Genetic variation in the *IFITM* locus and its phenotypic consequences

University of Cambridge

Jesus College

A thesis submitted for the degree of

Doctor of Philosophy

Carmen Lidia Díaz Soria

The Wellcome Trust Sanger Institute
Wellcome Genome Campus
Hinxton, Cambridge
CB10 1SA, UK

March 2017

For Alma, Pavel

and Isabel

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is my own work carried out under the supervision of Prof. Paul Kellam and Dr. Carl Anderson at the Wellcome Trust Sanger Institute. This dissertation does not exceed the word limit set out by the Degree Committee for the Faculty of Biology.

Carmen Lidia Diaz Soria
March 2017

Acknowledgements

I would like to thank my supervisor Prof. Paul Kellam for welcoming me into his group. Most importantly, I would like thank my secondary supervisor Dr. Carl Anderson for all his invaluable advice and support during the last stages of my PhD years and especially during the completion of my thesis. This dissertation would not have been possible without your support and encouragement.

I would also like to give special thanks Dr. Thomas Dan Otto for his guidance and encouragement over the years. I remember when I first approached you having no knowledge of data analysis or programming. You gave me the confidence to tackle programming and data analysis.

Without the support of colleagues and friends, this thesis would have been very difficult to complete. I am forever indebted to my colleagues Dr. Eva Serra, Dr. Loukas Moutsianas, and Dr. Javier Archury and my friends Dr. Manasa Ramakrishna, Dr. Matthew Young and Dr. Michal Szpak for their critical and constructive assessment of my work and their help with figure design.

I would also like to thank current and former team members for their continual advice and guidance, especially Liz Goode, Velislava Petrova, Dr. Jimmy Liu, Rachael Wash, Dr. Anne Palser and Dr. Irene Bassano

I also thank my internal and external collaborators including Dr. Paul Coupland for passing on his expertise and facilitating such fruitful collaboration. I also thank the WTSI Sequencing Pipelines, especially Dr. Sara Widaa and Sarah Sims for ensuring that my samples were sequenced in a prompt manner. I am also very grateful to my thesis committee members Dr. Leo James and Dr. Jeff Barrett, the WTSI Graduate Programme and the Committee of Graduate Studies, as well as the Wellcome Trust for my PhD studentship. I also wish to acknowledge external

collaborators especially Prof. Stuart Neil, Dr. Cristina Pomilla, Prof. Kholoud Porter, and Dr. Paul McLaren for facilitating the completion of the HIV project. I also thanks Prof. Martin Lloyd Hibberd and Dr. Eileen Pg for facilitating access to their dengue genotype data. Lastly, a special thanks to my fellow PhD students, especially Neneh, Mia, Pinky and TJ.

Most importantly, I would like to acknowledge my husband Pavel and my sister Isabel for giving me so much support during the thesis writing. To my daughter Alma, I hope that this work makes up for the time I had to spend away from you. I love you all very much.

Abstract

In the past few years, interferon-induced transmembrane (IFITM) proteins have been identified as important antiviral factors. The current understanding of IFITMs suggests that they localise within distinct cellular compartments from where they exert their broad antiviral role. For example, IFITM1 localises to the plasma membrane and restricts viruses that do not require endocytosis to infect host cells. In contrast, IFITM2 and IFITM3 are found in the early and late endosomes, respectively, and are potent inhibitors of viruses that depend on endosomal pathways for infection.

I begin this dissertation by providing some background on the biology and function of IFITM proteins, including details of *in vitro* assays that have helped elucidate *IFITMs* role as antiviral factors. I also describe some early candidate-gene association studies that have attempted to correlate genetic variation within these genes with variation in viