells, and the possible mechanisms underlying differential effects between T-cell subpopulations are not clear, an anti-apoptotic effect of CD97, on both intrinsic and extrinsic apoptotic pathways, has been reported in a human fibrosarcoma cell line. This effect requires a functional autoproteolysis domain, but appears to be independent of CD55 binding.[223] Further studies will be required to elucidate which of these possible mechanisms are responsible for the inefficient $CD8^+$ response.

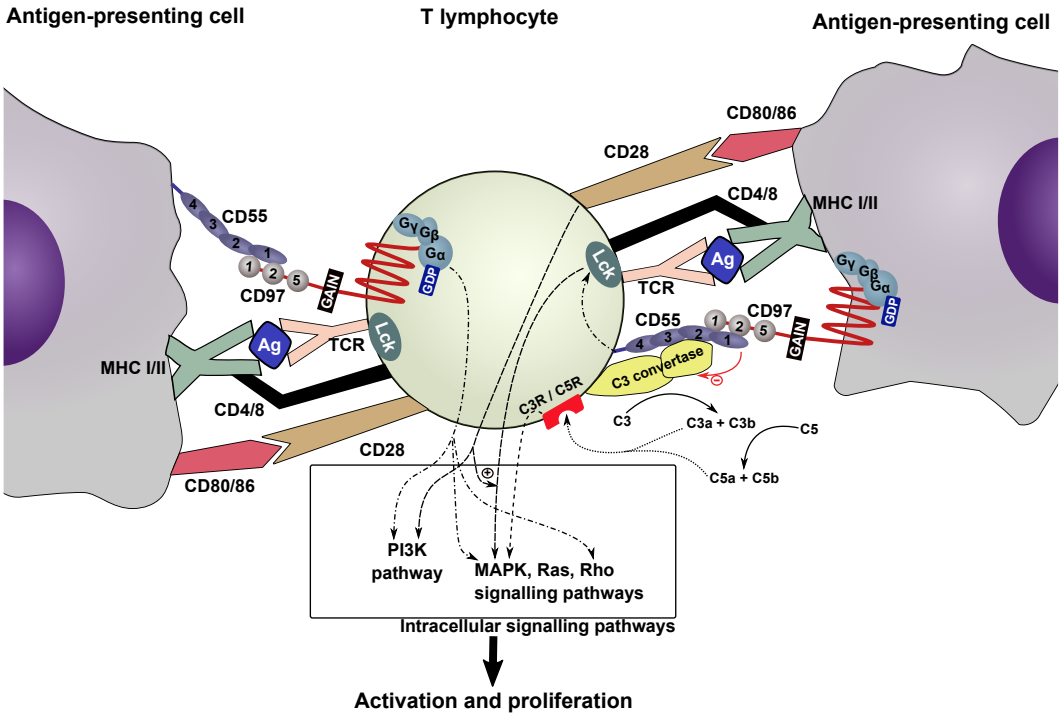

Figure 3.48 – A proposed model of CD55 and CD97 involvement in T-cell costimulation. Recognition of MHC-presented antigen by the T-cell receptor together with the CD4 (for MHC II) or CD8 (for MHC I) coreceptor leads to phosphorylation by Lck and initiation of intracellular signalling cascades leading to proliferation and activation. Classical costimulatory receptors such as CD28 augment this via a number of signalling pathways including the PI3K pathway. CD97-CD55 interaction could provide an alternative costimulatory signal in either direction, either via modulation of Lck and other kinase function by CD55, or by CD97-mediated activation of G-protein-coupled pathways such as MAPK or Rho kinase pathways. CD55 inhibition of complement activation could have an opposing effect, reducing activation via C3a and C5a and their associated receptors. Ag: antigen; C3R / C5R: complement C3 or C5 receptors; Lck: lymphocyte-specific protein tyrosine kinase; MAPK: mitogen activated protein kinase; MHC: Major histocompatibility complex; TCR: T-cell receptor

3.3.3 Relevance to human disease and limitations of the murine model for studying the role of CD97 in the host response to IAV infection

In this murine model of primary IAV infection in naïve animals, I have demonstrated a dysregulation in the T-cell response with CD97 deficiency, which results in a small transient change in disease severity. Although dysregulated

T-cell responses have previously been observed in *Cd55-/-* mice and effects of CD97 on T-cell function have been demonstrated *in vitro*, this is the first time that an effect of CD97 on the T-cell response has been reported *in vivo*. Since prior data on the role of CD97 in T-cell costimulation have been derived from human cells, these findings suggest that CD97 may also play a role in regulation of the T-cell response to IAV infection in humans. The T-cell response is likely to be critical in elimination of viral infection in humans as in mice, but cause versus effect of associations between deficits in T-cell immunity and severe disease is harder to establish in observational studies of human disease. For example reduced $CD4^+$ and increased $CD8^+$ T cells have been reported in severe compared to moderate H1N1 influenza pneumonia in children, but this is likely be a response to, rather than the cause of, increased infection severity.[285] A causative relationship is more plausible for the reduced $CD4^+$ and $CD8^+$ counts observed in a meta-analysis of T-cell responses in severe versus non-severe disease after SARS-CoV-2 infection, although this is not associated with a change in CD4/CD8 ratio.[286] These data are however based on peripheral blood counts (as lung or BALF samples are less frequently available from clinical cases), and so do not necessarily reflect the pulmonary T-cell response.

The genetic association between severe influenza and a variant causing reduced expression of CD55 has been assumed previously to result from increased complement-mediated host cell damage due to a partial loss of the complement inhibitory function of CD55.[70] This hypothesis, supported by an *in vitro* model, would certainly be a plausible explanation, possibly compounded by an increase in immunopathology due to complement-mediated T-cell reactivity as observed in *Cd55-/-* mice in other models. However, results from IAV infection models in *Cd55-/-* mice suggest that the effects of CD55 in influenza are more complex: CD55 deficiency results in attenuated disease severity and lung immunopathology for strains such as A/X-31 and A/Cal/04/09, but not A/Eng/195 or A/PR/8/34, in association with *decreased* BALF C3a concentration. This effect, the reverse of that predicted from human genetic data, is abolished in the absence of C3, suggesting that it is either complement-dependent or overwhelmed by the detrimental effects of complement deficiency. Effects on cell migration in these models differed between challenge strains, but included reduced early monocyte and neutrophil infiltration with strain A/X-31, and reduced $CD4^+$ and $CD8^+$ T-cell infiltration for recombinant strain A/PR/8/34-HK6.

Viral NA can furthermore cleave sialic acid residues on CD55, with variable efficiency between IAV strains, attenuating complement inhibitory function, providing a potential mechanism for the virus to modify immunopathology.[149] It is not known if sialic acid cleavage, which is more efficient for the A/Eng/195 neuraminidase than for that of A/Cal/04/09, modulates binding affinity for CD97.

Our results showing an effect of CD97, a known interacting partner for CD55, suggest that non-complement effects of CD55 could also contribute to the disease phenotype in CD55 deficiency. Loss of CD97-CD55 interaction-mediated T-cell costimulation and a consequent inefficient cytotoxic T-cell response could lead to a failure of viral clearance which could itself trigger severe disease or could have additive effects with complement-mediated immunopathology. Since the effects of CD55 on T-cell activation via complement inhibition and via direct costimulation have opposing effects, it is difficult to predict which could predominate in CD55 deficiency *in vivo*. However, our results in *Cd97-/-* mice are markedly different from those in *Cd55-/-* mice, with contrasting IAV strain specificity, opposing effects on weight loss and IFNγ production, and an effect on viral clearance that was not observed in the *Cd55-/-* model. Reduced T-cell infiltration in the *Cd55-/-* model, although not evaluated for all IAV strains, was inconsistent and seemed to depend on the HA segment present.[149] These differences indicate that effects independent of the CD97-CD55 interaction predominate in at least one of these models. CD55-independent effects of CD97 have not been excluded in the mouse model, and it is furthermore possible that increased CD97 expression resulting from CD55 deficiency (as seen in murine models[203]) could enhance such effects in patients with genetic variants in *CD55*.

One of the principal justifications for investigating genetic susceptibility to severe influenza is to identify key pathways contributing to pathogenesis that could be manipulated therapeutically. The small magnitude of effect of CD97 deficiency in the mouse model, and the uncertainty regarding whether the effects are detrimental or protective, depending in the stage of infection, suggests that CD97 signalling is unlikely to be a useful therapeutic target in influenza. Our findings are, however, relevant to possible targeting of this receptor in cancer therapy. As CD97 is upregulated on the surface of many tumor types and is correlated with aggressive behaviour, there has been some interest in direct targeting of the receptor to reduce invasion and metastasis.[186,204,287] Our results suggest that such targeting could impair the cytotoxic T-cell response which is an important

aspect of anti-tumour immunity, and so this potential detrimental effect should be carefully evaluated in attempts to develop and translate CD97 blockade as a therapeutic strategy.

While IAV challenge in mice is a useful model for the immune response to IAV infection, there are some differences in the course of infection which warrant caution in extrapolating this model to humans. As epithelial injury is less of a feature in murine IAV infection than in human disease, the murine model may lack sensitivity to assess contributions of host factors to epithelial pathology. Commonly used measures of overall severity in the mouse model, such as weight loss and mortality, are relatively crude and will reflect a number of non-specific factors such as inappetence and dehydration rather than specifically reflecting the severity of pulmonary pathology, and so may not be directly relevant to the respiratory failure which is the primary cause of death in severe cases in humans. A further difference which could be of prime importance in this instance, where a T-cell deficit is suspected, is the relative importance of adaptive immunity. In the naïve mouse model, innate immunity will predominate in the initial stages while animals mount a primary immune response. In humans, in contrast, prior exposure to other IAV strains is likely in many cases, and memory T-cell and B-cell responses will come into play earlier in the course of disease. In some circumstances this could be detrimental rather than protective: pre-existing non-neutralising complement-fixing antibodies were implicated as a possible contributory factor to immune-complex-mediated lung immunopathology in middle-aged patients in the 2009 H1N1 pandemic.[288] The naïve mouse model may thus underestimate the importance of dysregulation in the T-cell or B-cell responses. Although challenge-rechallenge models of secondary infection in mice are possible to address this, challenge dose titration and standardisation will be more difficult, and consequently high numbers of animals would be needed for an adequately powered experiment.

Similarly, there are differences in CD97 expression which could affect the relevance of the model to humans. CD97 knockout in mice is a more severe deficit than would occur with naturally-occurring genetic variation (barring rare loss-of-function mutations). This makes it more difficult to extrapolate from the model to the likely effects of genetic variants, but increases the sensitivity of the model to detect a role of the protein in the host response, and may be a better indicator of the potential impact of therapeutic targeting than a more subtle gene downregu-

lation would be. Lack of CD97 expression in alveolar macrophages in mice, and less frequent expression of CD55 in antigen-presenting cells (at least under unstimulated conditions) limit the ability of the model to detect effects dependent on specific interactions with these cell populations (or a specific direction of CD97-CD55 signalling in such interactions). Finally, there may be species differences in functional redundancy in adhesion G-protein-coupled receptors between species: for example, mice lack the closely-related EMR2 receptor which, while it does not bind CD55, could duplicate some non-CD55-mediated effects of CD97.

Despite these limitations, the mouse model provides a useful insight into the potential role of CD97 in the host response to IAV infection that would have been difficult to obtain from human observational studies, and provides a starting point for future functional studies.

3.3.4 Future directions

The precise association between genetic variants in the *CD97* gene in humans, changes in gene expression, and changes in T-cell phenotype or disease severity will be difficult to establish due to the high homology with the *EMR2* gene and consequent difficulty in distinguishing genotype at the *CD97* locus from that at the *EMR2* locus. This renders many analyses such as estimation of linkage disequilibrium or eQTL function unreliable, and so it is difficult to identify potential causative variants for which the disease-associated variant rs2302092 could be a proxy. Even if this is itself the causative variant, proof of this could be difficult. For example, if an effect on gene expression or function was to be demonstrated by CRISPR gene editing at the site, it would be difficult to design guides which would not also target *EMR2*, and consequently other measures would need to be taken to exclude effects of the other gene (such as additional gene knockout). Although such mechanistic genetic inference could conceptually be useful in establishing individual genetic risk, which could underpin the development of personalised therapeutics, the genetic association could alternatively be viewed as a clue to likely host pathways involved in a complex integrated response which could be relevant in all individuals regardless of genotype. As such, it would be more productive to focus future efforts on the function of the gene itself in the host response.

The mechanisms underlying the differential effects on the $CD4^+$ and $CD8^+$ T-cell response are not yet clear. Further *in vitro* experiments could help to establish whether CD97 deficiency differentially affects costimulation of $CD4^+$ versus $CD8^+$ T cells. Previous costimulation assays in human T cells have relied primarily on recombinant proteins or cross-linking antibodies to activate receptors in isolated T-cell cultures, or on the use of blocking antibodies to neutralise cell-to-cell interactions in co-cultured cells.[148,211,213] As antibodies can have unpredictable effects with regard to receptor activation or neutralisation, co-culture of relevant cells (for example dendritic cell / T-cell co-culture) with genetic manipulation of gene expression (for example, use of *Cd97-/-* and *Cd55-/-* mouse strains or CRISPR knockout in human cells) would be a more robust approach. Similarly, *in vitro* culture models could be used to assess effects of CD97 deficiency on survival or activation of apoptotic pathways in $CD4^+$ and $CD8^+$ T cells in response to stimuli such as TNFα and cycloheximide. These *in vitro* approaches would also facilitate identification of the signalling pathways involved, by quantifying expression or activation of key pathway mediators, or by assessing the impact of genetic or pharmacological blockade of candidate pathways.

In vivo mouse models could still have some use in investigation of the role of CD97 and CD55 in modulation of the T-cell response in influenza. For example, T-cell deficient *Cd97-/-* mice could be used to assess whether changes in weight loss and disease severity after IAV challenge are dependent on the T-cell response, *Cd97-/- Cd55-/-* double-knockout mice could be used to assess whether the change in T-cell ratio is CD55-dependent, and adoptive transfer experiments could be used to identify whether the cellular phenotype depends on CD97 expression in T cells or antigen-presenting cells. However, given the subtlety of the phenotype, particularly the clinical weight loss phenotype, large numbers of mice would be required to robustly demonstrate elimination of such a phenotype, especially with the additional experimental noise likely to be introduced by additional perturbations such as immune cell depletion or replacement. As such, these experiments would require careful justification under the principles of 'Reduction, Refinement and Replacement'.

The greatest translational potential in CD97 biology lies in cancer therapy rather than in therapeutic immunomodulation in influenza. As therapeutic targeting of CD97 could modulate the T-cell response, as suggested by our results in an influenza model and by *in vitro* results in human T cells, investigating the poten-

tial impact of this in a cancer model should be a high priority for future research, both in clinical and preclinical models. For example, although CD97 expression on tumours has been correlated with various measures of tumour behaviour and prognosis, to date there have been no reported attempts to correlate this with the development of a tumour-specific cytotoxic (or regulatory) T-cell response. Evaluation of the T-cell response would be justified in any attempt to modulate tumour behaviour via CD97 manipulation in animal models. If a T-cell perturbation is observed in a cancer model, as predicted from our viral infection model, elucidation of the underlying signalling mechanisms (for example with the experiments outlined above) will be invaluable in determining whether these effects can be ameliorated therapeutically.

Chapter 4

Searching for causative variants within a disease-associated locus: a CRISPR screen for cis-acting regulatory variants

4.1 Introduction

4.1.1 Linkage disequilibrium and the challenge of identifying causal variants

In Chapter 3, I examined a genetic association in which the gene as a whole was a plausible modulator of disease risk, but the exact variant underlying the association was not clear. This is a common scenario. Genetic association studies provide evidence merely of correlation between the genotype at one or more variants and the outcome of interest. This does not necessarily indicate any causal link between the lead variant and the outcome, even with complete control for all potential confounding factors such as population stratification (an ideal which is seldom perfectly realised), as the variants at a locus are not independent of each other due to linkage disequilibrium (LD). The set of possible causative variants can be refined to a certain degree by Bayesian credible set analysis of fine mapping data, which quantifies the posterior probability that

each of the variants in a locus is the causative factor, subject to certain assumptions (e.g. that exactly one genotyped variant is causative). The credible sets (i.e. the sets of variants with a specified probability of containing the causal variant) produced by such analyses narrow down the set of plausible contributory variants, but in many cases there is no single obvious candidate, and regulatory annotations are often lacking.[34] Thus without further experimental evidence, we often do not know which of the correlated variants in a linkage disequilibrium (LD) block are driving an observed association.

Identification of a single causative variant at a disease-associated locus is desirable for two main reasons. First, it will allow more accurate prediction of disease risk where linkage disequilibrium is not complete (i.e. $r^2 < 1$) between the index variant and the true causative variant. This is particularly important where linkage disequilibrium differs between populations. Secondly, it will increase our understanding of the molecular mechanisms underlying the association, for example whether the effect is on expression or function of a target protein, or which transcription factors or other interactors are likely to be involved in a regulatory association. This in turn may provide important information on mechanisms involved in pathogenesis that could be modulated for therapeutic (or conceptually prophylactic) benefit. Given the innate limitations of observational genetic studies, experimental manipulation is therefore warranted to demonstrate which variants are functional rather than just linked bystanders.

The majority of disease-associated loci in genome-wide association studies occur in the non-coding genome, giving rise to the assumption that there is a regulatory element driving the association.[34,46] We have developed a method of interrogating a disease-associated LD-block, based on previously described concepts of CRISPR screening in the non-coding genome, to identify putative regulatory elements which are likely to contain the true causative variant. In this chapter, as a proof of concept for the method, I use this screening technique to interrogate potential causative variants in a novel association between the macrophage regulatory protein SIRPα and schizophrenia, a multifactorial neurodevelopmental disorder.[289]

4.1.2 An association between SIRPα and schizophrenia

Signal regulatory protein alpha (SIRPα) has recently been implicated in the aetiology of schizophrenia by proteome-by-phenome Mendelian randomisation.[289] Mendelian randomisation aims to provide evidence of a causative link, rather than merely a correlation, between a measurable factor and a disease process, by using random segregation of alleles of naturally occurring genetic variants that influence the level of that factor (e.g. variants affecting the concentration of a biological mediator) as a 'randomising instrument'. Subject to certain assumptions, such as lack of an alternative mechanism linking the instrument to the outcome without involving the factor in question, and lack of association between the instrument and confounding factors, this can be seen as conceptually equivalent to a randomised controlled trial.[35] Single nucleotide polymorphism rs4813319 was identified as a robust protein quantitative locus (pQTL) for plasma SIRPα concentration, and thus a suitable randomising instrument, in two independent cohorts. Mendelian randomisation using summary data from meta-analysis of multiple genome-wide association studies subsequently showed a significant association (FDR < 0.05) between this instrument and schizophrenia, with the C allele (MAF 38% in a British population[160]), and hence reduced plasma SIRPα, carrying an increased risk.[289]

Schizophrenia is a complex neurodevelopmental disorder, typically manifesting in late adolescence or early adulthood. The pathophysiology of this syndrome is incompletely understood, but seems to be associated with atypical cell-to-cell communication or aberrant function of specific neural cell populations without consistent changes in gross brain structure.[290] Risk of developing schizophrenia is highly heritable, but multiple environmental factors (especially in early life) also influence the risk: for example, pregnancy complications or adverse perinatal events (such delivery complications), being born in winter, and being raised in an urban environment are all associated with an increased risk.[291,292] The genetic basis of schizophrenia risk has been studied extensively, with a large number of loci implicated in studies ranging from candidate gene association to genome-wide association studies and subsequent meta-analyses.[293,294] These include loci containing genes with obvious roles in neural function in pathways targeted by existing therapies, such as dopamine receptor D2, and others with less clearly defined roles in development or signal transduction. Interestingly,

GWAS hits are enriched for variants in enhancers active in immune cells as well as in brain tissue (but not in other irrelevant tissues such as fibroblasts), suggesting a possible role for immune dysregulation which is consistent with reported immunologic disturbances in individuals with schizophrenia.[293] With the exception of large rare copy number variants[295] and some other rare loss-of-function mutations with high penetrance such as translocations involving the *DISC1* (Disrupted in Schizophrenia) gene[296], most reported genetic associations consist of relatively common variants with low individual effect sizes, often with a difference in allele frequency of 2% or less between affected individuals and controls.[290] An association between *SIRPA* and schizophrenia has not been reported in previous genetic studies.

SIRPα is best known for its regulatory roles in the immune system. It is a transmembrane glycoprotein in the immunoglobulin family, thought to be involved in negative regulation of tyrosine kinase-mediated signal transduction pathways. SIRPα on myeloid cells interacts with the surface receptor CD47, and this interaction provides an anti-phagocytotic signal that protects normal host cells from destruction but also may be hijacked by viruses (including pox viruses and SARS-CoV2) or tumour cells to evade the immune response.[297–299] Beyond its roles in the immune system, a mechanistic role in the aberrant neurodevelopment thought to be responsible for schizophrenia is plausible for a number of reasons. It is expressed particularly highly in brain tissue, with transcript levels comparable to that in monocytes and macrophages.[50,64] In a proteomic survey of the prefrontal cortex, it was one of the most strongly down-regulated proteins in samples from patients with schizophrenia, consistent with the direction of effect in the Mendelian randomisation study.[300] Animal models of SIRPα and CD47 deficiency or dysfunction both show a range of cognitive and behavioural deficits, including prolonged immobility on a forced swim test, impaired memory retention and altered social behaviour.[301,302] Finally, it has been shown to play a key role in synaptic maturation and regulation of synaptic pruning, preventing excessive removal of synapses by microglia.[303,304] Excessive pruning in the prefrontal cortex, a process which is potentially therapeutically modifiable, has long been implicated in schizophrenia, and may also underlie a previously documented genetic association between schizophrenia and complement component C4.[305–307]

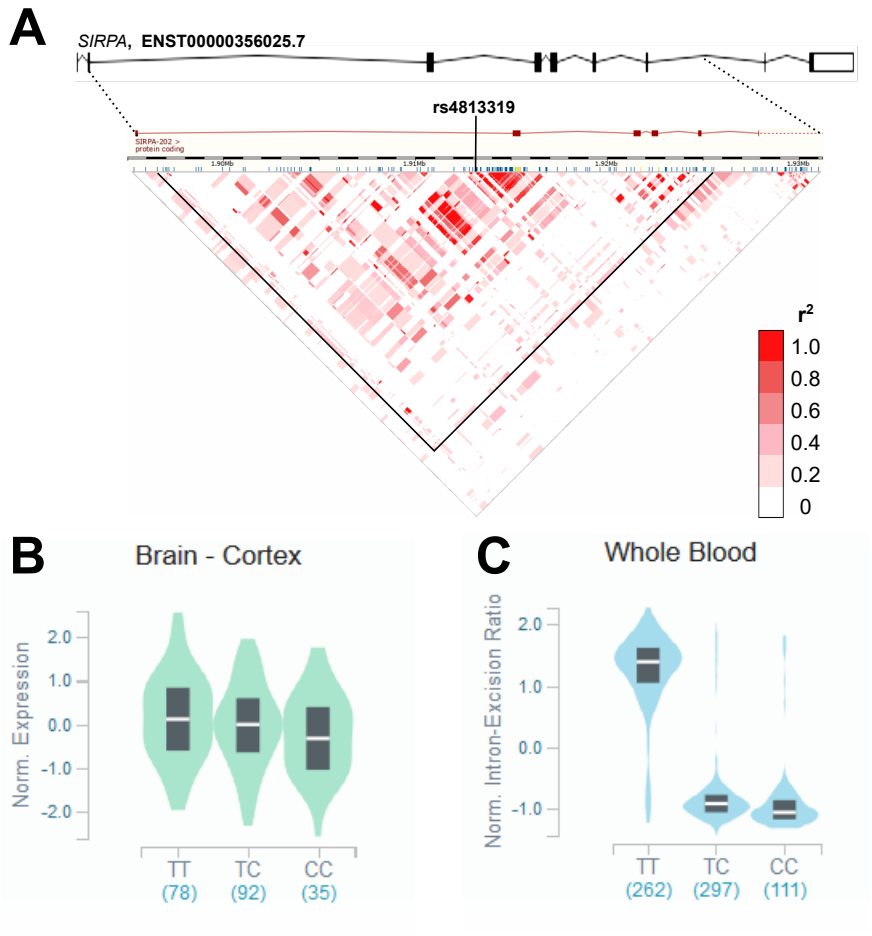

Figure 4.1 – LD block structure and expression effects of lead variant rs4813319. A: Structure of the most abundant *SIRPA* transcript, and location and linkage disequilibrium structure of the region containing rs4813319. The region delimited by the black lines contains all linked variants with a threshold of $r^2 > 0.2$. Data are from a British population in 1000 Genomes Phase 3 (n=91), accessed via the Ensembl browser.[160] B: Violin plot of effect of rs4813319 genotype on normalised *SIRPA* expression in cerebral cortex. The C allele is associated with increased schizophrenia risk. C: Effect of genotype on relative excision frequency of intron 3 of transcript ENST00000356025.7 in whole blood. Plots B and C use data from GTex version 8 and were generated using the GTex Portal; numbers in blue denote n.[64]

Although Mendelian randomisation provides evidence of a causative effect of SIRPα at the protein level, the index variant used as an instrument may not itself be the causative variant. Based on 1000 Genomes Data from a British population, 97 variants spanning approximately 30 kilobases are in linkage disequilibrium with the lead variant at a threshold of $r^2 > 0.2$ (Figure 4.1A). The

likelihood of each these variants being responsible for the observed association decreases as the strength of the linkage disequilibrium decreases, but these could all be considered as candidate causative variants. Most are located in the second and third introns and third exon of the most abundant transcript (ENST00000356025.7), and the protein-coding variants include ten missense mutations and one trinucleotide in-frame deletion (Table 4.1). These polymorphisms in the coding sequence are in a hyper-variable region of the N-terminal domain close to the CD47 binding site, but segregate away from the binding site itself. Protein variants differing in up to nine of these polymorphisms do not show differences in CD47 affinity.[308] Coupled with the strength of the evidence of a pQTL at this locus, underlying the initial association, this suggests that the causative element at this site is more likely to be a regulatory element than a functional protein-coding variant.

Data from GTex version 8 show that the index SNP and its linked variants modulate both *SIRPA* expression at the RNA level and splicing in a wide range of tissues including cerebral cortex: the 'risk' C allele is associated with a dose-dependent reduction in SIRPA expression, and reduced relative excision frequency of the normal third intron as measured by LeafCutter analysis of split RNA-Seq reads[309] (although isoforms with alternative splicing in this region are not commonly recognised).[64] The strong suspicion of the existence of a cis-acting causative regulatory element makes this disease-associated locus an ideal candidate for regulatory variant screening: this study aimed to interrogate the locus using CRISPR mutagenesis to identify the most likely candidate variants to be the true cause of the effect on schizophrenia risk.

Variant consequence	**Count**
Intronic variant	82
Synonymous variant	4
Missense mutation	10
In-frame deletion	1

Table 4.1 – Variant types in linkage disequilibrium with rs4813319.
97 variants are in linkage disequilibrium with the lead variant for this association (R^2>0.2), based on data for a British European population in 1000 Genomes.

4.1.3 Principles of CRISPR mutagenesis for identification of regulatory elements

The theoretical ideal method to confirm that a single variant is responsible for a change in gene expression would be to introduce a change from the non-risk to risk allele (and/or vice versa) *in situ* and to document an associated expression change. A number of methods exist to achieve this, including site-directed mutagenesis, CRISPR base editing and homology-directed repair of CRISPR/Cas9-induced double strand breaks.[310,311] Although high-throughput screens using homology-directed repair in a haploid cell line, or a CRISPR base editor in diploid cells, have recently been described for investigation of variants in the cancer-associated *BRCA1* gene[312,313], these techniques have a variety of limitations including low efficiency, limited range of inducible substitutions with base editors, limitations in target cell populations (e.g. need for haploid cells), potential for off-target effects and poor scalability for simultaneous investigation of multiple variants. Prime editing, a recently described technique combining CRISPR targeting with coupled primed reverse transcription to effect precise edits, may overcome many of these limitations, but has not yet been applied to parallel screen designs.[314]

An alternative method, slightly less specific but more efficient and scalable, is to use CRISPR mutagenesis to cause local sequence disruption at the site of the variant. The CRISPR (Clustered regularly interspaced short palindromic repeats) system is based on a naturally-occurring component of the adaptive immune response to bacteriophages in many bacteria. DNA fragments from bacteriophage genomes are incorporated into arrays of repeated elements in the bacterial genome; these fragments are later transcribed, and the resulting RNA molecules can be used to target a CRISPR-associated (Cas) nuclease enzyme to matching sequences in the DNA of subsequently invading bacteriophages, facilitating destruction of the foreign nucleic acid. Engineered versions of this system, most commonly incorporating single guide RNAs and the Cas9 nuclease enzyme, have a wide range of applications in molecular biology, enabling flexible and specific targeting of genome sequences.[173]

While the CRISPR/Cas9 system has been most commonly used to target the protein-coding genome, mutations resulting from nuclease activity can also be used to target the non-coding genome, as illustrated in Figure 4.2. If a single

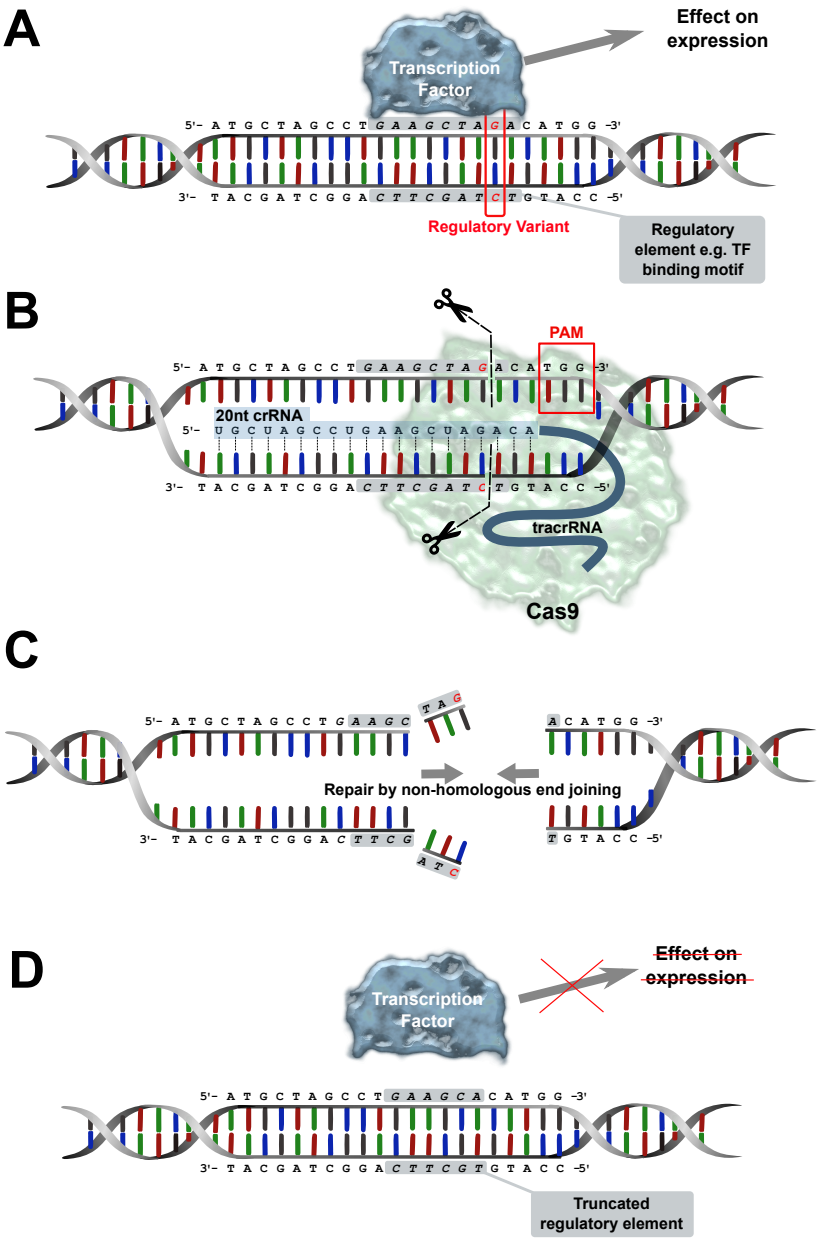

Figure 4.2 – CRISPR mutagenesis to identify regulatory elements. A: A regulatory variant (red) is located within a sequence-specific regulatory element (italic / shaded background), such as a transcription factor binding motif, that mediates a change in gene expression. B: An sgRNA, consisting of a 20nt crRNA complementary to the target site, and a linking tracrRNA, guides the Cas9 enzyme to the target sequence, adjacent to an NGG PAM motif. A double-strand cut is made 3 bp upstream of the PAM. C: The double-strand break is repaired by error-prone non-homologous end joining, with random indel formation. D: The repaired sequence has a truncated regulatory motif which in many cases will abolish or attenuate the effect on gene expression, e.g. by reducing transcription factor binding affinity.

genetic variant such as a SNP or short indel in the non-coding genome causes a change in gene expression, it must be located within a region of a regulatory element that is exquisitely sensitive to sequence changes (for example, a key transcription factor binding site). When Cas9 nuclease activity, guided to a target site by a specific guide RNA, causes double-stranded DNA breaks, these are in most cases repaired by error-prone non-homologous end-joining, which results in random deletions or short insertions at the site of the break. While these may not exactly replicate the effect of a genetic polymorphism, they will (assuming they are long enough to reach the critical residues of the regulatory region) ablate or attenuate any effect on gene expression that depends on the local integrity of the present sequence. For example, if a regulatory variant in an enhancer is part of the binding site of a transcriptional activator, sequence disruption is expected to reduce transcription factor binding affinity, and hence reduce transcription. Demonstration of expression change after targeting a sequence in the non-coding genome thus indicates that a regulatory element lies somewhere within the induced indel spectrum. The spectrum of indels produced varies between guides but is not random, depending largely on local sequence features such as microhomology[315], and while it leaves room for doubt as to the exact base or bases mediating a regulatory effect, it will at least narrow down the window of interest to tens rather than hundreds of bases.

Mutagenesis via CRISPR/Cas9 nuclease activity has the advantages of relatively flexible targeting and ease of implementation. Although still restricted by the necessity for a nearby protospacer-adjacent motif (PAM sequence; e.g. NGG for the commonly used *Streptococcus pyogenes*-derived spCas9), the range of indels produced will enable disruption of the target base for guides cutting within a short range on either side of the variant, allowing some flexibility in selecting a guides for optimal specificity and cutting efficiency. This system requires no additional repair template, and can be implemented in a variety of formats, including one and two-vector plasmid or lentiviral formats. It can easily be scaled to high throughput screens, and a number of screens have been reported specifically targeting regulatory elements in the non-coding genome. Reported screens have in most cases been based either on a saturation or tiling approach to a region of interest such as an enhancer or the region surrounding specific genes implicated in a physiologic or pathologic process, or have been targeted on specific elements such as transcription factor binding

sites. These screens, whether using gene expression or a functional phenotype such as cell growth as an outcome measure, have repeatedly demonstrated the ability of this type of CRISPR/Cas9-mediated mutagenesis to interfere with transcription factor binding and to identify regulatory features such as promoters and enhancers.[316–319]

In this chapter, I aim to adapt this methodology to identify candidate regulatory elements within the schizophrenia-associated LD block in the *SIRPA* gene.

4.1.4 Hypotheses and aims of this chapter

The principal hypotheses are that:

- The association between rs4813319 and *SIRPA* expression, which may play a causal role in schizophrenia, is mediated by a transcriptional regulatory function of one of the variants with which it is in linkage disequilibrium.
- A targeted non-coding CRISPR mutagenesis screen can help prioritise the most likely causal regulatory variants within a linkage disequilibrium block.

Specific objectives of the chapter are as follows:

- To develop a custom bioinformatic pipeline for design of sgRNA sequences for pooled targeted CRISPR/Cas9-mediated mutagenesis of variants in a linkage disequilibrium block.
- To investigate whether use of a Cas9 variant with an extended PAM range can improve screening efficiency by improving precision of targeting.
- To test the screening methodology using the association between rs4813319 and *SIRPA* for proof of concept.
- To prioritise the most likely candidate variants underlying the eQTL function at the locus containing rs4813319.

4.2 Results

4.2.1 Limited evidence of regulatory elements involving lead SNP rs4813319

To investigate whether index SNP rs4813319 could itself be the causative variant for the genetic association, evidence of local regulatory elements was evaluated in existing data sources. No candidate cis-acting regulatory elements (cCREs), defined on the basis of features including DNase I hypersensitivity and histone acetylation or methylation marks, overlapped this variant in the ENCODE database of DNA elements (Figure 4.3).[162,164] The nearest cCRE was a 229 bp region with characteristics of a distal enhancer, terminating 95 bp upstream of the lead variant. The evidence for activity of this cCRE is heterogeneous between tissues, but includes acetylation of histone protein H3 at lysine residue 27 (H3K27ac), a marker of enhancer activity[54], in multiple zones of cerebral cortex. The existence of this and other cCREs within the LD block is consistent with the observed association with protein and RNA expression, but lack of overlap suggests that r4813319 is not the most likely candidate to directly modify the function of such regulatory elements.

To complement this, evidence of transcription factor binding motifs, or evidence of transcription factor binding based on chromatin immunoprecipitation (ChIP), was retrieved from the Factorbook database[163], which includes data from 1813 human experiments including 682 transcription factors and 142 cell types, and incorporates data from ENCODE. Two ChIP peaks approximately 500-700 bp in length, for proteins argonaute-2 (a component of the RNA silencing system, for which the significance of DNA binding is uncertain) and RNA polymerase II phosphorylated subunit 2A, overlapped rs4813319 in single samples. No transcription factor binding motifs included this SNP. Again this provides limited evidence to support a direct causative role for this variant.

To test the direct effect of perturbation at this site, CRISPR mutagenesis was used to induce indel formation adjacent to the variant in THP-1 cells. THP-1 cells are a human monocytic leukaemia cell line, and were selected based on consistent high expression of SIRPα, potential biological relevance in relation to known functions of SIRPα in cells of the monocyte-macrophage system,

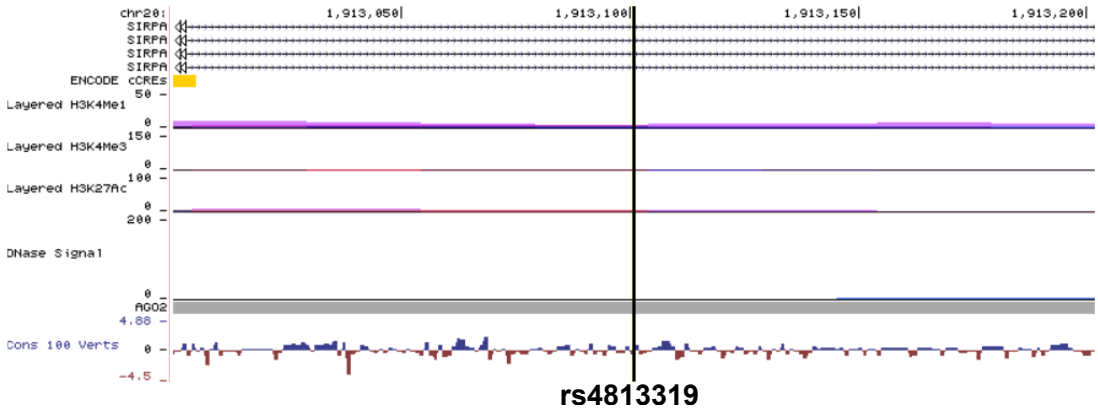

Figure 4.3 – Regulatory elements in the vicinity of rs4813319. UCSC Genome Browser[161] view of genomic features within ±100 bp of lead variant rs4813319. Histone methylation and acetylation marks from 7 cell lines, and DNase I hypersensitivity across 20 relevant brain and immune cell lines, are based on data from ENCODE.[162]

and ease of use of these cells in the laboratory (compared to primary cells). Enhancers and other regulatory elements can be highly tissue-specific, but although activity in the central nervous system is likely to be necessary for any effect on development of schizophrenia, the evidence of effects on plasma SIRPα concentrations[289] indicates that the regulatory effect is also likely to be present in peripheral tissues such as circulating leukocytes, and so a monocytic line is likely to be appropriate.

THP-1 cells were transduced with a lentiviral vector containing spCas9 and a guide RNA targeting as close as possible to rs4813319 (4 bp downstream), based on the availability of an NGG PAM sequence, and expression of SIRPα and SIRPβ1 was assessed by flow cytometry after seven days of puromycin selection. SIRPβ1 is a closely-related protein that has stimulatory rather than inhibitory effects and does not bind CD47.[308] Compared to negative controls transduced with vectors carrying either no sgRNA or a non-targeting sgRNA, targeted mutagenesis induced little change in the overall expression distribution and no significant change in median fluorescence intensity (Figure 4.4A-C), but a statistically significant increase in the small population of cells expressing very low levels of SIRPα (Figure 4.4B and D). Although this effect only involved a very small number of cells, and so may be vulnerable to experimental artifact, this could suggest that a small proportion of induced mutations could modu-

late expression. Sequencing of the targeted region showed moderate cutting efficiency with an estimated indel frequency of 27%, with predicted indels extending between one and eleven base pairs from, but not directly involving, the variant (Figure 4.4E): as such it is possible that any observed effects may not be directly attributable to the variant itself, and the overall sum of evidence for a direct causative effect of rs4813319 is weak.

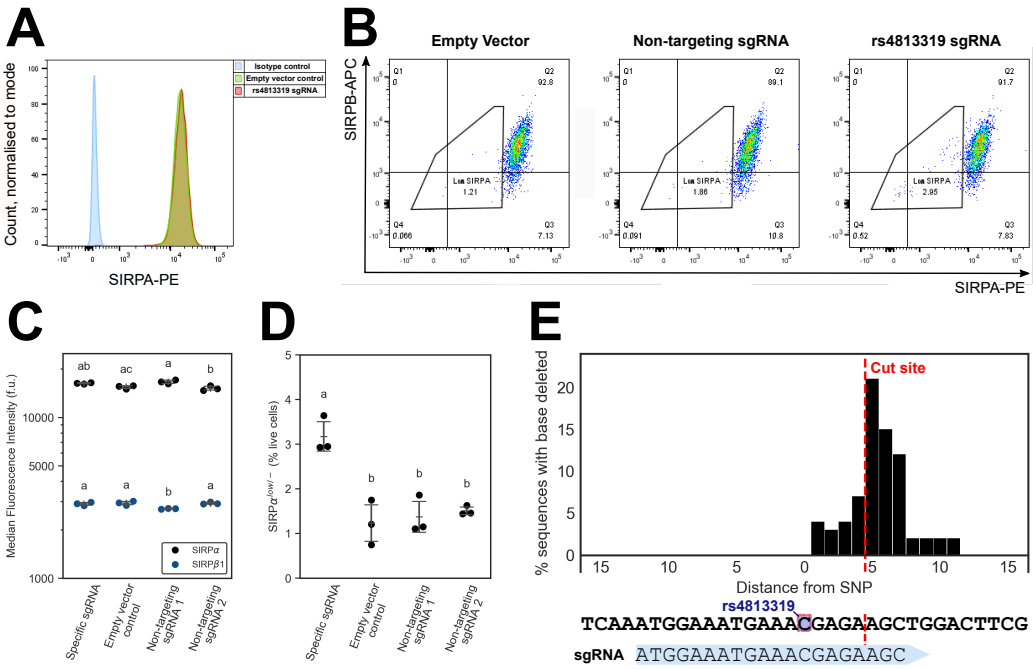

Figure 4.4 – CRISPR mutagenesis of rs4813319. A: Histogram of fluorescence distribution on flow cytometry after staining with anti-SIRPα antibody or isotype control, for THP-1 cells transduced with Lentivirus containing spCas9 and rs4813319-targeting sgRNA or no sgRNA (empty vector). B: Flow cytometry dot plots of SIRPα and SIRPβ1 expression with rs4183319-targeting or non-targeting guides or empty vector control. Plots in A and B have been gated on live single cells. C: Median fluorescence intensity of SIRPα and SIRPβ1 after transduction with targeting or non-targeting vectors as in A and B. D: Changes in the proportion of SIRPα$^-$ or low-SIRPα cells (see gating in B) with targeting or non-targeting vectors. In C and D, transductions have been performed in triplicate and groups compared with one-way ANOVA with Tukey's HSD post hoc test. Groups sharing a common letter are not significantly different ($p > 0.05$). E: Cut site and indel distribution (estimated by ICE analysis of Sanger sequencing data[179]) for the rs4813319-targeting sgRNA. Single nucleotide insertion was detected in an additional estimated 3% of sequences.

The presence of cCREs close to but not including rs4813319 suggests that one

of the linked variants in the vicinity is more likely to be the true causative regulatory variant, and prioritisation of these would help to find the true culprit. The CRISPR mutagenesis approach was thus extended to a high-throughput screen to target all possible variants within the LD block and identify which are located in sequences for which gene expression is vulnerable to sequence perturbation.

4.2.2 Screen design and pipeline development

For high-throughput focused targeting of sequences containing candidate regulatory variants within the 30 kb LD block, a pool of guide RNAs was designed to induce Cas9 cutting as close as possible to each variant. While algorithms for design of single guides or for genome-wide screens targeting protein-coding genes are commonly available, existing pipelines are less suitable for large-scale design of guides for the non-coding genome where the precise location is critical to success. I therefore created a custom design pipeline for this purpose, based on general principles of guide design previously reported for genome-wide knockout screens[172], but optimised to minimise distance from the targets.

Full details of the design pipeline are given in section 2.18.1. Briefly, the reference sequences for the variants and surrounding flank sequences were obtained from the variant database dbSNP.[159] All candidate 20-mer guides within the flank sequences were identified by the presence of an appropriate adjacent PAM sequence and a calculated cut site no more than 15 bp from the variant. For each candidate, on-target scores to determine likely cutting efficiency were calculated using the Doench *et al.* Rule Set 2 method.[173] All potential off-target matches with up to three mismatches and an appropriate PAM sequence elsewhere in the genome were identified by aligning candidates to the genome using Seqmap 1.0.8.[175], and the likelihood of off-target cutting for each of these matches was scored using a previously reported algorithm based on weighted effects of mismatches at each position in the 20-mer.[176] The percentage of guanosine and cytosine bases and the maximum length run of a single repeated nucleotide were also calculated for each candidate. Candidates were first filtered by the minimum criteria specified in Table 2.9. To rank the remaining candidates, numeric variables were binned (so that sorting was not based solely on a single continuous parameter) and sgRNAs were sorted by distance from the variant, on-target and off-target scores.

Distance from the cut site is critical to the chances of successful disruption of a regulatory element by this method, as illustrated by the indel spectrum in Figure 4.4E. Although the indel spectrum can be broad for some guides, and it may not be necessary to disrupt the target variant itself to interfere with function and hence enable identification of a regulatory element, closer targeting would increase the proportion of indels causing the desired mutations. The commonly used Cas9 variant spCas9 (from *Streptococcus pyogenes*) requires an NGG PAM sequence for efficient cutting (although nuclease activity is also seen at a much lower frequency for some other PAMs such as NAG), and this substantially restricts sgRNA options and thus the ability to cut close to the desired target. Recently, variants of Cas9 with an extended PAM range have been described, including xCas9, engineered from spCas9 by a process of *in vitro* phage-assisted evolution.[174] Although reported efficiency of xCas9 activity with different PAMs varied between guides and outcome measures (e.g. nuclease-mediated ablation of a reporter gene, indel frequency or Cas9-associated base editing), some activity was detected in at least one assay for around half of the 64 possible trinucleotide PAM sequences, with equal or greater efficiency to spCas9 at NGG PAMs and no increase in off-target effects. This extended PAM range would increase the available sgRNA options, and allow targeting closer to the variants of interest, with a large number of guides. The potential effects of adapting library designs for the xCas9 enzyme were therefore investigated.

At the time of development, no validated algorithms were available for xCas9 guide design, and so adaptation of the custom design pipeline to xCas9 required certain assumptions, including that on-target efficiency estimates (using a PAM-agnostic model) would still be valid. The off-target scoring algorithm was adapted to include matches with the extended PAM range. All 33 PAMs with evidence of xCas9 activity in at least one assay in the source publication were included, with weightings assigned according to a subjective evaluation of the consistency of evidence and maximal efficiency for each.[174] On-target scores were multiplied by these weightings, so that guides with high-efficiency PAMs such as NGG were preferred.

Including non-standard PAM sequences for xCas9 resulted in substantial improvements in guide coverage (Figure 4.5 A). While only 81/97 variants (84%) could be targeted within 15 bp by spCas9 with guides meeting the specified criteria (with a total of 222 unique guides, mean 3.0 per variant including guides

shared between neighbouring variants), 95/97 (98%) could be targeted using the extended PAM range (918 unique guides, mean 12.8 per variant, limiting to a maximum of 14 per variant). This improved coverage corresponded to reduced distances between the projected Cas9 cut site and the variant (Figure 4.5B), with the median of median cutting distances for guides targeting a specific variant decreasing from 7 to 4 bp. There was limited difference in on-target scores, indicating predicted Cas9 cutting efficiency before PAM weighting (Figure 4.5C, median of median scores 0.54 for xCas9 versus 0.53 for spCas9), or in off-target scores, indicating predicted specificity to the intended target (Figure 4.5D, median of median scores 0.49 for xCas9 versus 0.50 for spCas9).

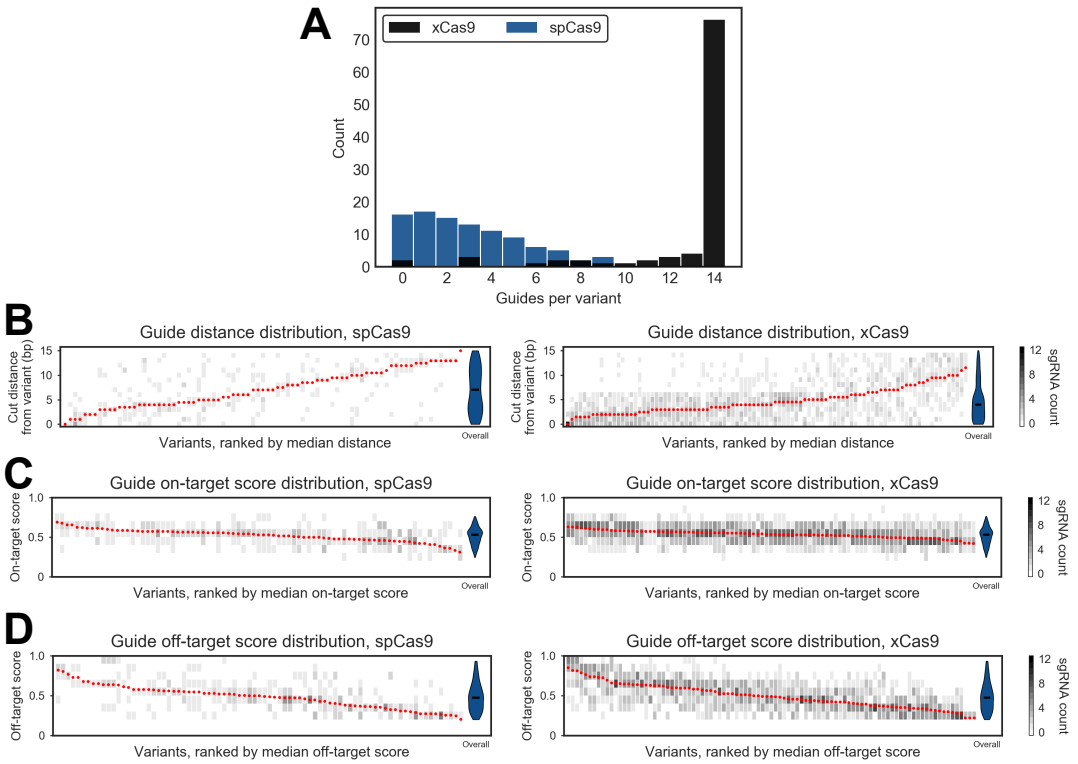

Figure 4.5 – Characteristics of libraries designed for xCas9 versus spCas9. A: Histogram of number of selected guides for each variant, for xCas9 versus spCas9. B: Distribution of distances from variant to sgRNA cut site for each variant. C: Distribution of guide on-target scores (Doench Rule Set 2) for each variant. D: Distribution of off-target scores for each variant (1.0 indicates perfect specificity). For plots B-D, variants have been ranked by the median value of the displayed measure (red points). On-target and off-target scores have been grouped in bins of width 0.1. Guide density within bin is indicated by shade. Overall library score distributions are shown by violin plots.

Due to the potential for closer variant targeting with xCas9, and hence increased chance of indels disrupting the variant of interest, designs for the extended PAM range were used to construct the sgRNA library. Guide sequences were synthesised, together with 78 human non-targeting guides selected randomly from a previously published list[177] as negative controls, and cloned into lentiviral vectors containing either xCas9 v3.7 or spCas9.

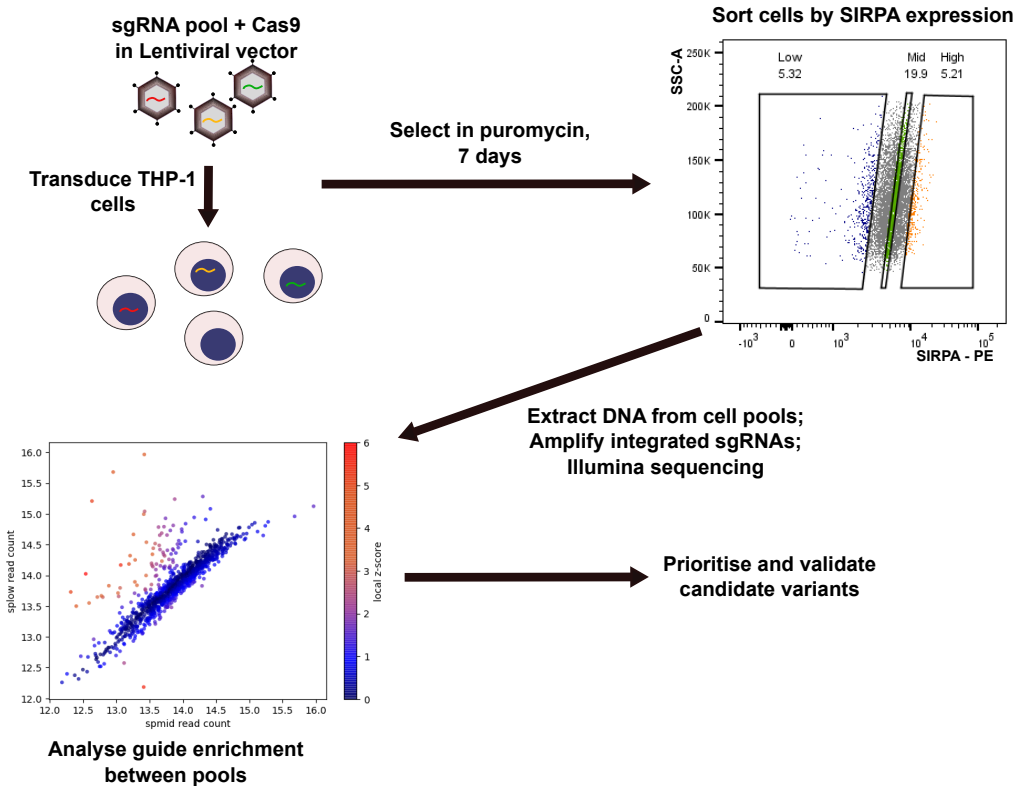

Figure 4.6 – Workflow for LD block-targeted regulatory CRISPR screen

These lentiviral libraries were used to perform pooled regulatory screens in THP-1 cells. The general workflow, as illustrated in Figure 4.6, is similar to that previously described.[317] Cells were transduced with the lentiviral libraries at a low MOI of 0.3 to ensure that few cells received more than one guide. Cells were subjected to antibiotic selection for seven days to remove untransduced cells and to allow time for maximal expression differences to develop. Cells were stained with fluorochrome-conjugated anti-SIRPα antibody to evaluate expression, and were sorted by fluorescence-activated cell sorting into three pools, consisting of cells with the lowest 5%, middle 20% and top 5% of fluorescence intensities. A cell population of at least 5000 cells per guide was

maintained at all stages of the process to ensure sufficient coverage. Integrated guide sequences from each of the pools were amplified by PCR and quantified by Illumina Next Generation Sequencing.

Guide enrichment between pools was analysed using a previously reported analysis pipeline[67], adapted to allow for targeting of more than one variant by some guides where variants are clustered close together. Briefly, this algorithm computes a z-score for the deviation of quantile-normalised guide read counts from the line of equality for a pairwise comparison, using sliding 'bins' of guide counts for the z-score calculations to account for heteroscedasticity across the range. Two-sided p-values for each variant are estimated by comparing the mean of associated z-scores, for all guides cutting within 15 bp of that target, to an empirical probability distribution derived from random permutations of an equal number of guide scores.

4.2.3 Guide enrichment depends on Cas9 variant and PAM sequence.

Sequenced libraries from the cell pools were of good quality with all 996 guides detected and a skew ratio (ratio of 90^{th}:10^{th} percentile read counts) of less than 3.5, compared to a recommended ratio of less than 10 for a genome-scale library[172], indicating highly homogeneous guide coverage.

Pairwise analysis of guide enrichment for a screen using the PAM-permissive version of the Cas9 enzyme, xCas9 v3.7, showed that normalised read counts were very closely correlated between cell pools expressing different levels of SIRPα, with only rare guides (including one non-targeting guide) deviating substantially from the line of equality (Figure 4.7).

In contrast, pairwise comparisons between cell pools for the screen conducted using the more PAM-restricted spCas9 (Figure 4.8), while still showing a high degree of correlation between pools for the majority of guides, showed clear enrichment of a number of guides. Minimal enrichment was observed for 77/78 non-targeting control sgRNAs, although a single control showed strong enrichment in the low-SIRPα pool, suggesting a possible unanticipated targeting effect (Figure 4.8B and F). Amongst targeting guides, there appeared to be a strong

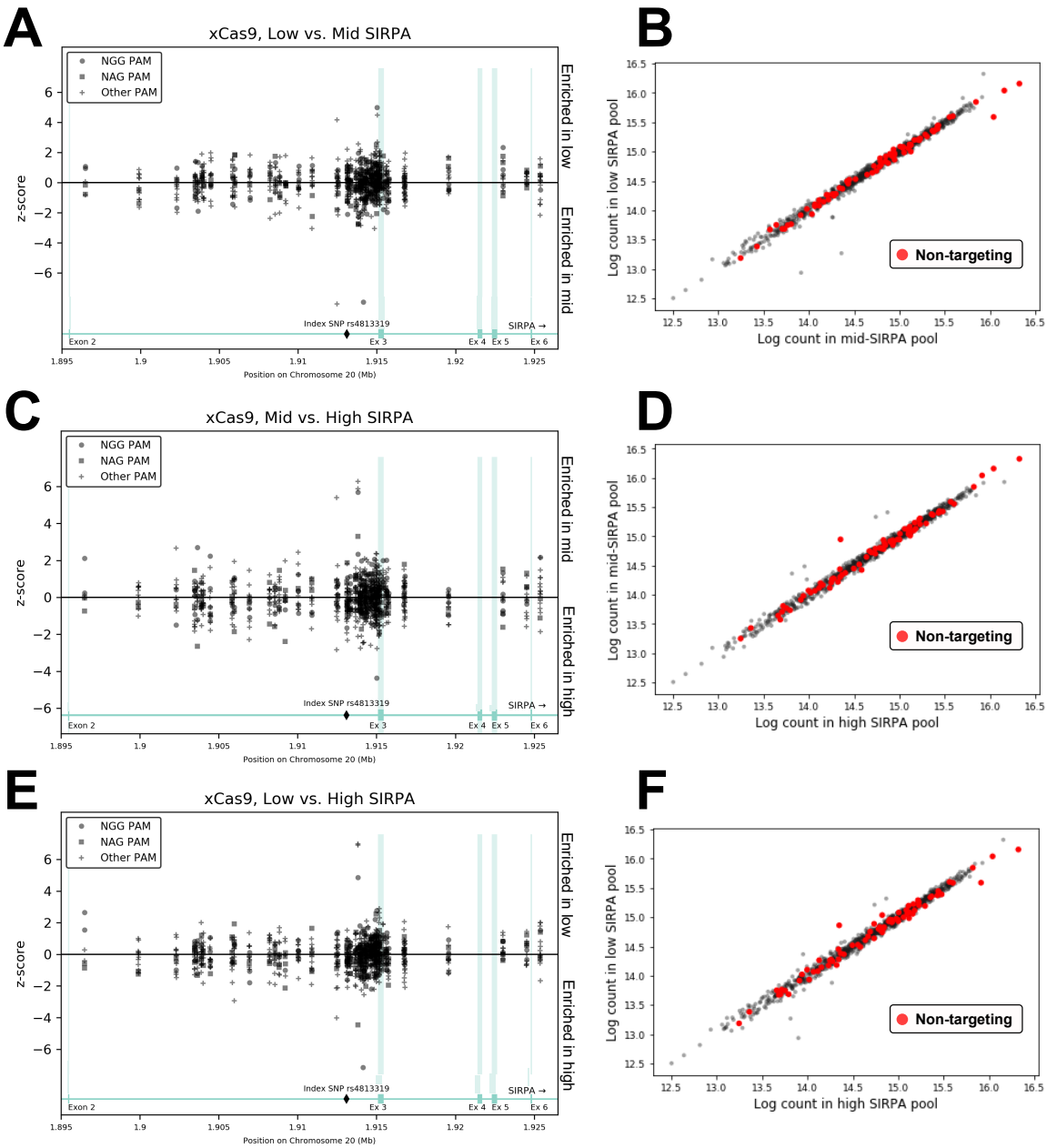

Figure 4.7 – Guide-level pairwise comparisons for a targeted regulatory screen using xCas9 v3.7 in THP-1 cells. A, C, E: sgRNA z-scores by position within the gene for pairwise comparisons of normalised read counts between cell pools with low vs. mid- (A), high vs. mid- (C) or high vs. low (E) SIRPα expression. Symbols denote PAM sequence type adjacent to the sgRNA target. Green shaded areas denote exons. B, D, F: Distributions of quantile-normalised read counts of targeting (grey) and non-targeting guides (red) for the same pairwise comparisons.

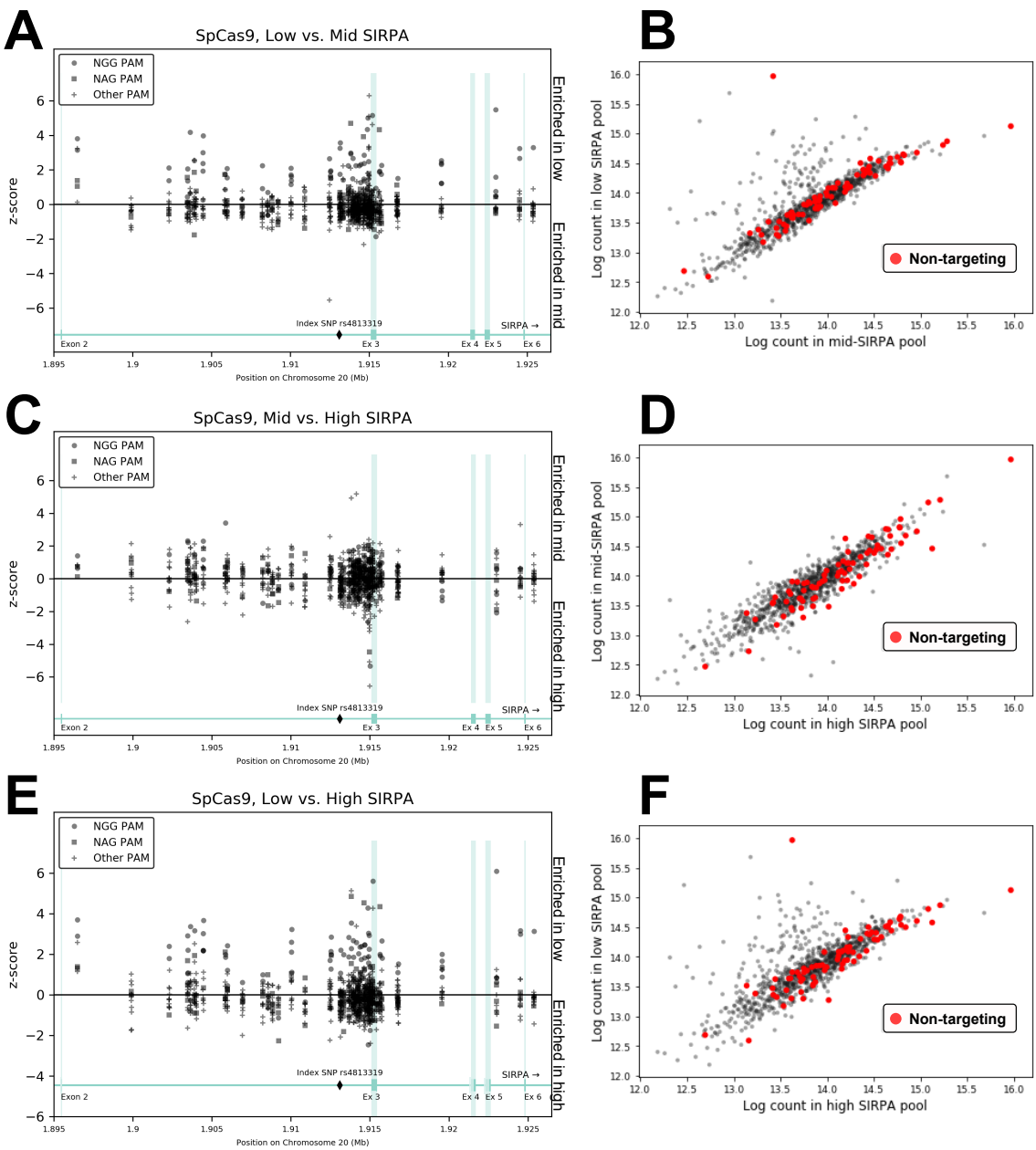

Figure 4.8 – Guide-level pairwise comparisons for a targeted regulatory screen using spCas9 in THP-1 cells. A, C, E: sgRNA z-scores by position within the gene for pairwise comparisons of normalised read counts between cell pools with low vs. mid- (A), high vs. mid- (C) or high vs. low (E) SIRPα expression. Symbols denote PAM sequence type adjacent to the sgRNA target. Green shaded areas denote exons. B, D, F: Distributions of quantile-normalised read counts of targeting (grey) and non-targeting guides (red) for the same pairwise comparisons.

bias towards enrichment in the low-SIRPα pool, suggesting that mutagenesis via Cas9 nuclease activity is more likely to reduce expression (e.g. by attenuating enhancer function) than to enhance it (e.g. by attenuating silencer or insulator function). Enrichment in the high-SIRPα compared to the mid-SIRPα pool was less apparent, and largely limited to a small number of guides just upstream of the third exon (Figure 4.8C and D).

For spCas9, guides targeting sequences adjacent to an NGG PAM sequence appeared to be over-represented among guides enriched in the low-SIRPα pool (Figure 4.8A). To confirm this apparent PAM bias, without prejudice as to direction of effect of targeting, distributions of maximum absolute z-score values (across all three pairwise comparisons) were compared by Anderson-Darling tests (Figure 4.9A) for targets with PAMs with expected high efficiency (NGG), low efficiency (NAG, NGA, NCG) or minimal activity (other PAMs and non-targeting guides) with spCas9. This did not support the null hypothesis that all scores were drawn from the same distribution (overall $p = 3 \times 10^{-13}$ by k-samples test), with significant differences between NGG and all other PAM groups except NCG but no other significant differences between groups ($p < 0.05$ on *post hoc* 2-sample tests with Bonferroni correction). This is consistent with the expected behaviour of spCas9, and suggests that observed enrichment is truly due to the Cas9 nuclease activity.

For xCas9, however, no such bias was observed, and the Anderson-Darling test indicated that score distributions for all PAM groups were drawn from the same population ($p = 0.69$; Figure 4.9). While lesser PAM bias would be expected for xCas9, the lack of difference between non-targeting guides and even expected high-efficiency PAM groups, combined with the observation that cutting at a subset of target sites is sufficient to cause enrichment with the alternative version of the enzyme, suggests that cutting efficiency was too low for reliable inference of expression effects induced by targeting. For this reason, further analysis focused primarily on results with spCas9 despite the reduced target coverage inherent in the PAM restriction.

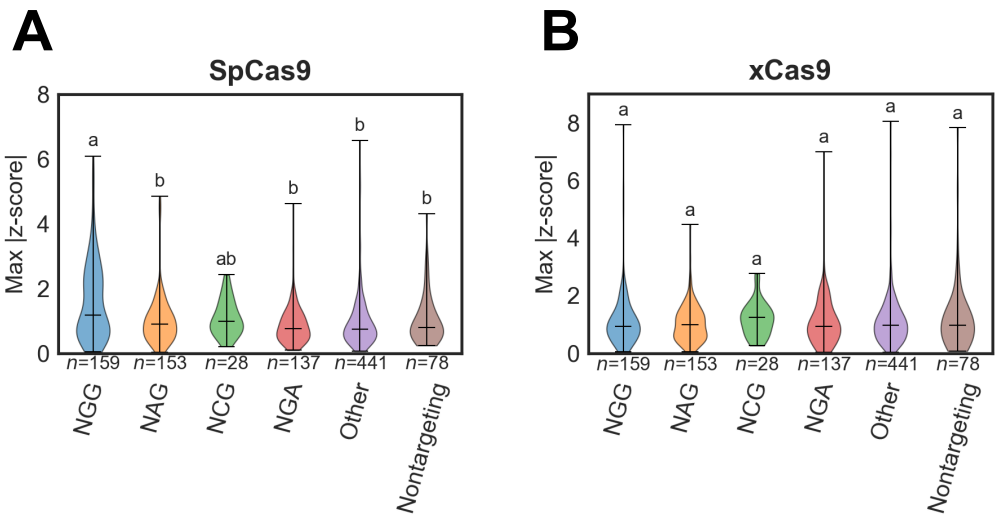

Figure 4.9 – PAM bias in guide enrichment. Maximum absolute z-scores across the three between-pool comparisons for guides with different PAM sequences adjacent to their targets, for spCas9 (A) and xCas9 (B). Common letters indicate no significant difference in empirical distributions between PAM groups (Bonferroni-adjusted $p < 0.05$ from pairwise 2-sample Anderson-Darling tests).

4.2.4 Target-level analysis implicates candidate regulatory variants.

Evidence of consistent enrichment of multiple sgRNAs targeting the same variant increases the confidence that the observed effects are a consequence of sequence disruption at that site, rather than due to chance or off-target effects for a single guide. Enrichment at the variant level was determined using summed z-scores for all guides expected to cut within 15 bp of the variant. Due to the strong NGG PAM preference shown by spCas9, only those guides with a target NGG PAM were counted towards the aggregated score for each variant.

After correction for multiple comparisons (FDR < 0.05), ten variants were significantly enriched in cells with low versus mid-level SIRPα expression (Figure 4.10, including two protein-coding variants (rs17853847 and rs17855612) and eight intronic variants. Of these, seven (including the protein-coding variants) also showed statistically significant enrichment in the comparison between high and low pools, while three others (rs6081134, rs6147627, and rs7274853) had

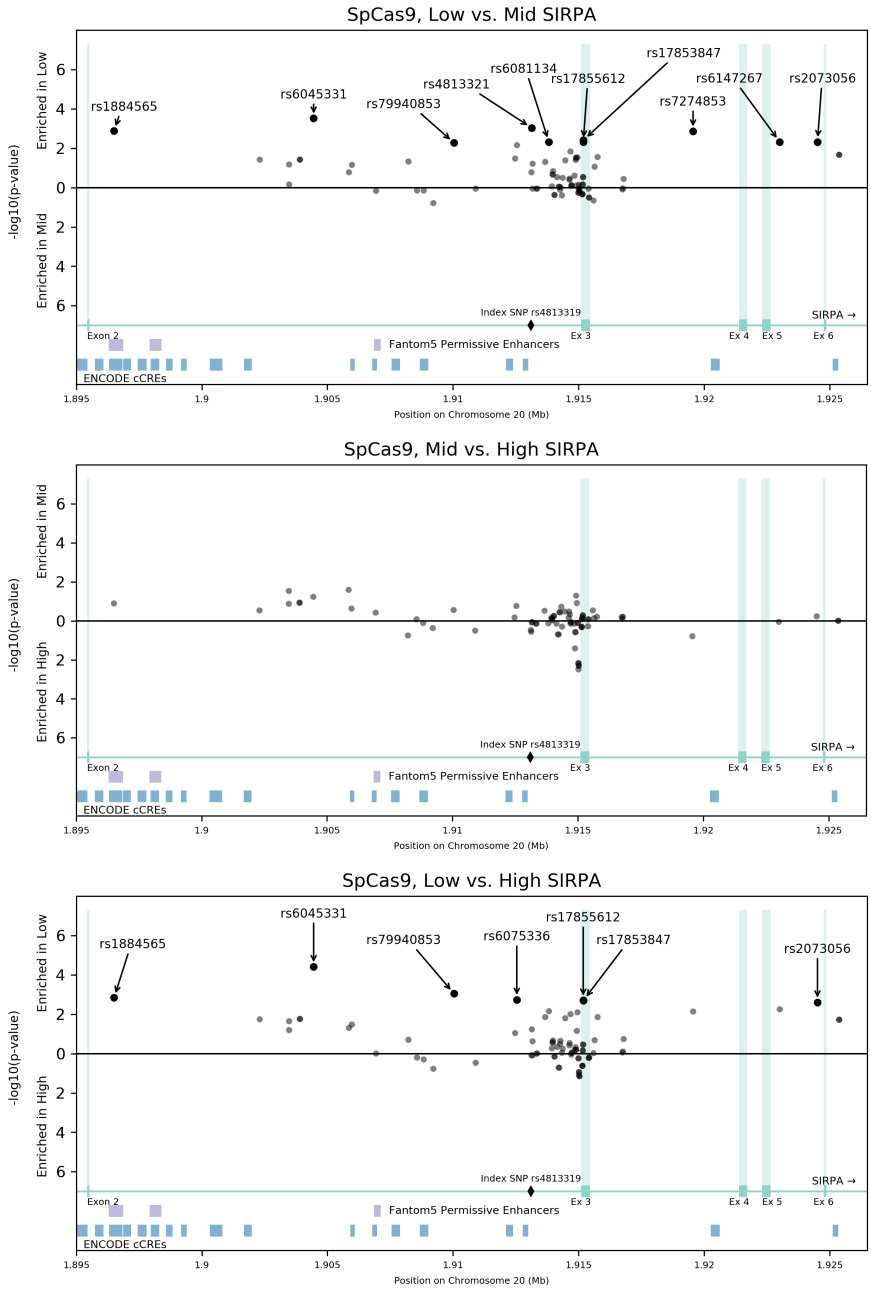

Figure 4.10 – Variant enrichment in a targeted regulatory screen using spCas9 in THP-1 cells. Two-sided p-values for mean z-scores (based on NGG PAMs only) for each targeted variant, estimated from an empirical probability distribution derived from random permutations of all guide scores, are shown in relation to the target position within the gene, for three pairwise comparisons between cell pools. Labelled variants are significant after correction for multiple comparisons (FDR < 0.05).

borderline FDR values (0.05 - 0.06). One additional variant (rs6075336) was significant in the high versus low comparison, while having a borderline FDR value in the low versus mid-SIRPα comparison. No significant differences were detected between pools with high versus mid-SIRPα expression.

Although the evidence for xCas9 activity was weak at the individual guide level, variant-level analysis was conducted for the xCas9 screen to determine whether aggregating weak effects of a large number of low-efficiency guides could increase power sufficiently to allow identification of enriched targets (Figure 4.11). Only three variants showed statistically significant enrichment in any pairwise comparison, and the direction of effect was not consistent: while they were significantly enriched in either the high or low-SIRPα pool compared to the mid-SIRPα pool, there was no significant difference between high and low pools, reducing the confidence in the results. This confirms that xCas9 was not useful in the context of this screen.

The lack of enrichment of most protein-coding variants was surprising, as random indel formation at these sites would be expected to induce frame-shift mutations in a proportion of cases, resulting in a dramatic drop in production of the correct protein. In a minority of cases, this could be attributed to lack of available targeting sgRNAs, especially with an NGG PAM requirement: even where guides are available, some false negatives are expected for a screen such as this as cutting efficiency for individual guides is variable and cannot always be predicted accurately. In other cases, deviation from the reference sequence could have compromised cutting efficiency. These variants all occur in clusters of multiple closely linked variants, typically three or four within a space of less than 20 base pairs. If one variant carries the non-reference allele, this will be true of all variants in the cluster; since the sgRNAs are designed from the reference genome sequence, these multiple deviations from the reference could have substantial detrimental effects on cutting efficiency of all sgRNAs targeting that specific region.

This is not however sufficient to explain the consistency of lack of enrichment in protein-coding variants with multiple appropriate sgRNAs (all in exon 3): for the nine variants with available Sanger sequencing data, the THP-1 cell line was homozygous for the reference allele in all cases. An alternative possibility is that alternative isoforms of this protein are produced either as a normal process

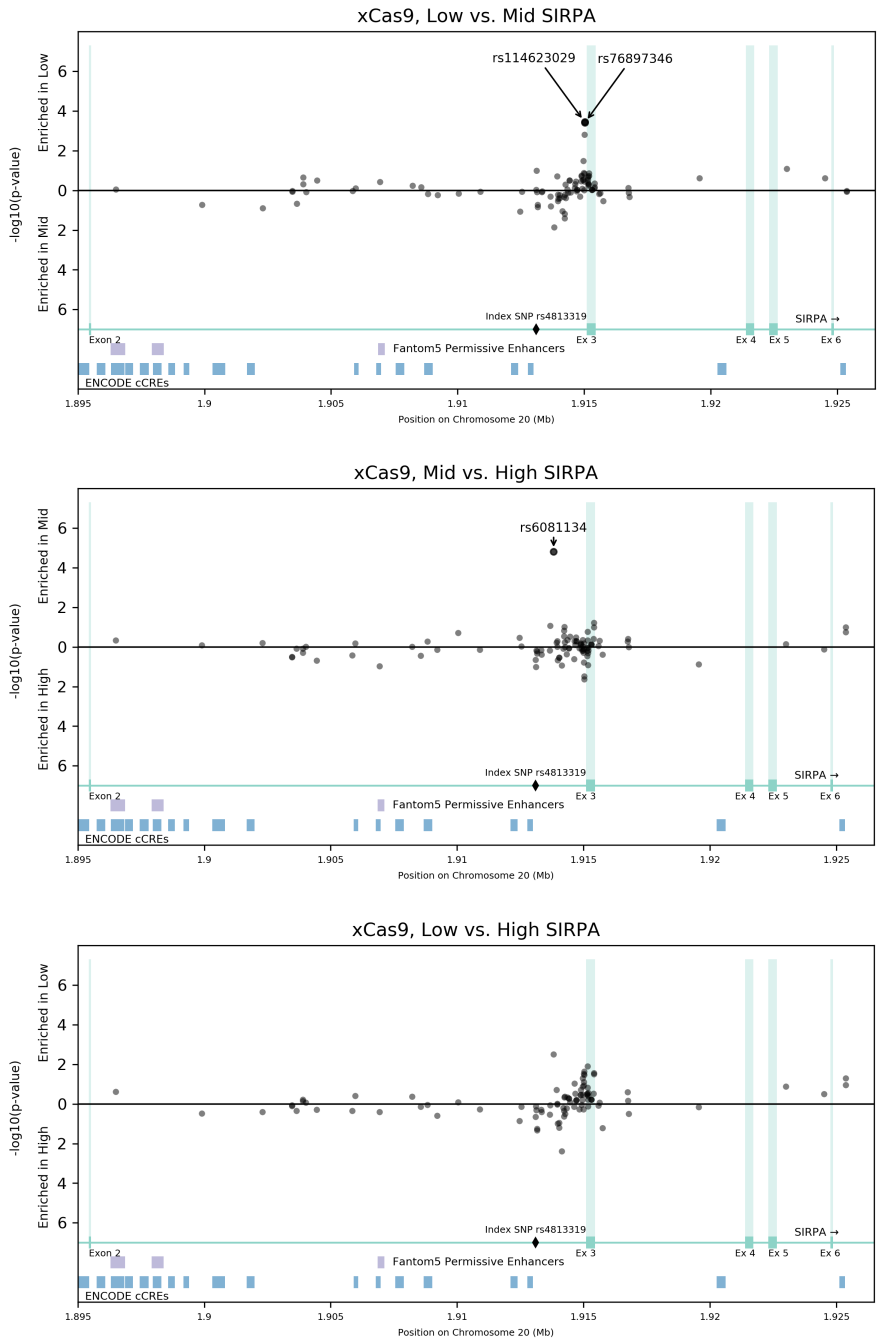

Figure 4.11 – Variant enrichment in a targeted regulatory screen using xCas9 in THP-1 cells. Two-sided p-values for mean z-scores (irrespective of PAM sequence) for each targeted variant, estimated from an empirical probability distribution derived from random permutations of all guide scores, are shown in relation to the target position within the gene, for three pairwise comparisons between cell pools. Labelled variants are significant after correction for multiple comparisons (FDR < 0.05).

in THP-1 cells or in response to mutagenesis in this exon: isoforms with variable starts to this exon have been identified in the Illumina Human Body Map 2.0[320] (for example in lung tissue), and CRISPR mutagenesis of early exons can in some cases lead to exon skipping rather than complete gene knockout[321], so this is plausible. The antibody used in the screen detects an unknown epitope in the extracellular domain of the protein, and so may not be sensitive to changes in isoform if its target domain remains intact.

	r^2 with lead SNP	Distance from index SNP (bp)	Expression change on perturbation	FDR for low vs mid	FDR for high vs low	Significant guides with NGG PAMs	Additional significant NAG or NGA guides
rs6045331	1.0	8,648	↓	**0.025**	**0.003**	4/5	0
rs1884565	0.28	16,597	↓	**0.029**	**0.028**	2/2	1
rs4813321	1.0	37	↓	**0.029**	0.20	3/3	0
rs7274853	0.42	6,462	↓	**0.029**	0.058	3/5	0
rs6081134	1.0	720	↓	**0.044**	0.058	2/3	1
rs6147267	0.22	9,909	↓	**0.044**	0.057	2/5	0
rs79940853	0.39	3,063	↓	**0.044**	**0.028**	3/5	0
rs2073056	0.37	11,416	↓	**0.044**	**0.030**	2/2	0
rs6075336	0.39	558	↓	0.052	**0.028**	4/4	1

Table 4.2 – Summary of evidence for 9 non-coding variants from a regulatory screen in THP-1 cells. Results are from the spCas9-based screen. FDR: Benjamini-Hochberg adjusted p-value. 'Significant guides' are defined here as guides with a raw p-value < 0.05 for enrichment in either low vs. mid or high vs. low comparisons, based on an empirical probability distribution derived from non-targeting guides only.

The evidence for the significantly enriched non-coding variants in the low-SIRPα pool on the spCas9 screen is summarised in table 4.2. These include a number of variants that are in complete or near-complete linkage disequilibrium with index variant rs4813319, but also other more distant, weakly linked variants that may be less plausible candidates as the causative factor for the observed association between genotype, expression and disease.

4.2.5 Double cutting has confounding effects on screen results.

Variant-level analyses for NGG PAMs showed a tendency for lower p-values towards the limits of the locus for pairwise comparisons involving the low-SIRPα pool. I hypothesised that this apparent bias in the analysis could be a result of double cutting in a subset of cells transduced with more than one guide.

Although the MOI of 0.3 used in this screen was consistent with general recommendations for CRISPR screens, to minimise multiple guide transduction, this MOI will still result in approximately one sixth of transduced cells carrying more than one guide. This is likely to be of little consequence in a genome-scale screen, due to random dispersion of a large number of possible guide pairings, and the low likelihood that any two guides will target the same chromosome. In this type of targeted screen, however, all guides are targeting the same locus, and so co-transduction of a pair of effective guides is likely to result in deletion of the intervening segment of DNA, including regulatory regions or even whole exons. Inclusion of non-targeting guides, or in this case ineffective guides with non-NGG PAM sequences, will reduce the probability of such an event. However, even if this only occurs in a small proportion of cells, large potential effects on expression could result in these cells comprising a substantial proportion of cells in the low expression pool, with consequent large enrichment scores.

To investigate if this was a likely explanation for the observed effects, sgRNA z-scores were compared to the median distance between their predicted cut site and the cut site for all other guides with NGG PAMs. Greater separation between two guides is in general more likely to result in deletion of functional elements when the intervening region is deleted, and so such functionally significant aberrant deletions are more likely to happen for guide pairs involving a guide on the extremes of the location distribution. As predicted, there was an approximately linear relationship between median distance and z-score, with increasing enrichment in lower SIRPα pools for greater distances (Figure 4.12). This correlation was strongest in the low versus high-SIRPα comparison (Pearson r = 0.35, p = 7 x 10^{-6}), but also significant for the low versus mid-SIRPα (Pearson r = 0.30, p = 0.0001) and mid- versus high-SIRPα (Pearson r = 0.24, p = 0.002) comparisons. Some guides targeting close to the index variant (where guide coverage was densest) appeared to deviate from the overall trend, with

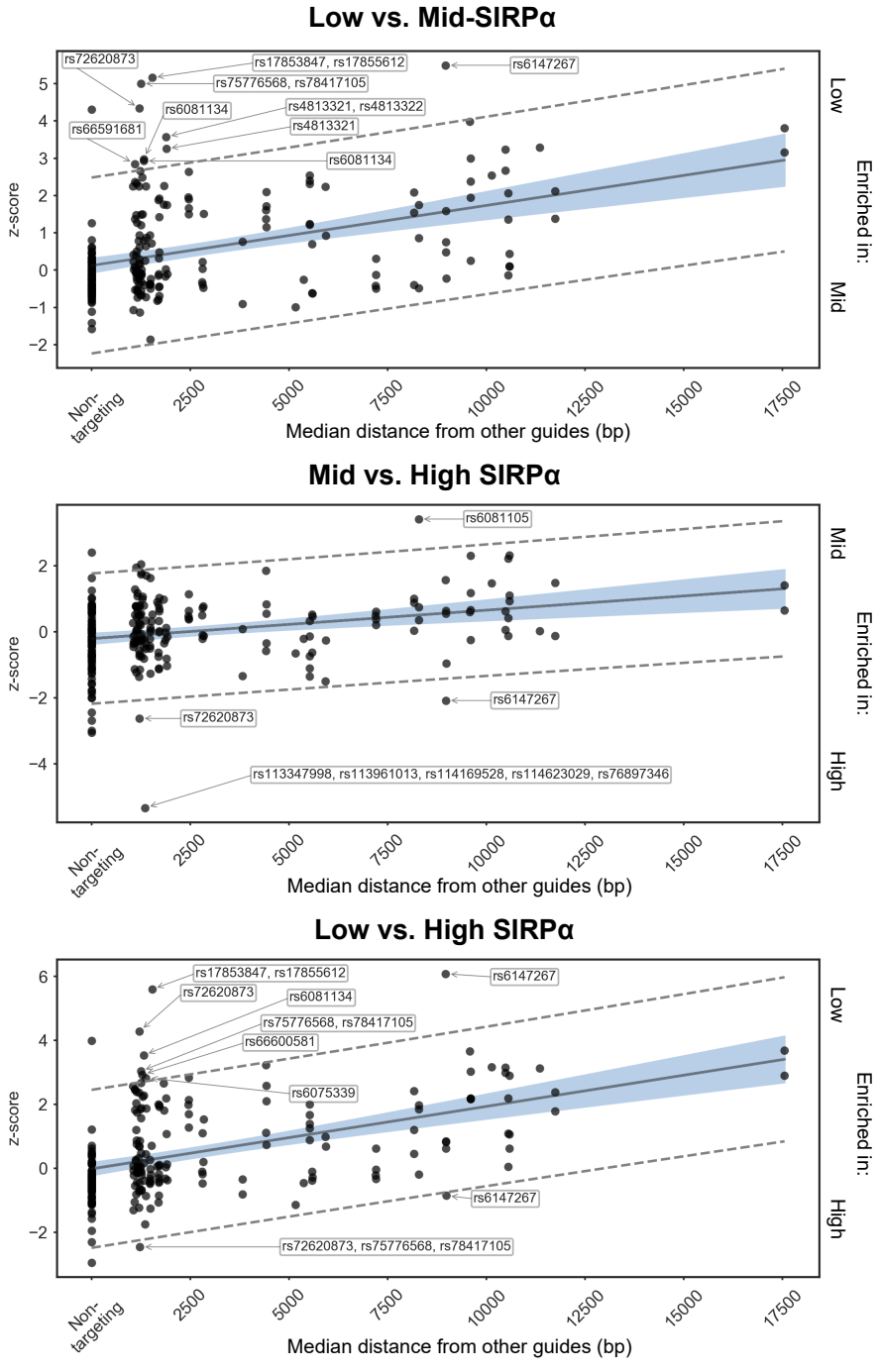

Figure 4.12 – Impact of guide dispersion on enrichment. Correlation between median distance from other guides and sgRNA z-score in pairwise cell pool comparisons. The grey line and blue shaded area indicate the fitted line (including non-targeting guides assigned a distance of 0, excluded from Pearson correlations) and 95% confidence band, while the dashed lines indicate the 95% prediction interval. Targets are given for top guides falling outside the prediction interval.

high enrichment scores despite low median distances: this region is the most likely to contain the true regulatory variant.

Statistical correction for this confounding effect is problematic as it cannot account for heterogeneity in cutting efficiency: ineffective guides would show lower enrichment in low-SIRPα pools than predicted based on their distance, and so could appear to be relatively enriched in higher SIRPα pools, distorting the probability distribution and increasing the chances of spurious observations. Correction based on a crude distance metric would furthermore not take into account likely differential effects of exact target position with respect to critical genomic regions (i.e. the likelihood that an exon or key regulatory region would be excised, versus an unannotated intronic region). Therefore, experimental design modification to limit this effect will be necessary in future screen replicates, but candidates can be prioritised to a certain extent based on whether guide enrichment falls outside the expected range for their location (see Figure 4.12).

Candidate variants close to the median location (and close to the index variant) that had sgRNAs with enrichment z-scores furthest outside the predicted range (Figure 4.12) included rs6081134, rs4813321 and the two significant protein-coding variants. Interestingly, some variants had individual targeting sgRNAs enriched in opposite directions, including the cluster of variants rs72620873, rs75776568 and rs78417105, located approximately 200 bp upstream of the exon junction and 160-180 bp upstream of the splicing branch point. It is likely that enrichment for one or more of these guides is spurious, and the cancelling effects mean that these variants are not enriched overall in target-level analyses. Bidirectional effects, although less likely, are however conceptually possible - if multiple functional elements with potential opposing effects are present in the vicinity, differing indel spectra of individual sgRNAs could result in differential disruption of these elements. More precise genomic editing would be required to dissect such finely localised effects if single guide enrichment was replicable.

4.2.6 Integration with other data sources assists with prioritisation of candidates.

Regulatory regions of the genome such as enhancers are often associated with chromatin features such as open chromatin or histone modifications. They may

also be identified by features such as bidirectional transcription or transcription factor binding sites. Integrating existing data sources on candidate regulatory elements can provide additional evidence on likely functionality of loci surrounding screen hits, and enable further prioritisation of candidate regulatory variants. To this end, the ENCODE, Factorbook, FANTOM5 and GTex databases were queried for evidence of chromatin features, transcription factor binding, bidirectional transcription and eQTL signals respectively across a wide range of cell types,[50,64,162,163] and the UCSC genome browser was searched for evidence of cross-species conservation, which when outside exons can also indicate likely regulatory regions.[42,161,166] Additionally, a published THP1-specific dataset was examined for evidence of local chromatin features at this locus.[322]

In THP-1 cells, before or after differentiation to macrophages, there was no evidence of open chromatin within this 30 kbp locus on the basis of an assay for transposase-accessible chromatin with sequencing (ATAC-Seq), a method which provides similar information to DNase hypersensitivity, using accessibility to a cleavage enzyme to identify chromatin state. Similarly, no ChIP-Seq peaks were found for the histone acetylation mark H3K27Ac, commonly associated with active enhancers.[54] Two regions of enriched binding within the locus were identified by ChIP-Seq for CCCTC-Binding Factor (CTCF), a transcription factor that is involved in establishment of three-dimensional chromatin structure and regulation of gene expression via promoter-enhancer interactions and silencer function.[62,323] These were, however, located towards the extremities of the locus and although the limit of each of the ChIP-Seq peaks lay within 150 bp of a variant in linkage disequilibrium ($r^2 \approx 0.3$) with the index SNP in the screen, no overlap was detected and the two closest variants were either not significant on the screen analysis (rs6081199), or not targetable in the screen due to homology with multiple other genomic sites (rs4814715).[322] This dataset, although the most specific to the cell type under study, thus proved uninformative in inference of regulatory elements within the target locus.

Overlap with other features suggestive of regulatory function is shown in Figure 4.13 and Table 4.3. Only one variant, rs1884565, lay within a candidate cis-acting regulatory region (cCRE) annotated in ENCODE, with proximal enhancer-like characteristics including evidence of DNase hypersensitivity, histone methylation and acetylation marks, multiple transcription factor binding peaks and positive conservation. This region is also identified as a robust enhancer in FAN-

TOM5, active principally in monocytes and monocyte-derived macrophages. The locus is thus highly likely to be a true enhancer. The variant itself may not necessarily be the causative element in this context, however: it is among the furthest from the index variant (with the potential for double cut confounding), and linkage disequilibrium between the two variants is relatively weak. It furthermore has relatively weak evidence of eQTL and sQTL function in GTex, compared to more strongly linked variants. The closest available specific sgRNAs had cut sites 10-12 base pairs away, and so may have disrupted sequence features not directly disrupted by the variant. There is limited data to suggest disruption of known motifs by the variant itself: for example, there is a VEZF1 binding site that overlaps the variant, but binding specificity is not predicted to change with the minor versus major allele.

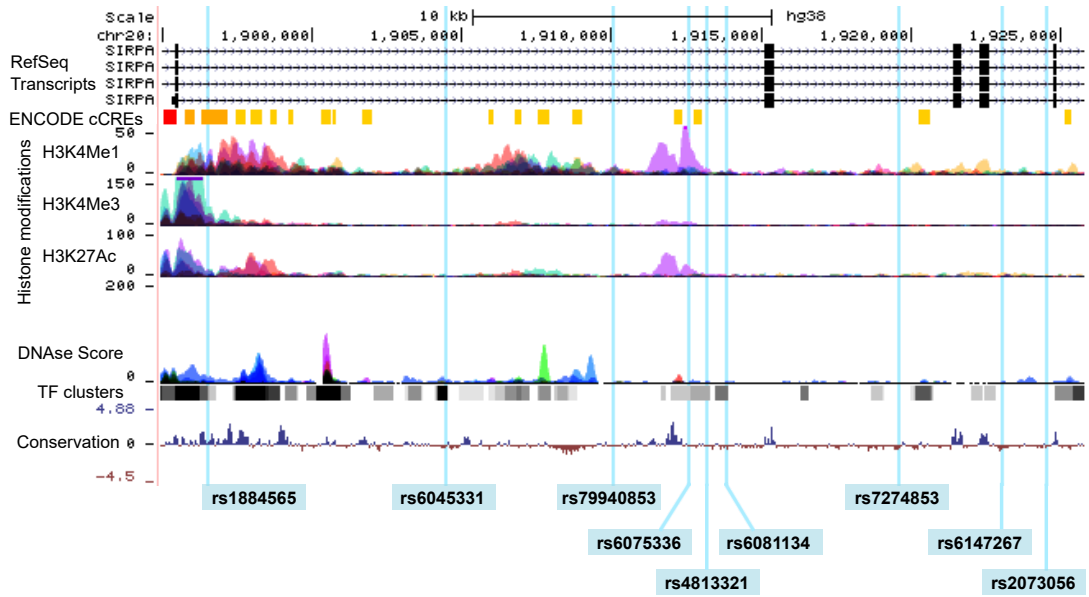

Figure 4.13 – Existing evidence of regulatory elements in the target locus in *SIRPA*. UCSC genome browser view, showing DNAse I hypersensivity, histone marks and candidate cis-acting regulatory elements from ENCODE, together with transcription factor occupancy ChIP peak clusters and PhyloP conservation score.

Of the other variants, rs6045331 and rs6075336 had the strongest supporting evidence of regulatory function. rs6045331 is located in a region with multiple transcription factor binding ChIP-Seq peaks, including for RNA polymerase II, CREB1 and multiple ATF transcription factors, and is also in a region of possible DNase hypersensitivity (with a lower signal than classic enhancers but higher

than most other variant sites). This variant is furthermore in perfect linkage disequilibrium with the index SNP and has the same evidence of eQTL function in GTex, but evidence for enrichment in the CRISPR screen must be interpreted with some caution given the physical distance from the lead variant. rs6075336 is much closer to the index variant, and overlaps histone acetylation and methylation marks compatible with regulatory function, as well as ChiP-Seq peaks for RNA polymerase II and CTCF, both often found at regulatory sites. This variant is however in weaker linkage disequilibrium with rs4813319, and has correspondingly weaker evidence of an influence on expression in GTex.

These two variants are among the top candidates to take forward for further validation. Others to consider include rs6081134 and rs4813321, which have little evidence of local open chromatin or histone marks, but have among the strongest evidence on the CRISPR screen, close linkage to the lead variant and some evidence of transcription factor peak overlap. In particular, RNA Binding Motif Protein 22 (RBM22) binding near to rs6081134 could be involved in regulation of splicing, and could provide a link to the differential splicing activity associated with this locus. Other more distant variants such as rs2073056 have little evidence to support the existence of local regulatory elements, which supports the hypothesis that their CRISPR screen enrichment could be artefactual. Validation of top candidates by single guide mutagenesis and specific base editing will now be necessary to confirm whether the variants themselves, rather than other neighbouring sequence elements, modulate gene expression or splicing.

	FANTOM5 Enhancer	ENCODE cCRE	TF ChIP peaks	TF motifs	DNase I hypersensitivity	H3K4Me1	H3K4Me3	H3K27Ac	Min. GTex *SIRPA* NES
rs6045331 (C/G)	×	×	POLR2A-pS5, GLIS1, CREB1, ATF2, ATF3, ATF7	None	±	-	-	-	-0.27
rs1884565 (A/G)	✓	✓	EBF1, POLR2A-pS5, ATF2/7, EZH2, PATZ1, ZNF692. ZFP69B, SUZ12	VEZF1	+	+	+	+	N.S.
rs4813321 (G/A)	×	×	POLR2A-pS5, AGO2	ORS2	±	+	-	-	-0.27
rs7274853 (G/T)	×	×	None	None	-	-	-	-	-0.16
rs6081134 (T/A)	×	×	RBM22	None	-	-	-	-	-0.26
rs6147267 (indel)	×	×	None	2 unannotated	-	-	-	±	N.S.
rs79940853 (C/T)	×	×	None	None	-	+	-	-	-0.16
rs2073056 (C/T)	×	×	POLR2A-pS5	None	-	-	-	-	-0.13
rs6075336 (G/A)	×	×	CTCF, POLR2A-pS5	None	-	++	+	+	-0.16

Table 4.3 – Regulatory features overlapping top 9 non-coding variants from a regulatory screen in THP-1 cells. DNAse hypersensitivity and histone marks are from ENCODE. TF: Transcription Factor; ChIP: chromatin immunoprecipitation; Min. GTex SIRPA NES: minimum normalised enrichment score (across all available tissues) for *SIRPA* in GTex v8.

4.2.7 Refinement of the screen design

This proof-of-concept experiment has enabled identification of a number of potential improvements to either the bioinformatic guide design pipeline or the overall experimental design.

First, the lack of apparent cutting efficiency of xCas9 in this case outweighed any potential benefits of the extended PAM range, and so future screens will only target NGG PAMs. Other engineered Cas9 variants with extended PAM ranges are available, such as Cas9-NG,[324] and it may be possible to expand variant-targeted library capabilities using some of these enzymes in future, subject to availability of robust models to predict targeting efficiency.

The confounding effect of longer deletions in the small proportion of cells transduced with more than one guide can be reduced by optimisation of the MOI, as well as by increasing the proportion of non-targeting negative controls. Estimation of the proportion of cells receiving more than perturbation, under a simple model where each cell is assumed to have an equal probability of transduction by each virus particle, shows a progressive reduction in this proportion as MOI is reduced (Figure 4.14). While approximately one sixth of transduced cells are expected to receive more than one guide under an MOI of 0.3, as used in this first experiment, this is reduced to approximately 5% of transduced cells for an MOI of 0.1. Our results suggest that a substantially lower MOI is desirable for this local targeted screen design than would be optimal for a genome-wide screen. The lower MOI limit will be dictated by practical concerns related to cell numbers needed and the growth characteristics of the cell type under investigation. For example, THP-1 cells as used in this experiment grow poorly below a density of 10^5 cells/ml and can exhibit altered differentiation characteristics above approximately 10^6 cells/ml, and so an MOI of 0.1, representing a theoretical fall from maximum to minimum cell density on addition of the selection antibiotic, may represent the practical limit for this cell type.

Sequencing of parts of the locus under investigation has shown that THP-1 cells contain the reference allele for some variants (e.g. rs17853847), and an alternative allele for others (e.g. rs4813319 itself). While the presence of an alternative allele does not necessarily preclude targeting by guides designed for the reference sequence, it will be detrimental to targeting efficiency in many cases, es-

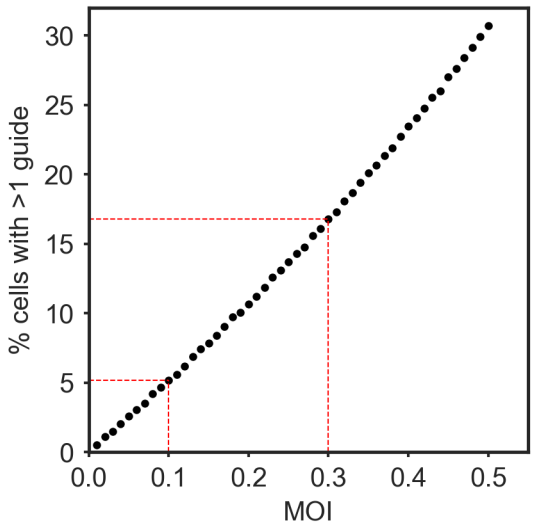

Figure 4.14 – Effect of MOI on multiple perturbation frequency. The proportion of successfully transduced cells receiving more than one lentiviral construct has been estimated using a computer simulation of 10^6 cells, assuming constant probability of any virus particle infecting any cell. The red dashed lines indicate the predicted effects of an MOI of 0.1 versus 0.3.

pecially if multiple mismatches are present within the 20-mer target sequence, or if mismatches occur close to the PAM sequence. Sequencing of the locus of interest in the cell type under investigation could help to generate a cell-specific guide library, but as an alternative where sequencing data are not immediately available or where more than one cell type may be used, guide designs can be generated to cover both reference and non-reference sequences, at least for common variants and those in linkage disequilibrium with the index variant. The library design pipeline has been updated to include this option, extracting data on both reference and alternative alleles from the dbSNP database.

I have implemented additional modifications to the updated design pipeline to further optimise filtering of ineffective guides or those with excessive off-target activity. Guides containing a 'TTTT' sequence will be removed, as this acts as a transcription termination signal when transcribing from RNA polymerase III promoters such as U6, used in common sgRNA vectors such as LentiCRISPR-v2. In the screen reported here, four of the guides targeting NGG PAMs contained a TTTT sequence. These had a low z-scores (maximum 1.1 across all pairwise

comparisons), and although the small number mandates caution in interpretation of any statistical comparison, the z-score distribution appeared to be different from the overall population (p = 0.035 on a 2-sample Anderson-Darling test), supporting the hypothesis that these sgRNAs are ineffective. Off-target scoring has also been updated to use the Doench cutting frequency determination (CFD) score, incorporating an optimised filtering threshold to exclude individual off-target matches with a very low probability of cutting from the scoring algorithm. This scoring system, which uses a matrix of weights for all possible base substitutions at each nucleotide position, in addition to weights for non-NGG PAMs, has been shown to have superior sensitivity and specificity in detecting likely off-target effects compared to the simpler position-weight matrix[176] used in the original designs.[173,319,325]

Additional efficiency improvements in the design pipeline have included automated calls to the current on-line version of the dbSNP database and most recent genome assembly, to avoid the need for large downloads of data files which may become obsolete, and filtering by PAM sequence at the earliest stage of the sgRNA selection, since flexible PAM selection is no longer required. Pipeline refinement is on-going, and the impact of these modifications will be assessed in future experiments. Updated and original versions of the design pipeline are available at https://github.com/baillielab/locus_crispr.

4.3 Discussion

4.3.1 CRISPR screen results reveal potential novel regulatory variants

From an initial candidate list of nearly 100 variants, all associated with differential *SIRPA* expression, but of which none had evidence of a direct role in regulation of gene expression, results of a targeted regulatory screen in THP-1 cells suggest a much narrower set of likely candidates to be the true causative variant. These include one variant in a typical proximal enhancer-like region, as identified by multiple existing databases using a variety of observational techniques, but also a number of other promising candidates in loci not previously

recognised as regulatory elements. While not always associated with expected chromatin features such as open chromatin, many of these are associated with regions of enriched transcription factor binding, providing a plausible mechanism of action. These direct experimental data on locus functionality complement other observations on correlates of regulatory function, but as with any high-throughput screen, results will require validation.

4.3.2 A workflow for candidate validation

To confirm that one of these polymorphisms directly affects gene expression will require demonstration of a change in gene expression after specific editing to change from one allele to the other, separating the effect of that variant from those of its neighbouring linked variants.

As a first step, to confirm that any enrichment is due to localised mutagenesis and not to inadvertent longer deletions from multiple guide transduction, expression change can be evaluated after targeting with each single guide. Sanger sequencing of the target prior to this will confirm which allele is carried by the cells, and thus identify if guide sequence modifications are desirable to improve targeting efficiency. A workflow for this stage of validation has been established, as performed for rs4813319 itself (Figure 4.4). This can be further expanded by assessment of gene expression at the transcript level using qRT-PCR, or by assessment of splicing variation at the RNA or protein level using isoform-specific qPCR or Western blot respectively. Alternative screening techniques could be used to provide corroborating data on regulatory function. For example, CRISPR interference, using a functionally inactive Cas9 enzyme together with a repressive domain, can block enhancer activity without the need for nuclease activity.[326] Alternatively, massively parallel reporter assays such as STARR-Seq (self-transcribing active regulatory region sequencing) can be used to identify enhancers and differential effects of genetic variants, although these assays depend on the regulatory element being operational in an artificial system outside their normal cellular and chromosomal context.[327]

Sequence editing for precise validation will pose some practical challenges in these cases. Base editing using well characterised CRISPR/Cas9-linked cytidine deaminase or deoxyadenosine deaminase enzymes is unlikely to be an

option for most of the top candidate variants, either due to the nature of the base change required, or to lack of PAM sequence suitably located to place the target within the required editing window, usually at positions 4 to 8 of the 20mer sgRNA. For example, rs6081134 and rs6045331 would both require a pyrimidine-purine transversion rather than the transitions produced by available enzymes (G to A, A to G, C to T or T to C), although this field is developing rapidly and a number of modified enzymes with different base conversion functions or editing windows have been described recently.[311] Editing by nuclease activity followed by homology-directed repair may be successful but could be limited by the need to introduce two constructs and poor tolerance of plasmid transfection or electroporation in THP-1 cells.

Prime editing, a recently-described CRISPR-based methodology which can effect any single base conversion as well as short insertions and deletions, may be the most promising and flexible technique for validation of candidate variants. Prime editors combine a partial Cas9 protein, with targeting capability and single-strand cutting ('nickase') function, with a reverse transcriptase domain. A prime-editing guide RNA (pegRNA) is used to target the protein to the required site with a 20mer sequence as for sgRNAs in other CRISPR systems, but also includes a template to guide reverse transcription of a short repair template. Finally, a second sgRNA induces nickase activity on the non-edited strand to promote preferential repair with the edited version.[314] Repair efficiency will vary with template design as well as nickase efficiency at the target site, but design tools are becoming increasingly available for this.[328] This technique may also allow editing of candidates that were not easily targetable with local indels induced by conventional spCas9 nuclease activity, as base conversions have been reported up to 29 bp from a PAM sequence.[314]

After replication of the screen to ensure that candidates are robust, single guide and prime editing experiments as above are being planned to validate the effects of the most promising variants on SIRPα expression and splicing.

4.3.3 Advantages and limitations of the targeted CRISPR screening approach

The targeted CRISPR screen design is conceptually attractive as a method of dissecting which of the many variants in an LD block are likely to be contributing to modified regulation of gene expression in a context where gene expression is linked to disease risk. It provides a means of direct perturbation which although less precise than base editing is more scalable, and provides complementary information to other primarily observational data on regulatory elements. While the methodology is similar to previously described non-coding CRISPR screens[316–319], the focus on a set of variants associated with a specific phenotype is a refinement to the regulatory screening concept that has not previously been reported. Limiting the potential targets to a small set, in which prior data suggests a high likelihood that a regulatory element exists that is acutely sensitive to small sequence perturbations, should improve the power and specificity of the screen, reducing the burden of multiple comparisons when screening a large number of irrelevant loci. This focus also aids interpretation of the results - screen hits (if validated) are likely to represent elements for which natural variation has biologically significant consequences.

Enhancers and other putative regulatory features in the genome are often identified by features such as open chromatin, histone acetylation or methylation marks, transcription factor occupancy (especially CTCF) or bidirectional transcription. Conservation across species may also help to identify regulatory sites.[42] Active enhancers are associated with both histone marks H3K27Ac and H3K4me1, whereas 'poised' enhancers tend to have low H3K27Ac but have a trimethylation mark at the corresponding residue (H3K27me3).[54,55] Chromatin marks may also be associated with other regulatory elements such as silencers.[329] These features are in many cases correlative rather than causative, and do not necessarily encompass all regulatory elements in the genome: for example, alternative classes of enhancers have been found with less well characterised histone marks, which are not associated with H3K27ac.[326,330] All these features (except conservation) vary markedly between cell types and between states for a single cell type, and so inference of regulatory element location may be unreliable where data are not available for the sample type of interest. Even where datasets do exist, false positives and false negatives are possible with

a single experiment, and this could explain the lack of open chromatin signal within this locus in a single dataset from THP-1 cells.[322]

Cell type specificity of regulatory elements could also limit the application of this screening method, as it requires the availability of a functionally relevant cell line. Whereas in this case there was a rational basis for selecting a monocytic cell line, in other cases where a gene is expressed in multiple cell types and the mechanistic link with disease is not clear, the appropriate cell line may not be obvious, or a suitable cell line may not exist (especially for non-human studies). Immortalised cell lines may furthermore behave differently from primary cells, which could reduce the sensitivity of the screen even with an apparently appropriate cell population.

As can be seen from Figure 4.13, multiple candidate regulatory elements can be present within a locus of interest, and moreover regions of specific chromatin marks, open chromatin or transcription factor occupancy sites (based on ChIP-Seq) tend to span hundreds of base pairs. Many potentially disease-associated variants within an LD block may thus overlap regulatory elements, with a single regulatory element potentially containing multiple variants, and so although the intersection may be helpful in prioritisation, it is insufficient to implicate specific polymorphisms as causative variants without direct evidence of modulation of gene expression with more precise spatial resolution. The targeted CRISPR mutagenesis screen aims to provide this evidence: in the absence of confounding effects, enrichment of a guide targeting close to a variant within a regulatory element indicates not only that that element is functional within the cell type and condition set studied, but also that the key functional sequence lies within a very short distance (usually up to 10-20 base pairs either side) of the target. This screen design moreover does not inherently assume that the effect of the locus on gene expression is mediated by a single variant. There is evidence that for some diseases such as common inflammatory diseases, multiple enhancer variants in linkage disequilibrium can act in concert to produce the observed effect.[331] This could be reflected by multiple independent signals in regulatory CRISPR screen results. The possibility of multiple causal variants should be taken into account when planning validation of short-listed candidates.

Like all screens, this technique has its limitations, and both false positives and false negatives are likely. Not all sequences can be targeted, either due to lack

of an appropriate PAM sequence or due to excessive homology with other sites in the genome. Limited guide availability within the target sequence surrounding each variant furthermore precludes harmonisation of sgRNA numbers between variants: they will thus vary in power and in susceptibility to experimental noise or off-target effects. Use of an extended PAM range Cas9 variant was designed to ameliorate these problems, but despite the hypothetical benefits, xCas9 did not prove useful in our study due to apparent lack of cutting efficiency.

In contrast to the equal or greater efficiency at NGG PAMs originally reported,[174] a number of other studies have reported reduced efficiency of xCas9 compared to spCas9.[332,333] As well as having a lower basal efficacy, xCas9 is more sensitive to target-sgRNA mismatches (which will reduce off-target effects but also increase the impact of deviations from the reference sequence), and to guide G-prefixing. As an initial guanine is necessary for efficient transcription from a U6 promoter, this is often included in the cloning flank sequences to prefix all 20-mer guides, as was done for our library. xCas9 has however been recently shown to be highly sensitive to mismatches between this 5' extension and the target sequence, reducing average indel frequencies more than two-fold, to less than 10%.[332] There is also evidence that PAM recognition by xCas9 (and indeed by some other Cas9 variants) may depend on more than the three nucleotides generally considered, and it may be necessary to consider potential PAMs of five or more nucleotides when selecting guides. The actual range of PAMs and thus sequences targetable may therefore be more restricted than previously thought. Use of more PAM-permissive Cas9 enzymes (or other Cas variants) could be a promising future improvement to the screening technique, once the targeting specificities of these enzymes become better characterised.

Variations from the reference sequence, as well as causing potential targeting inefficiency that can be prevented with modified guide design, may in some cases limit the expression change observable. In this study, the THP-1 cells used were homozygous for the minor 'risk' allele at the lead variant, but were homozygous for the 'non-risk' allele at other sites. If the index variant itself were the causative variant, the cell population would already be displaying the reduced expression profile associated with that polymorphism. The sensitivity of the screen to detect any further decrease in expression on mutagenesis of this site would then depend on whether the variant completely ablated local regulatory function (e.g. by preventing transcription factor binding), or merely

attenuated it (e.g. by partially reducing transcription factor binding affinity).

This screen design is primarily suitable for detection of regulatory effects on gene expression: since mutagenesis in the vicinity of protein-coding variants is likely to produce truncated or non-functional proteins rather than more precise point mutations, it cannot confirm or rule out a role for these variants. The approach is most applicable either to genetic associations without plausible candidate protein-coding variants in the credible variant set, or, as in this case, where previous experimental evidence has suggested that the protein variants present are unlikely to affect function.

Outcome measures for screening are not limited to protein production: more complex outcome phenotypes can be used, provided they can provide a basis for cell sorting, or for differential growth or survival. For example, if a target locus were implicated in viral replication, alternative outcome measures could include cell survival or production of viral protein, rather than the protein product of the gene itself. This could be especially useful to allow single-step screening of a locus involving multiple candidate genes, as long as an appropriate cellular model can be found. RNA expression could also be used as an alternative to protein expression via the use of reporter constructs such as a fluorescent protein linked to the transcript via an internal ribosome entry site. This approach was considered for the current screen but was not successful due the difficulties in effecting homology-directed repair to create such a construct in THP-1 cells. It will be a viable option in many cells that are more permissive to transfection.

Pleiotropy in source data could complicate both outcome measure selection and interpretation of results, much as it confounds analyses such as Mendelian randomisation. In this case, the documented eQTL at the target locus is pleiotropic for *SIRPA* and *SIRPB1*, with expression effects in opposite directions. It is thus possible that increased production of SIRPβ1 could be the actual outcome of biological interest and the causative factor underlying increased disease risk. This stimulatory receptor, closely related to SIRPα but lacking CD47 binding, has been reported to have increased expression in the prefrontal cortex of patients with schizophrenia, and a copy number variant increasing expression has been implicated in impulsive-disinhibited personality.[300,334] The *SIRPB1* gene is located approximately 300 kbp upstream of *SIRPA* on chromosome 20, and so it is conceivable that a long-range interaction allows a single regulatory element

within *SIRPA* to influence the expression of both genes directly, but it is perhaps more likely that the observed inverse relationship is caused by negative feedback at the protein level. In this case, quantification of either is likely to be an acceptable outcome measure for the screen, but the known roles of SIRPα in synapse development provide a mechanistic link that suggests that, of the two, expression change of SIRPα is more likely to be causal.

Given the GTex evidence that the lead variant is a splicing quantitative trait locus as well as an expression and protein quantitative trait locus, splice variation is one of the possible mechanisms that could link the LD block to disease. This would not be detected by an RNA reporter, but it is also uncertain whether this would be detected by antibody-mediated quantification of protein expression. Sensitivity of the described methodology to detect this depends on whether the isoforms differ in the domain to which the antibody binds, which is currently unknown as detailed epitope mapping has not been performed. The variant is linked to a change in splicing frequency of the third intron (GRCh38 20:1915455-1921395), which is of uncertain significance since none of the major isoforms reported by most sources, including GTex itself, have any variation in this part of the transcript.[64,335] A change in the relative frequency of an isoform with very low abundance is less likely to be biologically relevant than an overall change in gene expression. Although no strong candidate targets were identified at the splice site or splice branch point towards the end of the intron (with the possible exception of one cluster of variants approximately 160 - 180 bp upstream with inconsistent direction of enrichment for different sgRNAs), one of the lead candidates (rs6081134) did correspond to a binding site for the transcription factor RBM22, which is involved in spliceosome binding. An effect on splicing cannot therefore be ruled out, and it will be important to include assessment of differential splicing in downstream validation analyses.

Many of the limitations above can be mitigated either by modifications to the guide design and screening procedures, or by careful selection of the loci to which this methodology is applied.

4.3.4 Contribution to understanding of the genetic basis of schizophrenia

The reported association between SIRPα expression and schizophrenia risk[289] adds to the overall understanding of the complex developmental events that contribute to the pathogenesis of this syndrome. While the effect of SIRPα on synaptic maturation and pruning is likely to contribute to neurologic manifestations, reduced SIRPα could also play a role in some of the many immune derangements seen in some patients with schizophrenia, such as increased circulating concentrations of cytokines such as IL6, or even apparent decreased risk of some cancers.[336–339] Identification of the precise variant underpinning the association, and the processes that this variant disrupts, could give us important additional information about the underlying mechanisms: for example, disruption of RBM22 binding (e.g. by rs6081134) would implicate a role for the spliceosome, and hence suggest that differentially spliced isoforms could be of biological relevance *in vivo*, whereas other transcription factors with binding sites overlapping candidate variants can be induced by other potentially relevant stimuli (e.g. ATF3 by inflammatory cytokines).

Refining an association between schizophrenia and a broad locus to an association with a single variant could also have practical applications in the context of disease prediction. A number of polygenic risk scores have been developed for schizophrenia based on previously reported genetic associations, but these have limited sensitivity and specificity.[293] If proxy variants rather than true causative variants are used to construct these scores, accuracy will be limited by the degree of linkage disequilibrium between the proxy and causative variants. This can moreover vary markedly between ancestry groups. For example, rs6045331, one of the top candidate causative variants in the *SIRPA* locus, is in almost complete linkage disequilibrium in some populations (r^2 = 1.0 in a British population and 0.8 - 1.0 in other European, Asian and South American populations), but is a much weaker proxy in African populations (r^2 = 0.07 - 0.32). Therefore, any genetic risk estimate based on proxies can only be applied with confidence in the population in which it was derived, whereas although effects of specific variants could also vary to an extent between populations due to interaction with other genetic variants, a risk estimate based on causative variants is more likely to have external validity.

While validation of individual results is still pending, targeted screening of a set of disease-associated variants in an LD block has suggested some promising candidates as causal variants for an association between *SIRPA* and schizophrenia. Notwithstanding the limitations of this screening technique, many of which can be mitigated by refinements to the experimental design, this methodology provides a useful option for the interrogation of selected genomic loci where a regulatory variant is suspected. This could have particular applications in functional follow-up of associations from genome-wide association studies, to move our understanding beyond association, towards causation.

Chapter 5

Conclusions and future directions

5.1 Conclusions: how do these findings affect our understanding of causation for the observed genetic associations?

In this thesis, two contrasting approaches have been employed to address questions of causation and biological relevance underlying novel genetic associations between variants in immune system-associated genes and human disease.

First, I used an animal model to address causation at the whole gene and protein level, to assess whether presence or absence of the *CD97* gene product could cause a substantial change in the phenotype of influenza-associated disease. This is essential to our understanding of whether a genetic variant in this gene, whether causing a change in expression in either direction or a change in protein function, is plausible as a causal factor in the genetic association observed with naturally-occurring severe disease.

Our results showed a modest, transient effect on disease severity, but a more robust effect of CD97 deficiency on the balance of the T-cell response to viral challenge, with reduced efficiency of the $CD8^+$ response and reduced viral clearance. Notwithstanding the limitations of the murine model to recapitulate human severe disease, this shows that the gene can have an impact on an as-

pect of the adaptive immune response that is relevant to influenza and possibly to other conditions where an efficient $CD8^+$ T-cell response is critical, such as in the immune response to tumour cells. The evidence remains equivocal, however, as to whether the magnitude of this effect would be sufficient for CD97 deficiency to explain the apparent risk of severe influenza in patients carrying the rs2302092 G allele, or even whether reduced T-cell infiltration associated with reduced CD97 could in fact reduce tissue injury in later stages of the disease. Further work will be required to establish the molecular mechanisms underlying the effect on the T-cell response.

Secondly, I developed a method for screening for potential causative regulatory elements in a linkage disequilibrium block by targeted CRISPR/Cas9 mutagenesis of multiple non-coding variants in parallel, and applied this to a locus associated both with schizophrenia and with differential expression of *SIRPA*. Some potential limitations to the technique were identified, such as artefact from inadvertent excision of large segments of DNA, which can be improved upon with optimised experimental design in future applications. This approach was able to identify plausible candidate regulatory variants in the *SIRPA* locus, both in recognised and potential novel regulatory regions, where top candidate variants coincided with binding sites for transcription factors (e.g. ATF2, 3 and 7 for rs6045331) or splicing factors (e.g. RBM22 for rs6081134) with potential roles in regulation of gene expression and processing. While a single causal variant has not yet been identified, we now have a shortlist of prioritised candidates to take forward for validation with more precise gene editing techniques. Together with existing data on the association between SIRPα in schizophrenia and the likely roles of SIRPα in synaptic development, the evidence for a causative role for genetic variation in this gene on schizophrenia development is becoming more and more compelling. While it is difficult to quantify the exact extent of the contribution that this makes to a multifactorial disease, establishing a precise causal variant (and the regulatory processes which it perturbs) will provide an important insight into disease pathogenesis.

Figure 5.1 shows how our data and existing data contribute to our current understanding of potential causative pathways linking single variants in *CD97* and *SIRPA* with intermediate biological processes (T-cell regulation or synaptic development) and disease phenotypes, according to the integrated framework for observational and experimental data proposed in section 1.3.4 and Figure 1.3.

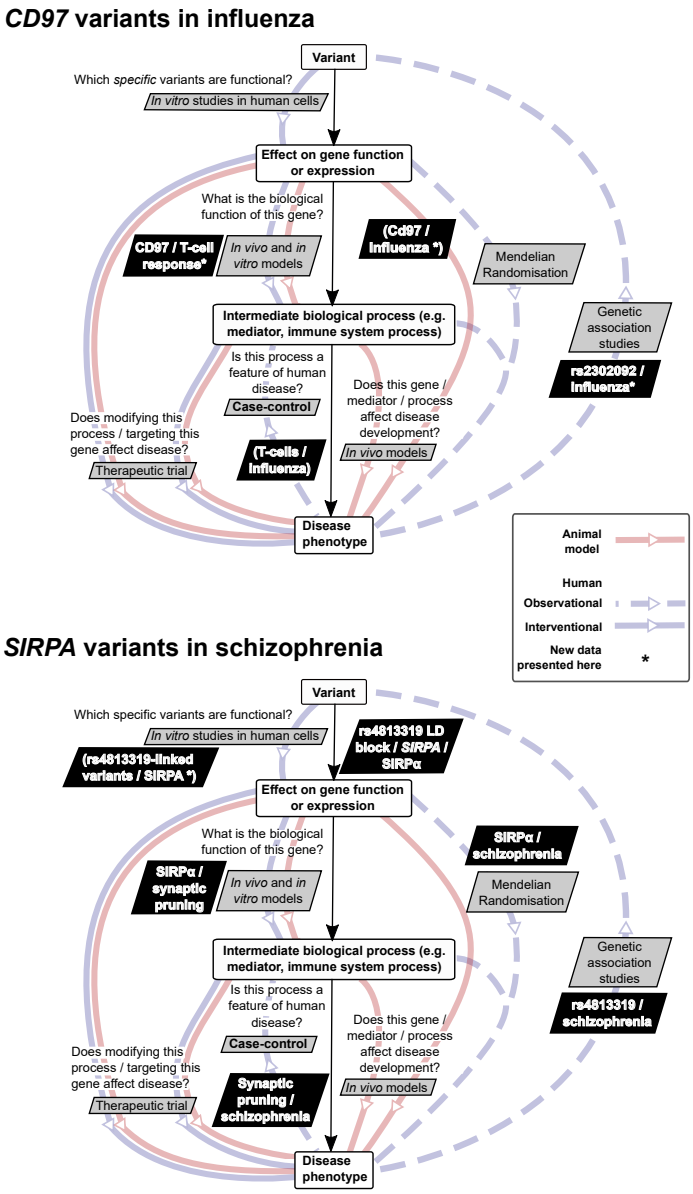

Figure 5.1 – Current state of knowledge on causative links between a *CD97* variant and severe influenza, and between a *SIRPA* variant and schizophrenia. Established links between the index genetic variants, their gene products, intermediate biological processes and disease phenotypes are shown, based on new data described in this thesis (*), interrogating the specific variants and biological processes involved in a putative mechanistic association, or on pre-existing data. Parentheses indicate incomplete or equivocal data.

5.2 Wider relevance of findings and techniques: applications to COVID-19 research

As the COVID-19 pandemic is one of the greatest global public health challenges currently facing us, the interpretation of the functional basis of genetic associations with severe COVID-19 is one of the most pressing challenges in functional genomics. We have access to an abundance of high quality genetic data unprecedented for human critical illness, and as there is intense interest and investment in research into urgently needed treatments for severe cases, we have the opportunity to harness genetic insights to rationalise therapeutic interventions. There are already real examples in which genetic association data could help inform selection of drugs for therapeutic trials: for example, associations with interferon receptors genes (*IFNAR2*) or interferon-simulated genes (*OAS1/2/3*) reinforce interest in interferon therapy, which is currently the subject of a number of clinical trials[340], while associations with *TYK2* and *TMPRSS2* warrant further investigation of these as potential therapeutic targets. The proteins encoded by these latter two genes could be of particular interest given that inhibitors are available as authorised medications for other conditions in some countries, and could be repurposed - TMPRSS2 inhibitor camostat mesilate is used for chronic pancreatitis, and inhibitors of JAK/TYK2 signalling such as baricitinib are in use for rheumatoid arthritis, with more specific drugs in clinical trials.[9,117]

There is little evidence that the variants or genes studied in this thesis contribute to COVID-19 severity. In a large GWAS of severe COVID-19, *SIRPA* variant rs4813319 was not associated with severe disease. While two intronic SNPs in *CD97* had nominally significant p-values for association ($p < 0.05$), including one at a CTCF binding site (but not including rs2302092, which was not genotyped and cannot be reliably imputed), effect sizes were very small, and this weak evidence would not survive correction for multiple comparisons. Regulation of the T-cell response by CD97 may well be relevant to the efficiency of SARS-CoV2 clearance, but the evidence does not support it as a promising therapeutic target. However, the approaches taken here are entirely applicable to COVID-associated loci. A particular case in point is the locus on chromosome 3 which has the strongest and most replicable association with severe disease, a locus containing a number of chemokine receptor genes as well as others

such as *FYCO1* and *LZTFL1* (see Figure 5.2).[9,341] Lead variant rs73064425 is in linkage disequilibrium with over 500 variants (at a threshold of $r^2 > 0.2$), spanning approximately 10 protein-coding genes. Any one of these genes could plausibly modify disease severity, and while there are only a small number of non-synonymous protein-coding variants in the candidate set (in the *FYCO1* gene), there are a large number of variants within known candidate cis-acting regulatory elements, which could plausibly be causal regulatory variants. While the effect size of the genetic association with this locus makes it the highest priority for further follow-up, the number of candidate genes and variants in the locus also makes it the most difficult to dissect.

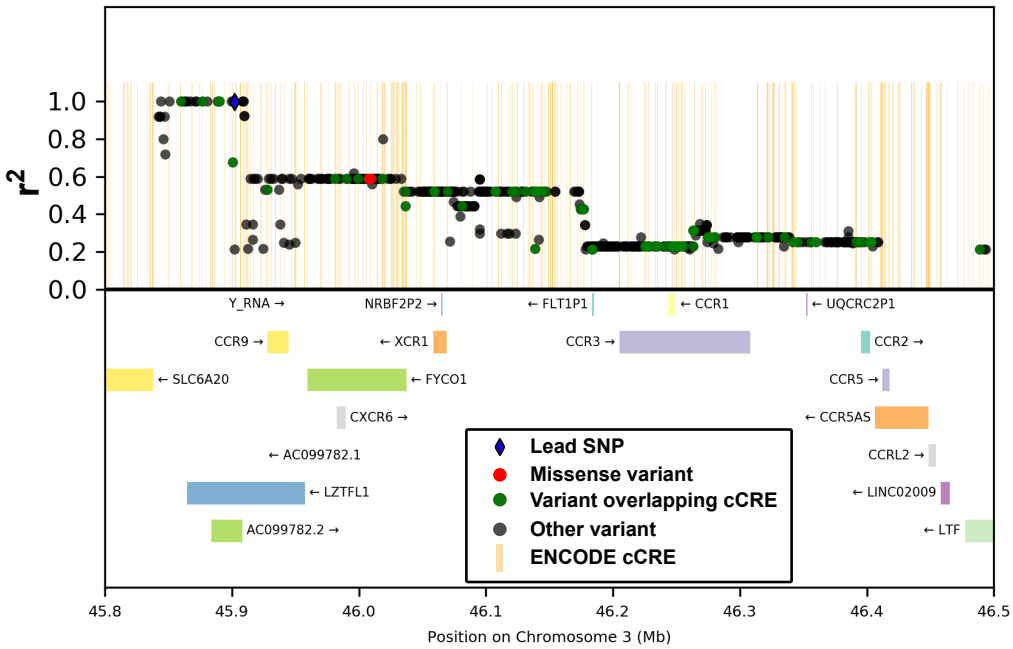

Figure 5.2 – Features of a linkage disequilibrium block in chromosome 3 associated with severe COVID-19 The locus plot shows all variants in linkage disequilibrium ($r^2 > 0.2$) with lead variant rs73064425, relative to the position of GENCODE gene annotations and known cis-acting candidate regulatory elements (cCREs) from ENCODE. Coordinates are based on genome assembly GRCh37.

The majority of genes in the chromosome 3 locus are involved in immune responses, including chemokine-induced leukocyte migration. Accurate modelling of their role in the host response to SARS-CoV2 infection could be difficult *in vitro*, and so one approach to investigation of the locus would be to test the

function of candidate genes in *in vivo* infection models in knockout mice, as has been done here for CD97. In addition to the usual limitations of an animal model, such as species variation in gene expression or incomplete recapitulation of human disease, this approach would have further limitations in this case. Firstly, the sheer number of candidate genes make an *in vivo* approach inefficient and undesirable from a 'reduction, refinement and replacement' perspective until a smaller selection of candidate genes can be prioritised. Secondly, unless suitable knockout mouse lines are already being bred, the time involved in creating, back-crossing and breeding a strain makes this approach less than ideal during a public health emergency.

Linkage disequilibrium block regulatory screening, as performed for the *SIRPA* locus in chapter 4, is an attractive alternative for interrogation of this locus. By targeting the set of candidate variants using CRISPR/Cas9, we will be able to establish which variants are located in regulatory elements that effect the expression of each target gene examined. While in this case each target gene will need to be assessed separately, the technique can be efficiently scaled to assess multiple targets. As with the *SIRPA* screen, this will allow us to prioritise candidate causal variants to validate individually (for example, via base editing or reporter assays). If we can demonstrate that one or more of the disease-associated variants influences expression of one of the candidate genes in the locus, that target gene can be prioritised for further mechanistic investigations. There are other examples where a functional phenotype, rather than gene expression, can be used as the outcome measure for this technique. For example, the association between severe disease and the locus containing the interferon-stimulated genes *OAS1*, *OAS2* and *OAS3* is almost certainly mediated by impaired viral restriction via these genes. The screen could therefore be adapted to assess the impact of variant-targeted mutagenesis on measures such as viral protein expression or survival in *in vitro* infection models. This could complement other more traditional methods of distinguishing between the genes, such as whole-gene knockout.

I have designed a pool of guides to target the COVID-associated variants in the chromosome 3 locus, using the modified design algorithm informed by the pilot experiments reported in this thesis. Work is underway to screen for regulatory elements in the locus, initially focusing on the chemokine receptor genes in a monocytic cell line. If successful, this extension of the methodology developed

in the program of work reported here has the potential to clarify a question of critical importance to global health, which has so far remained unsolved since this association was first reported in July 2020. Only with such mechanistic insight can we hope to leverage the biological clues provided by genetic associations for therapeutic gain.

References

[1] Mervyn Singer et al. “The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)”. In: *JAMA* 315.8 (2016), pages 801–810. DOI: 10.1001/jama.2016.0287.

[2] E. Bautista et al. “Clinical aspects of pandemic 2009 influenza A (H1N1) virus infection”. In: *N Engl J Med* 362.18 (2010), pages 1708–19. DOI: 10.1056/NEJMra1000449.

[3] R. Snacken, C. Quinten, I. Devaux, F. Plata, E. Broberg, P. Zucs, and A. Amato-Gauci. “Surveillance of hospitalised severe cases of influenza A(H1N1)pdm09 and related fatalities in nine EU countries in 2010-2011”. In: *Influenza Other Respir Viruses* 6.6 (2012), e93–6. DOI: 10.1111/j.1750-2659.2012.00406.x.

[4] Annemarie B Docherty et al. “Features of 20,133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study”. In: *BMJ* 369 (2020), page m1985. DOI: 10.1136/bmj.m1985.

[5] Thorkild I.A. Sørensen, Gert G. Nielsen, Per Kragh Andersen, and Thomas W. Teasdale. “Genetic and Environmental Influences on Premature Death in Adult Adoptees”. In: *New England Journal of Medicine* 318.12 (1988), pages 727–732. DOI: 10.1056/nejm198803243181202.

[6] F. S. Albright, P. Orlando, A. T. Pavia, G. G. Jackson, and L. A. Cannon Albright. “Evidence for a heritable predisposition to death due to influenza”. In: *J Infect Dis* 197.1 (2008), pages 18–24. DOI: 10.1086/524064.

[7] G. S. Cooke and A. V. Hill. “Genetics of susceptibility to human infectious disease”. In: *Nat Rev Genet* 2.12 (2001), pages 967–77. DOI: 10.1038/35103577.

[8] Frances M. K. Williams et al. “Self-Reported Symptoms of COVID-19, Including Symptoms Most Predictive of SARS-CoV-2 Infection, Are Heritable”. In: *Twin Research and Human Genetics* 23.6 (2021), pages 316–321. DOI: 10.1017/thg.2020.85.

[9] Erola Pairo-Castineira et al. “Genetic mechanisms of critical illness in Covid-19”. In: *Nature* (2020). DOI: 10.1038/s41586-020-03065-y.

[10] Qian Zhang et al. “Inborn errors of type I IFN immunity in patients with life-threatening COVID-19”. In: *Science* 370.6515 (2020), eabd4570. DOI: 10.1126/science.abd4570.

[11] M. Gottfredsson et al. “Lessons from the past: familial aggregation analysis of fatal pandemic influenza (Spanish flu) in Iceland in 1918”. In: *Proc Natl Acad Sci U S A* 105.4 (2008), pages 1303–8. DOI: 10.1073/pnas.0707659105.

[12] P. Horby, N. Y. Nguyen, S. J. Dunstan, and J. K. Baillie. “The role of host genetics in susceptibility to influenza: a systematic review”. In: *PLoS One* 7.3 (2012), e33180. DOI: 10.1371/journal.pone.0033180.

[13] Dimitrios P. Bogdanos, Daniel S. Smyk, Eirini I. Rigopoulou, Maria G. Mytilinaiou, Michael A. Heneghan, Carlo Selmi, and M. Eric Gershwin. “Twin studies in autoimmune disease: Genetics, gender and environment”. In: *Journal of Autoimmunity* 38.2 (2012), J156–J169. DOI: https://doi.org/10.1016/j.jaut.2011.11.003.

[14] Michael F. Seldin. “The genetics of human autoimmune disease: A perspective on progress in the field and future directions”. In: *Journal of autoimmunity* 64 (2015), pages 1–12. DOI: 10.1016/j.jaut.2015.08.015.

[15] Massimo Mangino, Mario Roederer, Margaret H. Beddall, Frank O. Nestle, and Tim D. Spector. “Innate and adaptive immune traits are differentially affected by genetic and environmental factors”. In: *Nature communications* 8 (2017), pages 13850–13850. DOI: 10.1038/ncomms13850.

[16] M. Roederer et al. “The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis”. In: *Cell* 161.2 (2015), pages 387–403. DOI: 10.1016/j.cell.2015.02.046.

[17] Petter Brodin et al. “Variation in the human immune system is largely driven by non-heritable influences”. In: *Cell* 160.1-2 (2015), pages 37–47. DOI: 10.1016/j.cell.2014.12.020.

[18] J. K. Baillie. “Translational genomics. Targeting the host immune response to fight infection”. In: *Science* 344.6186 (2014), pages 807–8. DOI: 10.1126/science.1255074.

[19] David A. Hume. “Macrophages as APC and the Dendritic Cell Myth”. In: *The Journal of Immunology* 181.9 (2008), pages 5829–5835. DOI: 10.4049/jimmunol.181.9.5829.

[20] T. Kawai and S. Akira. “The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors”. In: *Nat Immunol* 11.5 (2010), pages 373–84. DOI: 10.1038/ni.1863.

[21] H. H. Lin, D. E. Faunce, M. Stacey, A. Terajewicz, T. Nakamura, J. Zhang-Hoover, M. Kerley, M. L. Mucenski, S. Gordon, and J. Stein-Streilein. "The macrophage F4/80 receptor is required for the induction of antigen-specific efferent regulatory T cells in peripheral tolerance". In: *J Exp Med* 201.10 (2005), pages 1615–25. DOI: 10.1084/jem.20042307.

[22] Peter J. Murray. "Macrophage Polarization". In: *Annual Review of Physiology* 79.1 (2017), pages 541–566. DOI: 10.1146/annurev-physiol-022516-034339.

[23] J. K. Baillie et al. "Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new insights into genetic aetiology of inflammatory bowel disease". In: *PLoS Genet* 13.3 (2017), e1006641. DOI: 10.1371/journal.pgen.1006641.

[24] J. Kenneth Baillie et al. "Shared activity patterns arising at genetic susceptibility loci reveal underlying genomic and cellular architecture of human disease". In: *PLoS computational biology* 14.3 (2018), e1005934–e1005934. DOI: 10.1371/journal.pcbi.1005934.

[25] Laramie E. Duncan, Michael Ostacher, and Jacob Ballon. "How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete". In: *Neuropsychopharmacology* 44.9 (2019), pages 1518–1523. DOI: 10.1038/s41386-019-0389-5.

[26] J. B. Pingault, P. F. O'Reilly, T. Schoeler, G. B. Ploubidis, F. Rijsdijk, and F. Dudbridge. "Using genetic data to strengthen causal inference in observational research". In: *Nat Rev Genet* 19.9 (2018), pages 566–580. DOI: 10.1038/s41576-018-0020-3.

[27] Helena Furberg et al. "Genome-wide meta-analyses identify multiple loci associated with smoking behavior". In: *Nature Genetics* 42.5 (2010), pages 441–447. DOI: 10.1038/ng.571.

[28] Yohan Bossé and Christopher I. Amos. "A Decade of GWAS Results in Lung Cancer". In: *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 27.4 (2018), pages 363–379. DOI: 10.1158/1055-9965.EPI-16-0794.

[29] Donglei Hu and Elad Ziv. "Confounding in Genetic Association Studies and Its Solutions". In: *Pharmacogenomics in Drug Discovery and Development: From Bench to Bedside*. Edited by Qing Yan. Totowa, NJ: Humana Press, 2008, pages 31–39. ISBN: 978-1-59745-205-2. DOI: 10.1007/978-1-59745-205-2_3.

[30] Lotfi Slim, Clément Chatelain, Chloé-Agathe Azencott, and Jean-Philippe Vert. "Novel methods for epistasis detection in genome-wide association studies". In: *PLOS ONE* 15.11 (2020), e0242927. DOI: 10.1371/journal.pone.0242927.

[31] N. R. Wray. “Allele frequencies and the r2 measure of linkage disequilibrium: impact on design and interpretation of association studies”. In: *Twin Res Hum Genet* 8.2 (2005), pages 87–94. DOI: 10.1375/1832427053738827.

[32] Kui Zhang, Peter Calabrese, Magnus Nordborg, and Fengzhu Sun. “Haplotype Block Structure and Its Applications to Association Studies: Power and Study Designs”. In: *The American Journal of Human Genetics* 71.6 (2002), pages 1386–1394. DOI: https://doi.org/10.1086/344780.

[33] Daniel J. Schaid, Wenan Chen, and Nicholas B. Larson. “From genome-wide associations to candidate causal variants by statistical fine-mapping”. In: *Nature reviews. Genetics* 19.8 (2018), pages 491–504. DOI: 10.1038/s41576-018-0016-z.

[34] J. B. Maller et al. “Bayesian refinement of association signals for 14 loci in 3 common diseases”. In: *Nat Genet* 44.12 (2012), pages 1294–301. DOI: 10.1038/ng.2435.

[35] Stephen Burgess, Robert A. Scott, Nicholas J. Timpson, George Davey Smith, Simon G. Thompson, and Epic- InterAct Consortium. “Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors”. In: *European journal of epidemiology* 30.7 (2015), pages 543–552. DOI: 10.1007/s10654-015-0011-z.

[36] B. A. Ference, W. Yoo, I. Alesh, N. Mahajan, K. K. Mirowska, A. Mewada, J. Kahn, L. Afonso, Sr. Williams K. A., and J. M. Flack. “Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis”. In: *J Am Coll Cardiol* 60.25 (2012), pages 2631–9. DOI: 10.1016/j.jacc.2012.09.017.

[37] Luke J. O’Connor and Alkes L. Price. “Distinguishing genetic correlation from causation across 52 diseases and complex traits”. In: *Nature Genetics* 50.12 (2018), pages 1728–1734. DOI: 10.1038/s41588-018-0255-0.

[38] Andrei S. Rodin and Eric Boerwinkle. “Mining genetic epidemiology data with Bayesian networks I: Bayesian networks and example application (plasma apoE levels)”. In: *Bioinformatics* 21.15 (2005), pages 3273–3278. DOI: 10.1093/bioinformatics/bti505.

[39] Richard Howey, So-Youn Shin, Caroline Relton, George Davey Smith, and Heather J. Cordell. “Bayesian network analysis incorporating genetic anchors complements conventional Mendelian randomization approaches for exploratory analysis of causal relationships in complex data”. In: *PLOS Genetics* 16.3 (2020), e1008198. DOI: 10.1371/journal.pgen.1008198.

[40] Ali J. Marian. “Causality in genetics: the gradient of genetic effects and back to Koch’s postulates of causality”. In: *Circulation research* 114.2 (2014), e18–e21. DOI: 10.1161/CIRCRESAHA.114.302904.

[41] S. N. Chen, G. Czernuszewicz, Y. Tan, R. Lombardi, J. Jin, J. T. Willerson, and A. J. Marian. “Human molecular genetic and functional studies identify TRIM63, encoding Muscle RING Finger Protein 1, as a novel gene for human hypertrophic cardiomyopathy”. In: *Circ Res* 111.7 (2012), pages 907–19. DOI: 10.1161/circresaha.112.270207.

[42] M. Hemberg, J. M. Gray, N. Cloonan, S. Kuersten, S. Grimmond, M. E. Greenberg, and G. Kreiman. “Integrated genome analysis suggests that most conserved non-coding sequences are regulatory factor binding sites”. In: *Nucleic Acids Res* 40.16 (2012), pages 7858–69. DOI: 10.1093/nar/gks477.

[43] E. A. Boyle, Y. I. Li, and J. K. Pritchard. “An Expanded View of Complex Traits: From Polygenic to Omnigenic”. In: *Cell* 169.7 (2017), pages 1177–1186. DOI: 10.1016/j.cell.2017.05.038.

[44] R. C. Eisensmith and S. L. Woo. “Molecular basis of phenylketonuria and related hyperphenylalaninemias: mutations and polymorphisms in the human phenylalanine hydroxylase gene”. In: *Hum Mutat* 1.1 (1992), pages 13–23. DOI: 10.1002/humu.1380010104.

[45] S. J. Spier and E. P. Hoffman. “Hyperkalaemic periodic paralysis: Mother nature versus human nature”. In: *Equine Veterinary Education* 20.8 (2008), pages 401–405. DOI: https://doi.org/10.2746/095777308X334482.

[46] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits”. In: *Proc Natl Acad Sci U S A* 106.23 (2009), pages 9362–7. DOI: 10.1073/pnas.0903103106.

[47] Matthew T. Maurano et al. “Systematic localization of common disease-associated variation in regulatory DNA”. In: *Science (New York, N.Y.)* 337.6099 (2012), pages 1190–1195. DOI: 10.1126/science.1222794.

[48] Y. Han et al. “Integration of multiethnic fine-mapping and genomic annotation to prioritize candidate functional SNPs at prostate cancer susceptibility regions”. In: *Hum Mol Genet* 24.19 (2015), pages 5603–18. DOI: 10.1093/hmg/ddv269.

[49] Marc A. Schaub, Alan P. Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. “Linking disease associations with regulatory information in the human genome”. In: *Genome Research* 22.9 (2012), pages 1748–1759. DOI: 10.1101/gr.136127.111.

[50] Fantom Consortium The, Riken Pmi the, and Clst. “A promoter-level mammalian expression atlas”. In: *Nature* 507.7493 (2014), pages 462–470. DOI: 10.1038/nature13182http://www.nature.com/nature/journal/v507/n7493/abs/nature13182.html#supplementary-information.

[51] G. A. Maston, S. K. Evans, and M. R. Green. “Transcriptional regulatory elements in the human genome”. In: *Annu Rev Genomics Hum Genet* 7 (2006), pages 29–59. DOI: 10.1146/annurev.genom.7.080505.115623.

[52] Jennifer L Plank and Ann Dean. “Enhancer Function: Mechanistic and Genome-Wide Insights Come Together”. In: *Molecular Cell* 55.1 (2014), pages 5–14. DOI: 10.1016/j.molcel.2014.06.015.

[53] Stefan Schoenfelder and Peter Fraser. “Long-range enhancer–promoter contacts in gene expression control”. In: *Nature Reviews Genetics* 20.8 (2019), pages 437–455. DOI: 10.1038/s41576-019-0128-0.

[54] Menno P. Creyghton et al. “Histone H3K27ac separates active from poised enhancers and predicts developmental state”. In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pages 21931–21936. DOI: 10.1073/pnas.1016071107.

[55] A. Rada-Iglesias, R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka. “A unique chromatin signature uncovers early developmental enhancers in humans”. In: *Nature* 470.7333 (2011), pages 279–83. DOI: 10.1038/nature09692.

[56] R. E. Thurman et al. “The accessible chromatin landscape of the human genome”. In: *Nature* 489.7414 (2012), pages 75–82. DOI: 10.1038/nature11232.

[57] Scott Smemo et al. “Obesity-associated variants within FTO form long-range functional connections with IRX3”. In: *Nature* 507.7492 (2014), pages 371–375. DOI: 10.1038/nature13138.

[58] Petros Kolovos, Tobias A. Knoch, Frank G. Grosveld, Peter R. Cook, and Argyris Papantonis. “Enhancers and silencers: an integrated and simple model for their function”. In: *Epigenetics & Chromatin* 5.1 (2012), page 1. DOI: 10.1186/1756-8935-5-1.

[59] Stephen S. Gisselbrecht, Alexandre Palagi, Jesse V. Kurland, Julia M. Rogers, Hakan Ozadam, Ye Zhan, Job Dekker, and Martha L. Bulyk. “Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts”. In: *Molecular Cell* 77.2 (2020), 324–337.e8. DOI: https://doi.org/10.1016/j.molcel.2019.10.004.

[60] Yichao Cai et al. “H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions”. In: *Nature Communications* 12.1 (2021), page 719. DOI: 10.1038/s41467-021-20940-y.

[61] Juliet D French et al. “Functional Variants at the 11q13 Risk Locus for Breast Cancer Regulate Cyclin D1 Expression through Long-Range Enhancers”. In: *The American Journal of Human Genetics* 92.4 (2013), pages 489–503. DOI: https://doi.org/10.1016/j.ajhg.2013.01.002.

[62] Zhilian Jia, Jingwei Li, Xiao Ge, Yonghu Wu, Ya Guo, and Qiang Wu. “Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection”. In: *Genome Biology* 21.1 (2020), page 75. DOI: 10.1186/s13059-020-01984-7.

[63] J. Demars et al. “Analysis of the IGF2/H19 imprinting control region uncovers new genetic defects, including mutations of OCT-binding sequences, in patients with 11p15 fetal growth disorders”. In: *Hum Mol Genet* 19.5 (2010), pages 803–14. DOI: 10.1093/hmg/ddp549.

[64] François Aguet et al. “Genetic effects on gene expression across human tissues”. In: *Nature* 550.7675 (2017), pages 204–213. DOI: 10.1038/nature24277.

[65] Farhad Hormozdiari, Martijn van de Bunt, Ayellet V Segrè, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. “Colocalization of GWAS and eQTL Signals Detects Target Genes”. In: *The American Journal of Human Genetics* 99.6 (2016), pages 1245–1260. DOI: https://doi.org/10.1016/j.ajhg.2016.10.003.

[66] Andrew D. Johnston, Claudia A. Simões-Pires, Taylor V. Thompson, Masako Suzuki, and John M. Greally. “Functional genetic variants can mediate their regulatory effects through alteration of transcription factor binding”. In: *Nature Communications* 10.1 (2019), page 3472. DOI: 10.1038/s41467-019-11412-5.

[67] Bo Li et al. “Genome-wide CRISPR screen identifies host dependency factors for influenza A virus infection”. In: *Nature Communications* 11.1 (2020), page 164. DOI: 10.1038/s41467-019-13965-x.

[68] Nicholas Parkinson et al. “Dynamic data-driven meta-analysis for prioritisation of host genes implicated in COVID-19”. In: *Scientific Reports* 10.1 (2020), page 22303. DOI: 10.1038/s41598-020-79033-3.

[69] J. Han, J. T. Perez, C. Chen, Y. Li, A. Benitez, M. Kandasamy, Y. Lee, J. Andrade, B. tenOever, and B. Manicassamy. “Genome-wide CRISPR/-Cas9 Screen Identifies Host Factors Essential for Influenza Virus Replication”. In: *Cell Rep* 23.2 (2018), pages 596–607. DOI: 10.1016/j.celrep.2018.03.045.

[70] J. Zhou et al. “A functional variation in CD55 increases the severity of 2009 pandemic H1N1 influenza A virus infection”. In: *J Infect Dis* 206.4 (2012), pages 495–503. DOI: 10.1093/infdis/jis378.

[71] J. Zuniga et al. “Genetic variants associated with severe pneumonia in A/H1N1 influenza infection”. In: *Eur Respir J* 39.3 (2012), pages 604–10. DOI: 10.1183/09031936.00020611.

[72] Y. Chen et al. “Functional variants regulating LGALS1 (Galectin 1) expression affect human susceptibility to influenza A(H7N9)”. In: *Sci Rep* 5 (2015), page 8517. DOI: 10.1038/srep08517.

[73] K. Garcia-Etxebarria et al. “No Major Host Genetic Risk Factor Contributed to A(H1N1)2009 Influenza Severity”. In: *PLoS One* 10.9 (2015), e0135983. DOI: 10.1371/journal.pone.0135983.

[74] A. R. Everitt et al. “IFITM3 restricts the morbidity and mortality associated with influenza”. In: *Nature* 484.7395 (2012), pages 519–23. DOI: 10.1038/nature10921.

[75] S. S. Prabhu, T. T. Chakraborty, N. Kumar, and I. Banerjee. “Association between IFITM3 rs12252 polymorphism and influenza susceptibility and severity: A meta-analysis”. In: *Gene* 674 (2018), pages 70–79. DOI: 10.1016/j.gene.2018.06.070.

[76] E. Kaitlynn Allen et al. “SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans”. In: *Nature medicine* 23.8 (2017), pages 975–983. DOI: 10.1038/nm.4370.

[77] Jie Zhou et al. “Identification and characterization of GLDC as host susceptibility gene to severe influenza”. In: *EMBO Molecular Medicine* 11.1 (2019), e9528. DOI: https://doi.org/10.15252/emmm.201809528.

[78] Zhongshan Cheng et al. “Identification of TMPRSS2 as a Susceptibility Gene for Severe 2009 Pandemic A(H1N1) Influenza and A(H7N9) Influenza”. In: *The Journal of Infectious Diseases* 212.8 (2015), pages 1214–1221. DOI: 10.1093/infdis/jiv246.

[79] A. Maestri, V. A. Sortica, L. Tovo-Rodrigues, M. C. Santos, L. Barbagelata, M. R. Moraes, W. Alencar de Mello, L. Gusmão, R. C. Sousa, and S. Emanuel Batista Dos Santos. “Siaα2-3Galβ1- Receptor Genetic Variants Are Associated with Influenza A(H1N1)pdm09 Severity”. In: *PLoS One* 10.10 (2015), e0139681. DOI: 10.1371/journal.pone.0139681.

[80] S. E. Jørgensen, M. Christiansen, L. B. Ryø, H. H. Gad, J. Gjedsted, P. Staeheli, J. G. Mikkelsen, M. Storgaard, R. Hartmann, and T. H. Mogensen. “Defective RNA sensing by RIG-I in severe influenza virus infection”. In: *Clinical and experimental immunology* 192.3 (2018), pages 366–376. DOI: 10.1111/cei.13120.

[81] F. Hidaka, S. Matsuo, T. Muta, K. Takeshige, T. Mizukami, and H. Nunoi. “A missense mutation of the Toll-like receptor 3 gene in a patient with influenza-associated encephalopathy”. In: *Clin Immunol* 119.2 (2006), pages 188–94. DOI: 10.1016/j.clim.2006.01.005.

[82] H. K. Lim et al. “Severe influenza pneumonitis in children with inherited TLR3 deficiency”. In: *J Exp Med* 216.9 (2019), pages 2038–2056. DOI: 10.1084/jem.20181621.

[83] N. Lee et al. “IFITM3, TLR3, and CD55 Gene SNPs and Cumulative Genetic Risks for Severe Outcomes in Chinese Patients With H7N9/H1N1pdm09 Influenza”. In: *J Infect Dis* 216.1 (2017), pages 97–104. DOI: 10.1093/infdis/jix235.

[84] Nataliia O. Pryimenko, Tetiana M. Kotelevska, Tetiana I. Koval, Liudmyla M. Syzova, Halyna M. Dubynska, and Igor P Kaidashev. “Genetic polymorphism ARG753GLN of TLR-2, LEU412PHE of TLR-3, ASP299GLY of TLR-4 in patients with influenza and influenza-associated pneumonia”. In: *Wiadomosci lekarskie (Warsaw, Poland : 1960)* 72.12 cz 1 (2019), pages 2324–2328.

[85] Michael J. Ciancanelli et al. “Life-threatening influenza and impaired interferon amplification in human IRF7 deficiency”. In: *Science (New York, N.Y.)* 348.6233 (2015), pages 448–453. DOI: 10.1126/science.aaa1578.

[86] N. Hernandez et al. “Life-threatening influenza pneumonitis in a child with inherited IRF9 deficiency”. In: *J Exp Med* 215.10 (2018), pages 2567–2585. DOI: 10.1084/jem.20180628.

[87] María Bravo García-Morato et al. “Impaired control of multiple viral infections in a family with complete IRF9 deficiency”. In: *Journal of Allergy and Clinical Immunology* 144.1 (2019), 309–312.e10. DOI: 10.1016/j.jaci.2019.02.019.

[88] T. C. Carter et al. “Pilot screening study of targeted genetic polymorphisms for association with seasonal influenza hospital admission”. In: *J Med Virol* 90.3 (2018), pages 436–446. DOI: 10.1002/jmv.24975.

[89] Baihui Zhao et al. “Novel susceptibility loci for A(H7N9) infection identified by next generation sequencing and functional analysis”. In: *Scientific Reports* 10.1 (2020), page 11768. DOI: 10.1038/s41598-020-68675-y.

[90] N. Schmitz, M. Kurrer, M. F. Bachmann, and M. Kopf. “Interleukin-1 is responsible for acute lung immunopathology but increases survival of respiratory influenza virus infection”. In: *J Virol* 79.10 (2005), pages 6441–8. DOI: 10.1128/jvi.79.10.6441-6448.2005.

[91] Y. Liu, S. Li, G. Zhang, G. Nie, Z. Meng, D. Mao, C. Chen, X. Chen, B. Zhou, and G. Zeng. “Genetic variants in IL1A and IL1B contribute to the susceptibility to 2009 pandemic H1N1 influenza A virus”. In: *BMC Immunol* 14 (2013), page 37. DOI: 10.1186/1471-2172-14-37.

[92] R. A. García-Ramírez, A. Ramírez-Venegas, R. Quintana-Carrillo, E. Camarena Á, R. Falfán-Valencia, and J. M. Mejía-Aranguré. “TNF, IL6, and IL1B Polymorphisms Are Associated with Severe Influenza A (H1N1) Virus Infection in the Mexican Population”. In: *PLoS One* 10.12 (2015), e0144832. DOI: 10.1371/journal.pone.0144832.

[93] D. Damjanovic, M. Divangahi, K. Kugathasan, C. L. Small, A. Zganiacz, E. G. Brown, C. M. Hogaboam, J. Gauldie, and Z. Xing. “Negative regulation of lung inflammation and immunopathology by TNF-α during acute influenza infection”. In: *Am J Pathol* 179.6 (2011), pages 2963–76. DOI: 10.1016/j.ajpath.2011.09.003.

[94] J. M. Ferdinands, A. M. Denison, N. F. Dowling, H. A. Jost, M. L. Gwinn, L. Liu, S. R. Zaki, and D. K. Shay. “A pilot study of host genetic variants associated with influenza-associated deaths among children and young adults”. In: *Emerg Infect Dis* 17.12 (2011), pages 2294–302. DOI: 10.3201/eid1712.111002.

[95] A. Antonopoulou et al. “Role of tumor necrosis factor gene single nucleotide polymorphisms in the natural course of 2009 influenza A H1N1 virus infection”. In: *Int J Infect Dis* 16.3 (2012), e204–8. DOI: 10.1016/j.ijid.2011.11.012.

[96] J. Martinez-Ocaña et al. “Plasma cytokine levels and cytokine gene polymorphisms in Mexican patients during the influenza pandemic A(H1N1)pdm09”. In: *J Clin Virol* 58.1 (2013), pages 108–13. DOI: 10.1016/j.jcv.2013.05.013.

[97] Oliver Dienz, Jonathan G. Rud, Sheri M. Eaton, Paula A. Lanthier, Elianne Burg, Angela Drew, Janice Bunn, Benjamin T. Suratt, Laura Haynes, and Mercedes Rincon. “Essential role of IL-6 in protection against H1N1 influenza virus by promoting neutrophil survival in the lung”. In: *Mucosal immunology* 5.3 (2012), pages 258–266. DOI: 10.1038/mi.2012.2.

[98] Keer Sun, Luisa Torres, and Dennis W. Metzger. “A Detrimental Effect of Interleukin-10 on Protective Pulmonary Humoral Immunity during Primary Influenza A Virus Infection”. In: *Journal of Virology* 84.10 (2010), pages 5007–5014. DOI: 10.1128/jvi.02408-09.

[99] Estefanía Herrera-Ramos et al. “Surfactant protein A genetic variants associate with severe respiratory insufficiency in pandemic influenza A virus infection”. In: *Critical Care* 18.3 (2014), R127. DOI: 10.1186/cc13934.

[100] K. K. W. To et al. “Surfactant protein B gene polymorphism is associated with severe influenza”. In: *Chest* 145.6 (2014), pages 1237–1243. DOI: 10.1378/chest.13-1651.

[101] Mousumi Dutta, Prafulla Dutta, Subhash Medhi, Biswajyoti Borkakoty, and Dipankar Biswas. “Polymorphism of HLA class I and class II alleles in influenza A(H1N1)pdm09 virus infected population of Assam, Northeast India”. In: *Journal of Medical Virology* 90.5 (2018), pages 854–860. DOI: https://doi.org/10.1002/jmv.25018.

[102] Ramcés Falfán-Valencia, Arun Narayanankutty, Juan M. Reséndiz-Hernández, Gloria Pérez-Rubio, Alejandra Ramírez-Venegas, Karol J. Nava-Quiroz, Nora E. Bautista-Félix, Gilberto Vargas-Alarcón, Manuel D. J. Castillejos-López, and Andrés Hernández. “An Increased Frequency in HLA Class I Alleles and Haplotypes Suggests Genetic Susceptibility to Influenza A (H1N1) 2009 Pandemic: A Case-Control Study”. In: *Journal of Immunology Research* 2018 (2018), page 3174868. DOI: 10.1155/2018/3174868.

[103] S. Aranda-Romo, C. A. Garcia-Sepulveda, A. Comas-García, F. Lovato-Salas, M. Salgado-Bustamante, A. Gómez-Gómez, and D. E. Noyola. “Killer-cell immunoglobulin-like receptors (KIR) in severe A (H1N1) 2009 influenza infections”. In: *Immunogenetics* 64.9 (2012), pages 653–62. DOI: 10.1007/s00251-012-0623-3.

[104] Marlène Pasquet et al. “High frequency of GATA2 mutations in patients with mild chronic neutropenia evolving to MonoMac syndrome, myelodysplasia, and acute myeloid leukemia”. In: *Blood* 121.5 (2013), pages 822–829. DOI: https://doi.org/10.1182/blood-2012-08-447367.

[105] Ithaisa Sologuren et al. “Lethal Influenza in Two Related Adults with Inherited GATA2 Deficiency”. In: *Journal of Clinical Immunology* 38.4 (2018), pages 513–526. DOI: 10.1007/s10875-018-0512-0.

[106] T. C. Dawson, M. A. Beck, W. A. Kuziel, F. Henderson, and N. Maeda. “Contrasting effects of CCR5 and CCR2 deficiency in the pulmonary inflammatory response to influenza A virus”. In: *The American journal of pathology* 156.6 (2000), pages 1951–1959. DOI: 10.1016/S0002-9440(10)65068-7.

[107] Ariel Rodriguez et al. “Characterization In Vitro and In Vivo of a Pandemic H1N1 Influenza Virus from a Fatal Case”. In: *PLOS ONE* 8.1 (2013), e53515. DOI: 10.1371/journal.pone.0053515.

[108] A. Falcon, M. T. Cuevas, A. Rodriguez-Frandsen, N. Reyes, F. Pozo, S. Moreno, J. Ledesma, J. Martínez-Alarcón, A. Nieto, and I. Casas. “CCR5 deficiency predisposes to fatal outcome in influenza virus infection”. In: *J Gen Virol* 96.8 (2015), pages 2074–8. DOI: 10.1099/vir.0.000165.

[109] Dan Dou, Rebecca Revol, Henrik Östbye, Hao Wang, and Robert Daniels. “Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement”. In: *Frontiers in immunology* 9 (2018), pages 1581–1581. DOI: 10.3389/fimmu.2018.01581.

[110] Abraham L. Brass et al. “The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus”. In: *Cell* 139.7 (2009), pages 1243–1254. DOI: 10.1016/j.cell.2009.12.017.

[111] T. M. Desai, M. Marin, C. R. Chin, G. Savidis, A. L. Brass, and G. B. Melikyan. “IFITM3 restricts influenza A virus entry by blocking the formation of fusion pores following virus-endosome hemifusion”. In: *PLoS Pathog* 10.4 (2014), e1004048. DOI: 10.1371/journal.ppat.1004048.

[112] Shokouh Makvandi-Nejad et al. “Lack of Truncated IFITM3 Transcripts in Cells Homozygous for the rs12252-C Variant That is Associated With Severe Influenza Infection”. In: *The Journal of Infectious Diseases* 217.2 (2018), pages 257–262. DOI: 10.1093/infdis/jix512.

[113] Susana David, Vanessa Correia, Liliana Antunes, Ricardo Faria, José Ferrão, Paula Faustino, Baltazar Nunes, Fernando Maltez, João Lavinha, and Helena Rebelo de Andrade. “Population genetics of IFITM3 in Portugal and Central Africa reveals a potential modifier of influenza severity”. In: *Immunogenetics* 70.3 (2018), pages 169–177. DOI: 10.1007/s00251-017-1026-2.

[114] Adrienne G. Randolph et al. “Evaluation of IFITM3 rs12252 Association With Severe Pediatric Influenza Infection”. In: *The Journal of infectious diseases* 216.1 (2017), pages 14–21. DOI: 10.1093/infdis/jix242.

[115] Jéssica S. C. Martins, Maria L. A. Oliveira, Cristiana C. Garcia, Marilda M. Siqueira, and Aline R. Matos. “Investigation of Human IFITM3 Polymorphisms rs34481144A and rs12252C and Risk for Influenza A(H1N1)pdm09 Severity in a Brazilian Cohort”. In: *Frontiers in cellular and infection microbiology* 10 (2020), pages 352–352. DOI: 10.3389/fcimb.2020.00352.

[116] Hannah Limburg et al. “TMPRSS2 Is the Major Activating Protease of Influenza A Virus in Primary Human Airway Cells and Influenza B Virus in Human Type II Pneumocytes”. In: *Journal of Virology* 93.21 (2019), e00649–19. DOI: 10.1128/jvi.00649-19.

[117] Alessia David et al. “A common TMPRSS2 variant protects against severe COVID-19”. In: *medRxiv* (2021), page 2021.03.04.21252931. DOI: 10.1101/2021.03.04.21252931.

[118] M. Lucas-Hourani et al. “Inhibition of pyrimidine biosynthesis pathway suppresses viral growth through innate immunity”. In: *PLoS Pathog* 9.10 (2013), e1003678. DOI: 10.1371/journal.ppat.1003678.

[119] Akiko Iwasaki and Padmini S. Pillai. “Innate immunity to influenza virus infection”. In: *Nature reviews. Immunology* 14.5 (2014), pages 315–328. DOI: 10.1038/nri3665.

[120] A. M. Kell and Jr. Gale M. “RIG-I in RNA virus recognition”. In: *Virology* 479-480 (2015), pages 110–21. DOI: 10.1016/j.virol.2015.02.017.

[121] Susanna Esposito, Claudio Giuseppe Molteni, Silvia Giliani, Cinzia Mazza, Alessia Scala, Laura Tagliaferri, Claudio Pelucchi, Emilio Fossali, Alessandro Plebani, and Nicola Principi. “Toll-like receptor 3 gene polymorphisms and severity of pandemic A/H1N1/2009 influenza in otherwise healthy children”. In: *Virology journal* 9 (2012), pages 270–270. DOI: 10.1186/1743-422X-9-270.

[122] Y. H. C. Leung et al. “Highly pathogenic avian influenza A H5N1 and pandemic H1N1 virus infections have different phenotypes in Toll-like receptor 3 knockout mice”. In: *J Gen Virol* 95.Pt 9 (2014), pages 1870–1879. DOI: 10.1099/vir.0.066258-0.

[123] O. Haller, P. Staeheli, and G. Kochs. “Protective role of interferon-induced Mx GTPases against influenza viruses”. In: *Rev Sci Tech* 28.1 (2009), pages 219–31.

[124] Arindam Chakrabarti, Shuvojit Banerjee, Luigi Franchi, Yueh-Ming Loo, Jr. Gale Michael, Gabriel Núñez, and Robert H. Silverman. “RNase L activates the NLRP3 inflammasome during viral infections”. In: *Cell host & microbe* 17.4 (2015), pages 466–477. DOI: 10.1016/j.chom.2015.02.010.

[125] Xiaoyong Chen, Shasha Liu, Mohsan Ullah Goraya, Mohamed Maarouf, Shile Huang, and Ji-Long Chen. “Host Immune Response to Influenza A Virus Infection”. In: *Frontiers in immunology* 9 (2018), pages 320–320. DOI: 10.3389/fimmu.2018.00320.

[126] Kevan L. Hartshorn, Mitchell R. White, Virginia Shepherd, Ken Reid, Jens C. Jensenius, and E. C. Crouch. “Mechanisms of anti-influenza activity of surfactant proteins A and D: comparison with serum collectins”. In: *American Journal of Physiology-Lung Cellular and Molecular Physiology* 273.6 (1997), pages L1156–L1166. DOI: 10.1152/ajplung.1997.273.6.L1156.

[127] Amber Cardani, Adam Boulton, Taeg S. Kim, and Thomas J. Braciale. “Alveolar Macrophages Prevent Lethal Influenza Pneumonia By Inhibiting Infection Of Type-1 Alveolar Epithelial Cells”. In: *PLOS Pathogens* 13.1 (2017), e1006140. DOI: 10.1371/journal.ppat.1006140.

[128] Michelle D. Tate, Lisa J. Ioannidis, Ben Croker, Lorena E. Brown, Andrew G. Brooks, and Patrick C. Reading. “The role of neutrophils during mild and severe influenza virus infections of mice”. In: *PloS one* 6.3 (2011), e17618–e17618. DOI: 10.1371/journal.pone.0017618.

[129] Jeremy V. Camp and Colleen B. Jonsson. “A Role for Neutrophils in Viral Respiratory Disease”. In: *Frontiers in Immunology* 8 (2017), page 550. DOI: 10.3389/fimmu.2017.00550.

[130] Benjamin M. Tang et al. “Neutrophils-related host factors associated with severe disease and fatality in patients with influenza infection”. In: *Nature Communications* 10.1 (2019), page 3422. DOI: 10.1038/s41467-019-11249-y.

[131] Jerome P. Jayasekera, E. Ashley Moseman, and Michael C. Carroll. “Natural Antibody and Complement Mediate Neutralization of Influenza Virus in the Absence of Prior Immunity”. In: *Journal of Virology* 81.7 (2007), pages 3487–3494. DOI: 10.1128/jvi.02128-06.

[132] Kevin B. O’Brien, Thomas E. Morrison, David Y. Dundore, Mark T. Heise, and Stacey Schultz-Cherry. “A Protective Role for Complement C3 Protein during Pandemic 2009 H1N1 and H5N1 Influenza A Virus Infection”. In: *PLOS ONE* 6.3 (2011), e17377. DOI: 10.1371/journal.pone.0017377.

[133] A. Rattan, S. D. Pawar, R. Nawadkar, N. Kulkarni, G. Lal, J. Mullick, and A. Sahu. “Synergy between the classical and alternative pathways of complement is essential for conferring effective protection against the pandemic influenza A(H1N1) 2009 virus infection”. In: *PLoS Pathog* 13.3 (2017), e1006248. DOI: 10.1371/journal.ppat.1006248.

[134] Bianca L. Artiaga, Guan Yang, Tarun E. Hutchinson, Julia C. Loeb, Jürgen A. Richt, John A. Lednicky, Shahram Salek-Ardakani, and John P. Driver. “Rapid control of pandemic H1N1 influenza by targeting NKT-cells”. In: *Scientific Reports* 6.1 (2016), page 37999. DOI: 10.1038/srep37999.

[135] B. van Wilgenburg et al. “MAIT cells contribute to protection against lethal influenza infection in vivo”. In: *Nat Commun* 9.1 (2018), page 4706. DOI: 10.1038/s41467-018-07207-9.

[136] Ailar Sabbaghi, Seyed Mohammad Miri, Mohsen Keshavarz, Mehran Mahooti, Arghavan Zebardast, and Amir Ghaemi. “Role of γδ T cells in controlling viral infections with a focus on influenza virus: implications for designing novel therapeutic approaches”. In: *Virology Journal* 17.1 (2020), page 174. DOI: 10.1186/s12985-020-01449-0.

[137] K. Kai McKinstry, Tara M. Strutt, Yi Kuang, Deborah M. Brown, Stewert Sell, Richard W. Dutton, and Susan L. Swain. “Memory CD4+ T cells protect against influenza through multiple synergizing mechanisms”. In: *The Journal of Clinical Investigation* 122.8 (2012), pages 2847–2856. DOI: 10.1172/JCI63689.

[138] Endre Kiss-Toth, Edward Harlock, Darren Lath, Thomas Quertermous, and J. Mark Wilkinson. “A TNF Variant that Associates with Susceptibility to Musculoskeletal Disease Modulates Thyroid Hormone Receptor Binding to Control Promoter Activation”. In: *PLOS ONE* 8.9 (2013), e76034. DOI: 10.1371/journal.pone.0076034.

[139] Mei-Lin Yang et al. “Galectin-1 Binds to Influenza Virus and Ameliorates Influenza Virus Pathogenesis”. In: *Journal of Virology* 85.19 (2011), pages 10010–10020. DOI: 10.1128/jvi.00301-11.

[140] Sneha Sant et al. “HLA-B*27:05 alters immunodominance hierarchy of universal influenza-specific CD8+ T cells”. In: *PLOS Pathogens* 16.8 (2020), e1008714. DOI: 10.1371/journal.ppat.1008714.

[141] Tomer Hertz et al. “HLA targeting efficiency correlates with human T-cell response magnitude and with mortality from influenza A infection”. In: *Proceedings of the National Academy of Sciences* 110.33 (2013), pages 13492–13497. DOI: 10.1073/pnas.1221555110.

[142] David La, Chris Czarnecki, Hani El-Gabalawy, Anand Kumar, Adrienne F. A. Meyers, Nathalie Bastien, J. Neil Simonsen, Francis A. Plummer, and Ma Luo. “Enrichment of Variations in KIR3DL1/S1 and KIR2DL2/L3 among H1N1/09 ICU Patients: An Exploratory Study”. In: *PLOS ONE* 6.12 (2011), e29200. DOI: 10.1371/journal.pone.0029200.

[143] Y. Keynan, J. Juno, A. Meyers, T. B. Ball, A. Kumar, E. Rubinstein, and K. R. Fowke. “Chemokine receptor 5 Δ32 allele in patients with severe pandemic (H1N1) 2009”. In: *Emerg Infect Dis* 16.10 (2010), pages 1621–2. DOI: 10.3201/eid1610.100108.

[144] M. Sironi, R. Cagliani, C. Pontremoli, M. Rossi, G. Migliorino, M. Clerici, and A. Gori. “The CCR5Δ32 allele is not a major predisposing factor for severe H1N1pdm09 infection”. In: *BMC Res Notes* 7 (2014), page 504. DOI: 10.1186/1756-0500-7-504.

[145] A. Maestri, V. A. Sortica, D. L. Ferreira, J. de Almeida Ferreira, M. A. Amador, W. A. de Mello, S. E. Santos, and R. C. Sousa. “The His131Arg substitution in the FCGR2A gene (rs1801274) is not associated with the severity of influenza A(H1N1)pdm09 infection”. In: *BMC Res Notes* 9 (2016), page 296. DOI: 10.1186/s13104-016-2096-1.

[146] D. W. Lawrence, W. J. Bruyninckx, N. A. Louis, D. M. Lublin, G. L. Stahl, C. A. Parkos, and S. P. Colgan. “Antiadhesive role of apical decay-accelerating factor (CD55) in human neutrophil transmigration across mucosal epithelia”. In: *J Exp Med* 198.7 (2003), pages 999–1010. DOI: 10.1084/jem.20030380.

[147] Jianuo Liu, Takashi Miwa, Brendan Hilliard, Youhai Chen, John D. Lambris, Andrew D. Wells, and Wen-Chao Song. “The complement inhibitory protein DAF (CD55) suppresses T cell immunity in vivo”. In: *The Journal of experimental medicine* 201.4 (2005), pages 567–577. DOI: 10.1084/jem.20040863.

[148] M. Capasso, L. G. Durrant, M. Stacey, S. Gordon, J. Ramage, and I. Spendlove. “Costimulation via CD55 on human CD4+ T cells mediated by CD97”. In: *J Immunol* 177.2 (2006), pages 1070–7.

[149] Nuno Brito Santos, Zoé Enderlin Vaz da Silva, Catarina Gomes, Celso A. Reis, and Maria João Amorim. “Complement Decay-Accelerating Factor is a modulator of influenza A virus lung immunopathology”. In: *PLOS Pathogens* 17.7 (2021), e1009381. DOI: 10.1371/journal.ppat.1009381.

[150] Covid-Host Genetics Initiative The. “The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic”. In: *European Journal of Human Genetics* 28.6 (2020), pages 715–718. DOI: 10.1038/s41431-020-0636-6.

[151] Emilio Di Maria, Andrea Latini, Paola Borgiani, and Giuseppe Novelli. “Genetic variants of the human host influencing the coronavirus-associated phenotypes (SARS, MERS and COVID-19): rapid systematic review and field synopsis”. In: *Human Genomics* 14.1 (2020), page 30. DOI: 10.1186/s40246-020-00280-6.

[152] Ruth A. Elderfield et al. “Accumulation of human-adapting mutations during circulation of A(H1N1)pdm09 influenza virus in humans in the United Kingdom”. In: *Journal of virology* 88.22 (2014), pages 13269–13283. DOI: 10.1128/JVI.01636-14.

[153] Jianqiang Ye et al. “Variations in the hemagglutinin of the 2009 H1N1 pandemic virus: potential for strains with altered virulence phenotype?” In: *PLoS pathogens* 6.10 (2010), e1001145–e1001145. DOI: 10.1371/journal.ppat.1001145.

[154] M. Matrosovich, T. Matrosovich, J. Carr, N. A. Roberts, and H. D. Klenk. “Overexpression of the alpha-2,6-sialyltransferase in MDCK cells increases influenza virus sensitivity to neuraminidase inhibitors”. In: *J Virol* 77.15 (2003), pages 8418–25. DOI: 10.1128/jvi.77.15.8418-8425.2003.

[155] M. Paula Longhi, Anwen Williams, Matthew Wise, B. Paul Morgan, and Awen Gallimore. “CD59a deficiency exacerbates influenza-induced lung inflammation through complement-dependent and -independent mechanisms”. In: *European Journal of Immunology* 37.5 (2007), pages 1266–1274. DOI: 10.1002/eji.200636755.

[156] Andreas Untergasser, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, and Steven G. Rozen. “Primer3–new capabilities and interfaces”. In: *Nucleic acids research* 40.15 (2012), e115–e115. DOI: 10.1093/nar/gks596.

[157] Alexandre S. Stephens, Sebastien R. Stephens, and Nigel A. Morrison. “Internal control genes for quantitative RT-PCR expression analysis in mouse osteoblasts, osteoclasts and macrophages”. In: *BMC Research Notes* 4 (2011), pages 410–410. DOI: 10.1186/1756-0500-4-410.

[158] A. Frankish et al. “GENCODE reference annotation for the human and mouse genomes”. In: *Nucleic Acids Res* 47.D1 (2019), pages D766–d773. DOI: 10.1093/nar/gky955.

[159] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. “dbSNP: the NCBI database of genetic variation”. In: *Nucleic acids research* 29.1 (2001), pages 308–311. DOI: 10.1093/nar/29.1.308.

[160] Adam Auton et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pages 68–74. DOI: 10.1038/nature15393.

[161] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. “The human genome browser at UCSC”. In: *Genome Res* 12.6 (2002), pages 996–1006. DOI: 10.1101/gr.229102.

[162] Ian Dunham et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), pages 57–74. DOI: 10.1038/nature11247.

[163] J. Wang et al. “Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium”. In: *Nucleic Acids Res* 41.Database issue (2013), pages D171–6. DOI: 10.1093/nar/gks1221.

[164] C. A. Davis et al. “The Encyclopedia of DNA elements (ENCODE): data portal update”. In: *Nucleic Acids Res* 46.D1 (2018), pages D794–d801. DOI: 10.1093/nar/gkx1081.

[165] S. Fishilevich et al. “GeneHancer: genome-wide integration of enhancers and target genes in GeneCards”. In: *Database (Oxford)* 2017 (2017). DOI: 10.1093/database/bax028.

[166] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. “Detection of nonneutral substitution rates on mammalian phylogenies”. In: *Genome Res* 20.1 (2010), pages 110–21. DOI: 10.1101/gr.097857.109.

[167] Sunil Kumar, Giovanna Ambrosini, and Philipp Bucher. “SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity”. In: *Nucleic acids research* 45.D1 (2017), pages D139–D144. DOI: 10.1093/nar/gkw1064.

[168] J. E. Lattin et al. “Expression analysis of G Protein-Coupled Receptors in mouse macrophages”. In: *Immunome Res* 4 (2008), page 5. DOI: 10.1186/1745-7580-4-5.

[169] N. A. Mabbott, J. K. Baillie, H. Brown, T. C. Freeman, and D. A. Hume. “An expression atlas of human primary cells: inference of gene function from coexpression networks”. In: *BMC Genomics* 14 (2013), page 632. DOI: 10.1186/1471-2164-14-632.

[170] J. MacArthur et al. “The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)”. In: *Nucleic Acids Res* 45.D1 (2017), pages D896–d901. DOI: 10.1093/nar/gkw1133.

[171] Oriol Canela-Xandri, Konrad Rawlik, and Albert Tenesa. “An atlas of genetic associations in UK Biobank”. In: *Nature Genetics* 50.11 (2018), pages 1593–1599. DOI: 10.1038/s41588-018-0248-z.

[172] J. Joung, S. Konermann, J. S. Gootenberg, O. O. Abudayyeh, R. J. Platt, M. D. Brigham, N. E. Sanjana, and F. Zhang. “Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening”. In: *Nat Protoc* 12.4 (2017), pages 828–863. DOI: 10.1038/nprot.2017.016.

[173] John G. Doench et al. “Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9”. In: *Nature biotechnology* 34.2 (2016), pages 184–191. DOI: 10.1038/nbt.3437.

[174] Johnny H. Hu et al. “Evolved Cas9 variants with broad PAM compatibility and high DNA specificity”. In: *Nature* 556.7699 (2018), pages 57–63. DOI: 10.1038/nature26155.

[175] Hui Jiang and Wing Hung Wong. “SeqMap: mapping massive amount of oligonucleotides to the genome”. In: *Bioinformatics* 24.20 (2008), pages 2395–2396. DOI: 10.1093/bioinformatics/btn429.

[176] Patrick D. Hsu et al. “DNA targeting specificity of RNA-guided Cas9 nucleases”. In: *Nature Biotechnology* 31.9 (2013), pages 827–832. DOI: 10.1038/nbt.2647.

[177] N. E. Sanjana, O. Shalem, and F. Zhang. “Improved vectors and genome-wide libraries for CRISPR screening”. In: *Nat Methods* 11.8 (2014), pages 783–784. DOI: 10.1038/nmeth.3047.

[178] Jean-Paul Concordet and Maximilian Haeussler. “CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens”. In: *Nucleic Acids Research* 46.W1 (2018), W242–W245. DOI: 10.1093/nar/gky354.

[179] Tim Hsiau et al. “Inference of CRISPR Edits from Sanger Trace Data”. In: *bioRxiv* (2019), page 251082. DOI: 10.1101/251082.

[180] J. Hamann, B. Vogel, G. M. van Schijndel, and R. A. van Lier. “The seven-span transmembrane receptor CD97 has a cellular ligand (CD55, DAF)”. In: *J Exp Med* 184.3 (1996), pages 1185–9.

[181] J. Dunning et al. “Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza”. In: *Nat Immunol* 19.6 (2018), pages 625–635. DOI: 10.1038/s41590-018-0111-5.

[182] J. X. Gray, M. Haino, M. J. Roth, J. E. Maguire, P. N. Jensen, A. Yarme, M. A. Stetler-Stevenson, U. Siebenlist, and K. Kelly. “CD97 is a processed, seven-transmembrane, heterodimeric receptor associated with inflammation”. In: *J Immunol* 157.12 (1996), pages 5438–47.

[183] C. C. Hsiao, H. Y. Chen, G. W. Chang, and H. H. Lin. “GPS autoproteolysis is required for CD97 to up-regulate the expression of N-cadherin that promotes homotypic cell-cell aggregation”. In: *FEBS Lett* 585.2 (2011), pages 313–8. DOI: 10.1016/j.febslet.2010.12.005.

[184] Y. M. Qian, M. Haino, K. Kelly, and W. C. Song. “Structural characterization of mouse CD97 and study of its specific interaction with the murine decay-accelerating factor (DAF, CD55)”. In: *Immunology* 98.2 (1999), pages 303–11.

[185] W. Eichler. “CD97 isoform expression in leukocytes”. In: *J Leukoc Biol* 68.4 (2000), pages 561–7.

[186] Y. Ward et al. “Platelets Promote Metastasis via Binding Tumor CD97 Leading to Bidirectional Signaling that Coordinates Transendothelial Migration”. In: *Cell Rep* 23.3 (2018), pages 808–822. DOI: 10.1016/j.celrep.2018.03.092.

[187] S. Gordon, J. Hamann, H. H. Lin, and M. Stacey. “F4/80 and the related adhesion-GPCRs”. In: *Eur J Immunol* 41.9 (2011), pages 2472–6. DOI: 10.1002/eji.201141715.

[188] J. Hamann, C. Stortelers, E. Kiss-Toth, B. Vogel, W. Eichler, and R. A. van Lier. “Characterization of the CD55 (DAF)-binding site on the seven-span transmembrane receptor CD97”. In: *Eur J Immunol* 28.5 (1998), pages 1701–7. DOI: 10.1002/(SICI)1521-4141(199805)28:05<1701::AID-IMMU1701>3.0.CO;2-2.

[189] J. Hamann, C. van Zeventer, A. Bijl, C. Molenaar, K. Tesselaar, and R. A. van Lier. “Molecular cloning and characterization of mouse CD97”. In: *Int Immunol* 12.4 (2000), pages 439–48.

[190] M. Stacey, G. W. Chang, J. Q. Davies, M. J. Kwakkenbos, R. D. Sanderson, J. Hamann, S. Gordon, and H. H. Lin. “The epidermal growth factor-like domains of the human EMR2 receptor mediate cell attachment through chondroitin sulfate glycosaminoglycans”. In: *Blood* 102.8 (2003), pages 2916–24. DOI: 10.1182/blood-2002-11-3540.

[191] M. J. Kwakkenbos, W. Pouwels, M. Matmati, M. Stacey, H. H. Lin, S. Gordon, R. A. van Lier, and J. Hamann. “Expression of the largest CD97 and EMR2 isoforms on leukocytes facilitates a specific interaction with chondroitin sulfate on B cells”. In: *J Leukoc Biol* 77.1 (2005), pages 112–9. DOI: 10.1189/jlb.0704402.

[192] T. Wang, Y. Ward, L. Tian, R. Lake, L. Guedez, W. G. Stetler-Stevenson, and K. Kelly. “CD97, an adhesion receptor on inflammatory cells, stimulates angiogenesis through binding integrin counterreceptors on endothelial cells”. In: *Blood* 105.7 (2005), pages 2836–44. DOI: 10.1182/blood-2004-07-2878.

[193] E. Wandel, A. Saalbach, D. Sittig, C. Gebhardt, and G. Aust. “Thy-1 (CD90) is an interacting partner for CD97 on activated endothelial cells”. In: *J Immunol* 188.3 (2012), pages 1442–50. DOI: 10.4049/jimmunol.1003944.

[194] W. Eichler, J. Hamann, and G. Aust. “Expression characteristics of the human CD97 antigen”. In: *Tissue Antigens* 50.5 (1997), pages 429–38.

[195] G. Aust, E. Wandel, C. Boltze, D. Sittig, A. Schütz, L. C. Horn, and M. Wobus. “Diversity of CD97 in smooth muscle cells”. In: *Cell Tissue Res* 324.1 (2006), pages 139–47. DOI: 10.1007/s00441-005-0103-2.

[196] G. Aust et al. “Mice overexpressing CD97 in intestinal epithelial cells provide a unique model for mammalian postnatal intestinal cylindrical growth”. In: *Mol Biol Cell* 24.14 (2013), pages 2256–68. DOI: 10.1091/mbc.E13-04-0175.

[197] H. Veninga et al. “Analysis of CD97 expression and manipulation: antibody treatment but not gene targeting curtails granulocyte migration”. In: *J Immunol* 181.9 (2008), pages 6574–83.

[198] G. Boulday, J. Hamann, J. P. Soulillou, and B. Charreau. “CD97-decay-accelerating factor interaction is not involved in leukocyte adhesion to endothelial cells”. In: *Transplantation* 73.3 (2002), pages 429–36.

[199] W. Eichler, A. Lohrenz, K. U. Simon, S. Krohn, J. Lange, S. Bürger, and I. Liebscher. “The role of ADGRE5/CD97 in human retinal pigment epithelial cell growth and survival”. In: *Ann N Y Acad Sci* 1456.1 (2019), pages 64–79. DOI: 10.1111/nyas.14210.

[200] Y. Y. Lu, M. J. Sweredoski, D. Huss, R. Lansford, S. Hess, and D. A. Tirrell. “Prometastatic GPCR CD97 is a direct target of tumor suppressor microRNA-126”. In: *ACS Chem Biol* 9.2 (2014), pages 334–8. DOI: 10.1021/cb400704n.

[201] J. Wilfinger, S. Seuter, T. P. Tuomainen, J. K. Virtanen, S. Voutilainen, T. Nurmi, V. D. de Mello, M. Uusitupa, and C. Carlberg. “Primary vitamin D receptor target genes as biomarkers for the vitamin D3 status in the hematopoietic system”. In: *J Nutr Biochem* 25.8 (2014), pages 875–84. DOI: 10.1016/j.jnutbio.2014.04.002.

[202] M. Wobus, E. Wandel, S. Prohaska, S. Findeiss, K. Tschöp, and G. Aust. “Transcriptional regulation of the human CD97 promoter by Sp1/Sp3 in smooth muscle cells”. In: *Gene* 413.1-2 (2008), pages 67–75. DOI: 10.1016/j.gene.2008.01.021.

[203] O. N. Karpus et al. “Shear stress-dependent downregulation of the adhesion-G protein-coupled receptor CD97 on circulating leukocytes upon contact with its ligand CD55”. In: *J Immunol* 190.7 (2013), pages 3740–8. DOI: 10.4049/jimmunol.1202192.

[204] Y. Yin, X. Xu, J. Tang, W. Zhang, G. Zhangyuan, J. Ji, L. Deng, S. Lu, H. Zhuo, and B. Sun. “CD97 Promotes Tumor Aggressiveness Through the Traditional G Protein-Coupled Receptor-Mediated Signaling in Hepatocellular Carcinoma”. In: *Hepatology* 68.5 (2018), pages 1865–1878. DOI: 10.1002/hep.30068.

[205] J. C. Leemans, A. A. te Velde, S. Florquin, R. J. Bennink, K. de Bruin, R. A. van Lier, T. van der Poll, and J. Hamann. “The epidermal growth factor-seven transmembrane (EGF-TM7) receptor CD97 is required for neutrophil migration and host defense”. In: *J Immunol* 172.2 (2004), pages 1125–31.

[206] E. N. Kop, J. Adriaansen, T. J. Smeets, M. J. Vervoordeldonk, R. A. van Lier, J. Hamann, and P. P. Tak. “CD97 neutralisation increases resistance to collagen-induced arthritis in mice”. In: *Arthritis Res Ther* 8.5 (2006), R155. DOI: 10.1186/ar2049.

[207] M. van Pel, H. Hagoort, M. J. Kwakkenbos, J. Hamann, and W. E. Fibbe. “Differential role of CD97 in interleukin-8-induced and granulocyte-colony stimulating factor-induced hematopoietic stem and progenitor cell mobilization”. In: *Haematologica* 93.4 (2008), pages 601–4. DOI: 10.3324/haematol.11606.

[208] T. Wang, L. Tian, M. Haino, J. L. Gao, R. Lake, Y. Ward, H. Wang, U. Siebenlist, P. M. Murphy, and K. Kelly. “Improved antibacterial host defense and altered peripheral granulocyte homeostasis in mice lacking the adhesion class G protein receptor CD97”. In: *Infect Immun* 75.3 (2007), pages 1144–53. DOI: 10.1128/IAI.00869-06.

[209] H. Veninga, D. M. de Groot, N. McCloskey, B. M. Owens, M. C. Dessing, J. S. Verbeek, S. Nourshargh, H. van Eenennaam, A. M. Boots, and J. Hamann. “CD97 antibody depletes granulocytes in mice under conditions of acute inflammation via a Fc receptor-dependent mechanism”. In: *J Leukoc Biol* 89.3 (2011), pages 413–21. DOI: 10.1189/jlb.0510280.

[210] H. Veninga, R. M. Hoek, A. F. de Vos, A. M. de Bruin, F. Q. An, T. van der Poll, R. A. van Lier, M. E. Medof, and J. Hamann. “A novel role for CD55 in granulocyte homeostasis and anti-bacterial host defense”. In: *PLoS One* 6.10 (2011), e24431. DOI: 10.1371/journal.pone.0024431.

[211] R. J. Abbott, I. Spendlove, P. Roversi, H. Fitzgibbon, V. Knott, P. Teriete, J. M. McDonnell, P. A. Handford, and S. M. Lea. “Structural and functional characterization of a novel T cell receptor co-regulatory protein complex, CD97-CD55”. In: *J Biol Chem* 282.30 (2007), pages 22023–32. DOI: 10.1074/jbc.M702588200.

[212] I. Spendlove and R. Sutavani. “The role of CD97 in regulating adaptive T-cell responses”. In: *Adv Exp Med Biol* 706 (2010), pages 138–48.

[213] R. V. Sutavani, R. G. Bradley, J. M. Ramage, A. M. Jackson, L. G. Durrant, and I. Spendlove. “CD55 costimulation induces differentiation of a discrete T regulatory type 1 cell population with a stable phenotype”. In: *J Immunol* 191.12 (2013), pages 5895–903. DOI: 10.4049/jimmunol.1301458.

[214] S. Wang, Z. Sun, W. Zhao, Z. Wang, M. Wu, Y. Pan, H. Yan, and J. Zhu. “CD97/ADGRE5 Inhibits LPS Induced NF-κB Activation through PPAR-γ Upregulation in Macrophages”. In: *Mediators Inflamm* 2016 (2016), page 1605948. DOI: 10.1155/2016/1605948.

[215] R. M. Hoek, D. de Launay, E. N. Kop, A. S. Yilmaz-Elis, F. Lin, K. A. Reedquist, J. S. Verbeek, M. E. Medof, P. P. Tak, and J. Hamann. “Deletion of either CD55 or CD97 ameliorates arthritis in mouse models”. In: *Arthritis Rheum* 62.4 (2010), pages 1036–42. DOI: 10.1002/art.27347.

[216] J. Hamann, H. Veninga, D. M. de Groot, L. Visser, C. L. Hofstra, P. P. Tak, J. D. Laman, A. M. Boots, and H. van Eenennaam. “CD97 in leukocyte trafficking”. In: *Adv Exp Med Biol* 706 (2010), pages 128–37.

[217] Y. He, W. Wang, L. Xu, L. Li, J. Liu, M. Feng, and H. Bu. “Immunohistochemical Expression and Prognostic Significance of CD97 and its Ligand DAF in Human Cervical Squamous Cell Carcinoma”. In: *Int J Gynecol Pathol* 34.5 (2015), pages 473–9. DOI: 10.1097/PGP.0000000000000200.

[218] Z. He, H. Wu, Y. Jiao, and J. Zheng. “Expression and prognostic value of CD97 and its ligand CD55 in pancreatic cancer”. In: *Oncol Lett* 9.2 (2015), pages 793–797. DOI: 10.3892/ol.2014.2751.

[219] V. P. Vaikari, J. Yang, S. Wu, and H. Alachkar. “CD97 expression is associated with poor overall survival in acute myeloid leukemia”. In: *Exp Hematol* 75 (2019), 64–73.e4. DOI: 10.1016/j.exphem.2019.06.474.

[220] D. Liu, B. Trojanowicz, Y. Radestock, T. Fu, K. Hammje, L. Chen, and C. Hoang-Vu. “Role of CD97 isoforms in gastric carcinoma”. In: *Int J Oncol* 36.6 (2010), pages 1401–8.

[221] M. Safaee et al. “Overexpression of CD97 confers an invasive phenotype in glioblastoma cells and is associated with decreased survival of glioblastoma patients”. In: *PLoS One* 8.4 (2013), e62765. DOI: 10.1371/journal.pone.0062765.

[222] C. C. Hsiao, W. C. Wang, W. L. Kuo, H. Y. Chen, T. C. Chen, J. Hamann, and H. H. Lin. “CD97 inhibits cell migration in human fibrosarcoma cells by modulating TIMP-2/MT1- MMP/MMP-2 activity–role of GPS autoproteolysis and functional cooperation between the N- and C-terminal fragments”. In: *FEBS J* 281.21 (2014), pages 4878–91. DOI: 10.1111/febs.13027.

[223] C. C. Hsiao, K. Keysselt, H. Y. Chen, D. Sittig, J. Hamann, H. H. Lin, and G. Aust. “The Adhesion GPCR CD97/ADGRE5 inhibits apoptosis”. In: *Int J Biochem Cell Biol* 65 (2015), pages 197–208. DOI: 10.1016/j.biocel.2015.06.007.

[224] G. H. Martin et al. “CD97 is a critical regulator of acute myeloid leukemia stem cell function”. In: *J Exp Med* 216.10 (2019), pages 2362–2377. DOI: 10.1084/jem.20190598.

[225] Wen-Ye Tjong and Hsi-Hsien Lin. “The RGD motif is involved in CD97/ADGRE5-promoted cell adhesion and viability of HT1080 cells”. In: *Scientific reports* 9.1 (2019), pages 1517–1517. DOI: 10.1038/s41598-018-38045-w.

[226] S. Becker, E. Wandel, M. Wobus, R. Schneider, S. Amasheh, D. Sittig, C. Kerner, R. Naumann, J. Hamann, and G. Aust. “Overexpression of CD97 in intestinal epithelial cells of transgenic mice attenuates colitis by strengthening adherens junctions”. In: *PLoS One* 5.1 (2010), e8507. DOI: 10.1371/journal.pone.0008507.

[227] T. Zyryanova et al. “Skeletal muscle expression of the adhesion-GPCR CD97: CD97 deletion induces an abnormal structure of the sarcoplasmatic reticulum but does not impair skeletal muscle function”. In: *PLoS One* 9.6 (2014), e100513. DOI: 10.1371/journal.pone.0100513.

[228] H. Yeon Won, S. Hwan Mun, B. Shin, and S. K. Lee. “Contradictory Role of CD97 in Basal and Tumor Necrosis Factor-Induced Osteoclastogenesis In Vivo”. In: *Arthritis Rheumatol* 68.5 (2016), pages 1301–13. DOI: 10.1002/art.39538.

[229] M. J. Kwakkenbos, E. N. Kop, M. Stacey, M. Matmati, S. Gordon, H. H. Lin, and J. Hamann. “The EGF-TM7 family: a postgenomic view”. In: *Immunogenetics* 55.10 (2004), pages 655–66. DOI: 10.1007/s00251-003-0625-2.

[230] Y. Ward, R. Lake, J. J. Yin, C. D. Heger, M. Raffeld, P. K. Goldsmith, M. Merino, and K. Kelly. “LPA receptor heterodimerizes with CD97 to amplify LPA-initiated RHO-dependent signaling and invasion in prostate cancer cells”. In: *Cancer Res* 71.23 (2011), pages 7301–11. DOI: 10.1158/0008-5472.CAN-11-2381.

[231] N. Bhudia, S. Desai, N. King, N. Ancellin, D. Grillot, A. A. Barnes, and S. J. Dowell. “G Protein-Coupling of Adhesion GPCRs ADGRE2/EMR2 and ADGRE5/CD97, and Activation of G Protein Signalling by an Anti-EMR2 Antibody”. In: *Sci Rep* 10.1 (2020), page 1004. DOI: 10.1038/s41598-020-57989-6.

[232] H. Shen, M. Jin, S. Gu, Y. Wu, M. Yang, and X. Hua. “CD97 Is Decreased in Preeclamptic Placentas and Promotes Human Trophoblast Invasion Through PI3K/Akt/mTOR Signaling Pathway”. In: *Reprod Sci* 27.8 (2020), pages 1553–1561. DOI: 10.1007/s43032-020-00183-w.

[233] Yulia Nefedova, Pingyan Cheng, Daniele Gilkes, Michelle Blaskovich, Amer A. Beg, Said M. Sebti, and Dmitry I. Gabrilovich. “Activation of dendritic cells via inhibition of Jak2/STAT3 signaling”. In: *Journal of immunology (Baltimore, Md. : 1950)* 175.7 (2005), pages 4338–4346. DOI: 10.4049/jimmunol.175.7.4338.

[234] Ga Bin Park and Daejin Kim. “MicroRNA-503-5p Inhibits the CD97-Mediated JAK2/STAT3 Pathway in Metastatic or Paclitaxel-Resistant Ovarian Cancer Cells”. In: *Neoplasia (New York, N.Y.)* 21.2 (2019), pages 206–215. DOI: 10.1016/j.neo.2018.12.005.

[235] Bowang Chen, John W. Cole, and Caspar Grond-Ginsbach. “Departure from Hardy Weinberg Equilibrium and Genotyping Error”. In: *Frontiers in Genetics* 8.167 (2017). DOI: 10.3389/fgene.2017.00167.

[236] Lihua Shi, Juan C. Perin, Jeremy Leipzig, Zhe Zhang, and Kathleen E. Sullivan. “Genome-wide analysis of interferon regulatory factor I binding in primary human monocytes”. In: *Gene* 487.1 (2011), pages 21–28. DOI: 10.1016/j.gene.2011.07.004.

[237] Gleb Kichaev, Gaurav Bhatia, Po-Ru Loh, Steven Gazal, Kathryn Burch, Malika K. Freund, Armin Schoech, Bogdan Pasaniuc, and Alkes L. Price. “Leveraging Polygenic Functional Enrichment to Improve GWAS Power”. In: *American journal of human genetics* 104.1 (2019), pages 65–75. DOI: 10.1016/j.ajhg.2018.11.008.

[238] K. Suhre et al. “Connecting genetic risk to disease end points through the human blood plasma proteome”. In: *Nat Commun* 8 (2017), page 14357. DOI: 10.1038/ncomms14357.

[239] Chunlei Wu, Xuefeng Jin, Ginger Tsueng, Cyrus Afrasiabi, and Andrew I. Su. “BioGPS: building your own mash-up of gene annotations and expression profiles”. In: *Nucleic Acids Research* 44.D1 (2015), pages D313–D316. DOI: 10.1093/nar/gkv1104.

[240] H. Yoshida et al. “The cis-Regulatory Atlas of the Mouse Immune System”. In: *Cell* 176.4 (2019), 897–912.e20. DOI: 10.1016/j.cell.2018.12.036.

[241] E. Arner et al. “Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells”. In: *Science* 347.6225 (2015), pages 1010–4. DOI: 10.1126/science.1259418.

[242] M. Berger and M. E. Medof. “Increased expression of complement decay-accelerating factor during activation of human neutrophils”. In: *J Clin Invest* 79.1 (1987), pages 214–20. DOI: 10.1172/jci112786.

[243] Jeffery K. Taubenberger and David M. Morens. “The pathology of influenza virus infections”. In: *Annual review of pathology* 3 (2008), pages 499–522. DOI: 10.1146/annurev.pathmechdis.3.121806.154316.

[244] K. R. Short, E. J. Kroeze, R. A. Fouchier, and T. Kuiken. “Pathogenesis of influenza-induced acute respiratory distress syndrome”. In: *Lancet Infect Dis* 14.1 (2014), pages 57–69. DOI: 10.1016/s1473-3099(13)70286-x.

[245] S. Herold, C. Becker, K. M. Ridge, and G. R. Budinger. “Influenza virus-induced lung injury: pathogenesis and implications for treatment”. In: *Eur Respir J* 45.5 (2015), pages 1463–78. DOI: 10.1183/09031936.00186214.

[246] Pål Voltersvik et al. “Pulmonary changes in Norwegian fatal cases of pandemic influenza H1N1 (2009) infection: a morphologic and molecular genetic study”. In: *Influenza and Other Respiratory Viruses* 10.6 (2016), pages 525–531. DOI: 10.1111/irv.12410.

[247] P. Staeheli, R. Grob, E. Meier, J. G. Sutcliffe, and O. Haller. “Influenza virus-susceptible mice carry Mx genes with a large deletion or a nonsense mutation”. In: *Molecular and cellular biology* 8.10 (1988), pages 4518–4523. DOI: 10.1128/mcb.8.10.4518-4523.1988.

[248] Masaya Fukushi, Tateki Ito, Teruaki Oka, Toshio Kitazawa, Tohru Miyoshi-Akiyama, Teruo Kirikae, Makoto Yamashita, and Koichiro Kudo. “Serial Histopathological Examination of the Lungs of Mice Infected with Influenza A Virus PR8 Strain”. In: *PLOS ONE* 6.6 (2011), e21207. DOI: 10.1371/journal.pone.0021207.

[249] Alexander Karlas et al. “Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication”. In: *Nature* 463.7282 (2010), pages 818–822. DOI: 10.1038/nature08760.

[250] Renate König et al. “Human host factors required for influenza virus replication”. In: *Nature* 463.7282 (2010), pages 813–817. DOI: 10.1038/nature08699.

[251] Troy D. Cline, Donald Beck, and Elizabeth Bianchini. “Influenza virus replication in macrophages: balancing protection and pathogenesis”. In: *The Journal of general virology* 98.10 (2017), pages 2401–2412. DOI: 10.1099/jgv.0.000922.

[252] Sara Clohisey et al. “Comprehensive Characterization of Transcriptional Activity during Influenza A Virus Infection Reveals Biases in Cap-Snatching of Host RNA Sequences”. In: *Journal of Virology* 94.10 (2020), e01720–19. DOI: 10.1128/jvi.01720-19.

[253] Troy D. Cline, Erik A. Karlsson, Bradley J. Seufzer, and Stacey Schultz-Cherry. “The hemagglutinin protein of highly pathogenic H5N1 influenza viruses overcomes an early block in the replication cycle to promote productive replication in macrophages”. In: *Journal of virology* 87.3 (2013), pages 1411–1419. DOI: 10.1128/JVI.02682-12.

[254] Shauna A. Marvin, Marion Russier, C. Theodore Huerta, Charles J. Russell, and Stacey Schultz-Cherry. “Influenza Virus Overcomes Cellular Blocks To Productively Replicate, Impacting Macrophage Function”. In: *Journal of virology* 91.2 (2017), e01417–16. DOI: 10.1128/JVI.01417-16.

[255] C. Pommerenke, E. Wilk, B. Srivastava, A. Schulze, N. Novoselova, R. Geffers, and K. Schughart. “Global transcriptome analysis in influenza-infected mouse lungs reveals the kinetics of innate and adaptive host immune responses”. In: *PLoS One* 7.7 (2012), e41169. DOI: 10.1371/journal.pone.0041169.

[256] H. E. Ghoneim, P. G. Thomas, and J. A. McCullers. “Depletion of alveolar macrophages during influenza infection facilitates bacterial superinfections”. In: *J Immunol* 191.3 (2013), pages 1250–9. DOI: 10.4049/jimmunol.1300014.

[257] Rebecca S. McHugh, Matthew J. Whitters, Ciriaco A. Piccirillo, Deborah A. Young, Ethan M. Shevach, Mary Collins, and Michael C. Byrne. “CD4⁺CD25⁺ Immunoregulatory T Cells: Gene Expression Analysis Reveals a Functional Role for the Glucocorticoid-Induced TNF Receptor”. In: *Immunity* 16.2 (2002), pages 311–323. DOI: 10.1016/S1074-7613(02)00280-7.

[258] D. Jankovic, D. G. Kugler, and A. Sher. “IL-10 production by CD4+ effector T cells: a mechanism for self-regulation”. In: *Mucosal Immunology* 3.3 (2010), pages 239–246. DOI: 10.1038/mi.2010.8.

[259] N. Song et al. “C5a receptor1 inhibition alleviates influenza virus-induced acute lung injury”. In: *Int Immunopharmacol* 59 (2018), pages 12–20. DOI: 10.1016/j.intimp.2018.03.029.

[260] M. M. Hufford, T. S. Kim, J. Sun, and T. J. Braciale. “Antiviral CD8+ T cell effector activities in situ are regulated by target cell type”. In: *J Exp Med* 208.1 (2011), pages 167–80. DOI: 10.1084/jem.20101850.

[261] R. Schoenborn Jamie and B. Wilson Christopher. “Regulation of Interferon-γ During Innate and Adaptive Immune Responses”. In: *Advances in Immunology* 96 (2007), pages 41–101. DOI: https://doi.org/10.1016/S0065-2776(07)96002-2.

[262] M. M. Hufford, T. S. Kim, J. Sun, and T. J. Braciale. “The effector T cell response to influenza infection”. In: *Curr Top Microbiol Immunol* 386 (2015), pages 423–55. DOI: 10.1007/82_2014_397.

[263] N. Sarween, A. Chodos, C. Raykundalia, M. Khan, A. K. Abbas, and L. S. Walker. “CD4+CD25+ cells controlling a pathogenic CD4 response inhibit cytokine differentiation, CXCR-3 expression, and tissue invasion”. In: *J Immunol* 173.5 (2004), pages 2942–51. DOI: 10.4049/jimmunol.173.5.2942.

[264] S. M. Haeryfar, R. J. DiPaolo, D. C. Tscharke, J. R. Bennink, and J. W. Yewdell. “Regulatory T cells suppress CD8+ T cell responses induced by direct priming and cross-priming and moderate immunodominance disparities”. In: *J Immunol* 174.6 (2005), pages 3344–51. DOI: 10.4049/jimmunol.174.6.3344.

[265] Inês Antunes and George Kassiotis. “Suppression of innate immune pathology by regulatory T cells during Influenza A virus infection of immunodeficient mice”. In: *Journal of virology* 84.24 (2010), pages 12564–12575. DOI: 10.1128/JVI.01559-10.

[266] S. Oliphant, J. L. Lines, M. L. Hollifield, and B. A. Garvy. “Regulatory T Cells Are Critical for Clearing Influenza A Virus in Neonatal Mice”. In: *Viral Immunol* 28.10 (2015), pages 580–9. DOI: 10.1089/vim.2015.0039.

[267] Asher Maroof, Yvonne M. Yorgensen, Yufeng Li, and Jay T. Evans. “Intranasal Vaccination Promotes Detrimental Th17-Mediated Immunity against Influenza Infection”. In: *PLOS Pathogens* 10.1 (2014), e1003875. DOI: 10.1371/journal.ppat.1003875.

[268] A. N. Aljurayyan et al. “A critical role of T follicular helper cells in human mucosal anti-influenza response that can be enhanced by immunological adjuvant CpG-DNA”. In: *Antiviral Res* 132 (2016), pages 122–30. DOI: 10.1016/j.antiviral.2016.05.021.

[269] Haiya Wu, Verena Haist, Wolfgang Baumgärtner, and Klaus Schughart. “Sustained viral load and late death in Rag2-/- mice after influenza A virus infection”. In: *Virology journal* 7 (2010), pages 172–172. DOI: 10.1186/1743-422X-7-172.

[270] G. Palladino, K. Mozdzanowska, G. Washko, and W. Gerhard. “Virus-neutralizing antibodies of immunoglobulin G (IgG) but not of IgM or IgA isotypes can cure influenza virus pneumonia in SCID mice”. In: *Journal of virology* 69.4 (1995), pages 2075–2081. DOI: 10.1128/JVI.69.4.2075-2081.1995.

[271] D. J. Topham, R. A. Tripp, and P. C. Doherty. “CD8+ T cells clear influenza virus by perforin or Fas-dependent processes”. In: *The Journal of Immunology* 159.11 (1997), page 5197.

[272] D. J. Topham and P. C. Doherty. “Clearance of an influenza A virus by CD4+ T cells is inefficient in the absence of B cells”. In: *J Virol* 72.1 (1998), pages 882–5.

[273] G. Kassiotis, D. Gray, Z. Kiafard, J. Zwirner, and B. Stockinger. “Functional specialization of memory Th cells revealed by expression of integrin CD49b”. In: *J Immunol* 177.2 (2006), pages 968–75. DOI: 10.4049/jimmunol.177.2.968.

[274] D. Moskophidis and D. Kioussis. “Contribution of virus-specific CD8+ cytotoxic T cells to virus clearance or pathologic manifestations of influenza virus infection in a T cell receptor transgenic mouse model”. In: *The Journal of experimental medicine* 188.2 (1998), pages 223–232. DOI: 10.1084/jem.188.2.223.

[275] Michelle D. Tate, Andrew G. Brooks, and Patrick C. Reading. “The role of neutrophils in the upper and lower respiratory tract during influenza virus infection of mice”. In: *Respiratory research* 9.1 (2008), pages 57–57. DOI: 10.1186/1465-9921-9-57.

[276] Soowon Kang, Hailey M. Brown, and Seungmin Hwang. “Direct Antiviral Mechanisms of Interferon-Gamma”. In: *Immune network* 18.5 (2018), e33–e33. DOI: 10.4110/in.2018.18.e33.

[277] M. Q. Nicol, G. M. Campbell, D. J. Shaw, I. Dransfield, Y. Ligertwood, P. M. Beard, A. A. Nash, and B. M. Dutia. “Lack of IFNgamma signaling attenuates spread of influenza A virus in vivo and leads to reduced pathogenesis”. In: *Virology* 526 (2019), pages 155–164. DOI: 10.1016/j.virol.2018.10.017.

[278] D. Califano, Y. Furuya, S. Roberts, D. Avram, A. N. J. McKenzie, and D. W. Metzger. “IFN-γ increases susceptibility to influenza A infection through suppression of group II innate lymphoid cells”. In: *Mucosal Immunol* 11.1 (2018), pages 209–219. DOI: 10.1038/mi.2017.41.

[279] Jens Zerrahn, Werner Held, and David H. Raulet. “The MHC Reactivity of the T Cell Repertoire Prior to Positive and Negative Selection”. In: *Cell* 88.5 (1997), pages 627–636. DOI: 10.1016/S0092-8674(00)81905-4.

[280] H. Raedler, M. Yang, P. N. Lalli, M. E. Medof, and P. S. Heeger. “Primed CD8(+) T-cell responses to allogeneic endothelial cells are controlled by local complement activation”. In: *Am J Transplant* 9.8 (2009), pages 1784–95. DOI: 10.1111/j.1600-6143.2009.02723.x.

[281] M. Suresh, Hector Molina, Maria S. Salvato, Dimitrios Mastellos, John D. Lambris, and Matyas Sandor. “Complement Component 3 Is Required for Optimal Expansion of CD8 T Cells During a Systemic Viral Infection”. In: *The Journal of Immunology* 170.2 (2003), page 788. DOI: 10.4049/jimmunol.170.2.788.

[282] M. M. Hufford, G. Richardson, H. Zhou, B. Manicassamy, A. García-Sastre, R. I. Enelow, and T. J. Braciale. “Influenza-infected neutrophils within the infected lungs act as antigen presenting cells for anti-viral CD8(+) T cells”. In: *PLoS One* 7.10 (2012), e46581. DOI: 10.1371/journal.pone.0046581.

[283] A. M. Shenoy-Scaria, J. Kwong, T. Fujita, M. W. Olszowy, A. S. Shaw, and D. M. Lublin. “Signal transduction through decay-accelerating factor. Interaction of glycosyl-phosphatidylinositol anchor and protein tyrosine kinases p56lck and p59fyn 1”. In: *The Journal of Immunology* 149.11 (1992), page 3535.

[284] Paulus Mrass, Sreenivasa Rao Oruganti, G. Matthew Fricke, Justyna Tafoya, Janie R. Byrum, Lihua Yang, Samantha L. Hamilton, Mark J. Miller, Melanie E. Moses, and Judy L. Cannon. “ROCK regulates the intermittent mode of interstitial T cell migration in inflamed lungs”. In: *Nature Communications* 8.1 (2017), page 1010. DOI: 10.1038/s41467-017-01032-2.

[285] J. E. Kim, S. Bauer, K. S. La, K. H. Lee, J. T. Choung, K. H. Roh, C. K. Lee, and Y. Yoo. “CD4+/CD8+ T lymphocytes imbalance in children with severe 2009 pandemic influenza A (H1N1) pneumonia”. In: *Korean J Pediatr* 54.5 (2011), pages 207–11. DOI: 10.3345/kjp.2011.54.5.207.

[286] Haipeng Zhang and Ti Wu. “CD4+T, CD8+T counts and severe COVID-19: A meta-analysis”. In: *The Journal of infection* 81.3 (2020), e82–e84. DOI: 10.1016/j.jinf.2020.06.036.

[287] Y. Ward et al. “CD97 amplifies LPA receptor signaling and promotes thyroid cancer progression in a mouse model”. In: *Oncogene* 32.22 (2013), pages 2726–2738. DOI: 10.1038/onc.2012.301.

[288] Ana Clara Monsalvo et al. “Severe pandemic 2009 H1N1 influenza disease due to pathogenic immune complexes”. In: *Nature medicine* 17.2 (2011), pages 195–199. DOI: 10.1038/nm.2262.

[289] A. D. Bretherick et al. “Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits”. In: *PLoS Genet* 16.7 (2020), e1008785. DOI: 10.1371/journal.pgen.1008785.

[290] René S. Kahn et al. “Schizophrenia”. In: *Nature Reviews Disease Primers* 1.1 (2015), page 15067. DOI: 10.1038/nrdp.2015.67.

[291] C. B. Pedersen and P. B. Mortensen. “Evidence of a dose-response relationship between urbanicity during upbringing and schizophrenia risk”. In: *Arch Gen Psychiatry* 58.11 (2001), pages 1039–46. DOI: 10.1001/archpsyc.58.11.1039.

[292] M. Cannon, P. B. Jones, and R. M. Murray. “Obstetric complications and schizophrenia: historical and meta-analytic review”. In: *Am J Psychiatry* 159.7 (2002), pages 1080–92. DOI: 10.1176/appi.ajp.159.7.1080.

[293] “Biological insights from 108 schizophrenia-associated genetic loci”. In: *Nature* 511.7510 (2014), pages 421–7. DOI: 10.1038/nature13595.

[294] Chenxing Liu et al. “The schizophrenia genetics knowledgebase: a comprehensive update of findings from candidate gene studies”. In: *Translational Psychiatry* 9.1 (2019), page 205. DOI: 10.1038/s41398-019-0532-4.

[295] Dheeraj Malhotra and Jonathan Sebat. “CNVs: harbingers of a rare variant revolution in psychiatric genetics”. In: *Cell* 148.6 (2012), pages 1223–1241. DOI: 10.1016/j.cell.2012.02.039.

[296] D. H. Blackwood, A. Fordyce, M. T. Walker, D. M. St Clair, D. J. Porteous, and W. J. Muir. “Schizophrenia and affective disorders–cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family”. In: *American journal of human genetics* 69.2 (2001), pages 428–433. DOI: 10.1086/321969.

[297] Stephen B. Willingham et al. “The CD47-signal regulatory protein alpha (SIRPa) interaction is a therapeutic target for human solid tumors”. In: *Proceedings of the National Academy of Sciences* 109.17 (2012), pages 6662–6667. DOI: 10.1073/pnas.1121623109.

[298] M. A. Morrissey, N. Kern, and R. D. Vale. “CD47 Ligation Repositions the Inhibitory Receptor SIRPA to Suppress Integrin Activation and Phagocytosis”. In: *Immunity* 53.2 (2020), 290–302.e6. DOI: 10.1016/j.immuni.2020.07.008.

[299] Michal Caspi Tal et al. “Upregulation of CD47 Is a Host Checkpoint Response to Pathogen Recognition”. In: *mBio* 11.3 (2020), e01293–20. DOI: 10.1128/mBio.01293-20.

[300] Daniel Martins-de-Souza, Wagner F. Gattaz, Andrea Schmitt, Christiane Rewerts, Giuseppina Maccarrone, Emmanuel Dias-Neto, and Christoph W. Turck. “Prefrontal cortex shotgun proteome analysis reveals altered calcium homeostasis and immune system imbalance in schizophrenia”. In: *European Archives of Psychiatry and Clinical Neuroscience* 259.3 (2009), pages 151–163. DOI: 10.1007/s00406-008-0847-2.

[301] H. Ohnishi et al. “Stress-evoked tyrosine phosphorylation of signal regulatory protein α regulates behavioral immobility in the forced swim test”. In: *J Neurosci* 30.31 (2010), pages 10472–83. DOI: 10.1523/jneurosci.0257-10.2010.

[302] Hisatsugu Koshimizu, Keizo Takao, Takashi Matozaki, Hiroshi Ohnishi, and Tsuyoshi Miyakawa. “Comprehensive behavioral analysis of cluster of differentiation 47 knockout mice”. In: *PloS one* 9.2 (2014), e89584–e89584. DOI: 10.1371/journal.pone.0089584.

[303] Anna B. Toth, Akiko Terauchi, Lily Y. Zhang, Erin M. Johnson-Venkatesh, David J. Larsen, Michael A. Sutton, and Hisashi Umemori. “Synapse maturation by activity-dependent ectodomain shedding of SIRPα”. In: *Nature neuroscience* 16.10 (2013), pages 1417–1425. DOI: 10.1038/nn.3516.

[304] E. K. Lehrman et al. “CD47 Protects Synapses from Excess Microglia-Mediated Pruning during Development”. In: *Neuron* 100.1 (2018), 120–134.e6. DOI: 10.1016/j.neuron.2018.09.017.

[305] A. Sekar et al. “Schizophrenia risk from complex variation of complement component 4”. In: *Nature* 530.7589 (2016), pages 177–83. DOI: 10.1038/nature16549.

[306] D. Inta, U. E. Lang, S. Borgwardt, A. Meyer-Lindenberg, and P. Gass. “Microglia Activation and Schizophrenia: Lessons From the Effects of Minocycline on Postnatal Neurogenesis, Neuronal Survival and Synaptic Pruning”. In: *Schizophr Bull* 43.3 (2017), pages 493–496. DOI: 10.1093/schbul/sbw088.

[307] C. M. Sellgren et al. “Increased synapse elimination by microglia in schizophrenia patient-derived models of synaptic pruning”. In: *Nat Neurosci* 22.3 (2019), pages 374–385. DOI: 10.1038/s41593-018-0334-7.

[308] Deborah Hatherley, Stephen C. Graham, Jessie Turner, Karl Harlos, David I. Stuart, and A. Neil Barclay. “Paired Receptor Specificity Explained by Structures of Signal Regulatory Proteins Alone and Complexed with CD47”. In: *Molecular Cell* 31.2 (2008), pages 266–277. DOI: https://doi.org/10.1016/j.molcel.2008.05.026.

[309] Yang I. Li, David A. Knowles, Jack Humphrey, Alvaro N. Barbeira, Scott P. Dickinson, Hae Kyung Im, and Jonathan K. Pritchard. “Annotation-free quantification of RNA splicing using LeafCutter”. In: *Nature Genetics* 50.1 (2018), pages 151–158. DOI: 10.1038/s41588-017-0004-9.

[310] Alexis C. Komor, Yongjoo B. Kim, Michael S. Packer, John A. Zuris, and David R. Liu. “Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage”. In: *Nature* 533.7603 (2016), pages 420–424. DOI: 10.1038/nature17946.

[311] Andrew V. Anzalone, Luke W. Koblan, and David R. Liu. “Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors”. In: *Nature Biotechnology* 38.7 (2020), pages 824–844. DOI: 10.1038/s41587-020-0561-9.

[312] Gregory M. Findlay, Riza M. Daza, Beth Martin, Melissa D. Zhang, Anh P. Leith, Molly Gasperini, Joseph D. Janizek, Xingfan Huang, Lea M. Starita, and Jay Shendure. “Accurate classification of BRCA1 variants with saturation genome editing”. In: *Nature* 562.7726 (2018), pages 217–222. DOI: 10.1038/s41586-018-0461-z.

[313] J. Kweon, A. H. Jang, H. R. Shin, J. E. See, W. Lee, J. W. Lee, S. Chang, K. Kim, and Y. Kim. "A CRISPR-based base-editing screen for the functional assessment of BRCA1 variants". In: *Oncogene* 39.1 (2020), pages 30–35. DOI: 10.1038/s41388-019-0968-2.

[314] Andrew V. Anzalone et al. "Search-and-replace genome editing without double-strand breaks or donor DNA". In: *Nature* 576.7785 (2019), pages 149–157. DOI: 10.1038/s41586-019-1711-4.

[315] Anob M. Chakrabarti, Tristan Henser-Brownhill, Josep Monserrat, Anna R. Poetsch, Nicholas M. Luscombe, and Paola Scaffidi. "Target-Specific Precision of CRISPR-Mediated Genome Editing". In: *Molecular Cell* 73.4 (2019), 699–713.e6. DOI: https://doi.org/10.1016/j.molcel.2018.11.031.

[316] M. C. Canver et al. "BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis". In: *Nature* 527.7577 (2015), pages 192–7. DOI: 10.1038/nature15521.

[317] Y. Diao, B. Li, Z. Meng, I. Jung, A. Y. Lee, J. Dixon, L. Maliskova, K. L. Guan, Y. Shen, and B. Ren. "A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening". In: *Genome Res* 26.3 (2016), pages 397–405. DOI: 10.1101/gr.197152.115.

[318] N. E. Sanjana, J. Wright, K. Zheng, O. Shalem, P. Fontanillas, J. Joung, C. Cheng, A. Regev, and F. Zhang. "High-resolution interrogation of functional elements in the noncoding genome". In: *Science* 353.6307 (2016), pages 1545–1549. DOI: 10.1126/science.aaf7613.

[319] Josh Tycko et al. "Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements". In: *Nature communications* 10.1 (2019), pages 4063–4063. DOI: 10.1038/s41467-019-11955-7.

[320] Jennifer Harrow et al. "GENCODE: The reference human genome annotation for The ENCODE Project". In: *Genome Research* 22.9 (2012), pages 1760–1774. DOI: 10.1101/gr.135350.111.

[321] Tingting Sui, Yuning Song, Zhiquan Liu, Mao Chen, Jichao Deng, Yuanyuan Xu, Liangxue Lai, and Zhanjun Li. "CRISPR-induced exon skipping is dependent on premature termination codon mutations". In: *Genome Biology* 19.1 (2018), page 164. DOI: 10.1186/s13059-018-1532-z.

[322] D. H. Phanstiel, K. Van Bortle, D. Spacek, G. T. Hess, M. S. Shamim, I. Machol, M. I. Love, E. L. Aiden, M. C. Bassik, and M. P. Snyder. "Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development". In: *Mol Cell* 67.6 (2017), 1037–1048.e6. DOI: 10.1016/j.molcel.2017.08.006.

[323] Gang Ren, Wenfei Jin, Kairong Cui, Joseph Rodrigez, Gangqing Hu, Zhiying Zhang, Daniel R. Larson, and Keji Zhao. "CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression". In: *Molecular cell* 67.6 (2017), 1049–1058.e6. DOI: 10.1016/j.molcel.2017.08.026.

[324] Hiroshi Nishimasu et al. “Engineered CRISPR-Cas9 nuclease with expanded targeting space”. In: *Science* 361.6408 (2018), pages 1259–1262. DOI: 10.1126/science.aas9129.

[325] Maximilian Haeussler et al. “Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR”. In: *Genome Biology* 17.1 (2016), page 148. DOI: 10.1186/s13059-016-1012-2.

[326] M. M. Pradeepa, G. R. Grimes, Y. Kumar, G. Olley, G. C. Taylor, R. Schneider, and W. A. Bickmore. “Histone H3 globular domain acetylation identifies a new class of enhancers”. In: *Nat Genet* 48.6 (2016), pages 681–6. DOI: 10.1038/ng.3550.

[327] C. D. Arnold, D. Gerlach, C. Stelzer, M. Boryń Ł, M. Rath, and A. Stark. “Genome-wide quantitative enhancer activity maps identified by STARR-seq”. In: *Science* 339.6123 (2013), pages 1074–7. DOI: 10.1126/science.1232542.

[328] R. D. Chow, J. S. Chen, J. Shen, and S. Chen. “A web tool for the design of prime-editing guide RNAs”. In: *Nat Biomed Eng* (2020). DOI: 10.1038/s41551-020-00622-8.

[329] H. Kang, M. N. Shokhirev, Z. Xu, S. Chandran, J. R. Dixon, and M. W. Hetzer. “Dynamic regulation of histone modifications and long-range chromosomal interactions during postmitotic transcriptional reactivation”. In: *Genes Dev* 34.13-14 (2020), pages 913–930. DOI: 10.1101/gad.335794.119.

[330] K. M. Dorighi, T. Swigut, T. Henriques, N. V. Bhanu, B. S. Scruggs, N. Nady, 2nd Still C. D., B. A. Garcia, K. Adelman, and J. Wysocka. “Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation”. In: *Mol Cell* 66.4 (2017), 568–576.e4. DOI: 10.1016/j.molcel.2017.04.018.

[331] Olivia Corradin, Alina Saiakhova, Batool Akhtar-Zaidi, Lois Myeroff, Joseph Willis, Richard Cowper-Sal lari, Mathieu Lupien, Sanford Markowitz, and Peter C. Scacheri. “Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits”. In: *Genome research* 24.1 (2014), pages 1–13. DOI: 10.1101/gr.164079.113.

[332] Hui Kwon Kim, Sungtae Lee, Younggwang Kim, Jinman Park, Seonwoo Min, Jae Woo Choi, Tony P. Huang, Sungroh Yoon, David R. Liu, and Hyongbum Henry Kim. “High-throughput analysis of the activities of xCas9, SpCas9-NG and SpCas9 at matched and mismatched target sequences in human cells”. In: *Nature Biomedical Engineering* 4.1 (2020), pages 111–124. DOI: 10.1038/s41551-019-0505-1.

[333] M. Legut, Z. Daniloski, X. Xue, D. McKenzie, X. Guo, H. H. Wessels, and N. E. Sanjana. “High-Throughput Screens of PAM-Flexible Cas9 Variants for Gene Knockout and Transcriptional Modulation”. In: *Cell Rep* 30.9 (2020), 2859–2868.e5. DOI: 10.1016/j.celrep.2020.02.010.

[334] M. Laplana, J. L. Royo, L. F. García, A. Aluja, J. L. Gomez-Skarmeta, and J. Fibla. “SIRPB1 copy-number polymorphism as candidate quantitative trait locus for impulsive-disinhibited personality”. In: *Genes Brain Behav* 13.7 (2014), pages 653–62. DOI: 10.1111/gbb.12154.

[335] I. S. Yang, H. Son, S. Kim, and S. Kim. “ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer”. In: *BMC Genomics* 17.1 (2016), page 631. DOI: 10.1186/s12864-016-2852-6.

[336] Jianguang Ji, Kristina Sundquist, Yi Ning, Kenneth S. Kendler, Jan Sundquist, and Xiangning Chen. “Incidence of cancer in patients with schizophrenia and their first-degree relatives: a population-based study in Sweden”. In: *Schizophrenia bulletin* 39.3 (2013), pages 527–536. DOI: 10.1093/schbul/sbs065.

[337] Golam M. Khandaker, Lesley Cousins, Julia Deakin, Belinda R. Lennox, Robert Yolken, and Peter B. Jones. “Inflammation and immunity in schizophrenia: implications for pathophysiology and treatment”. In: *The lancet. Psychiatry* 2.3 (2015), pages 258–270. DOI: 10.1016/S2215-0366(14)00122-9.

[338] Koby Kidder, Zhen Bian, Lei Shi, and Yuan Liu. “Inflammation Unrestrained by SIRPα Induces Secondary Hemophagocytic Lymphohistiocytosis Independent of IFN-γ”. In: *The Journal of Immunology* (2020), ji2000652. DOI: 10.4049/jimmunol.2000652.

[339] Tracy C. Kuo et al. “Targeting the myeloid checkpoint receptor SIRPα potentiates innate and adaptive immune responses to promote anti-tumor activity”. In: *Journal of hematology and oncology* 13.1 (2020), pages 160–160. DOI: 10.1186/s13045-020-00989-w.

[340] Phillip D. Monk et al. “Safety and efficacy of inhaled nebulised interferon beta-1a (SNG001) for treatment of SARS-CoV-2 infection: a randomised, double-blind, placebo-controlled, phase 2 trial”. In: *The Lancet Respiratory Medicine* 9.2 (2021), pages 196–206. DOI: 10.1016/S2213-2600(20)30511-7.

[341] D. Ellinghaus et al. “Genomewide Association Study of Severe Covid-19 with Respiratory Failure”. In: *N Engl J Med* 383.16 (2020), pages 1522–1534. DOI: 10.1056/NEJMoa2020283.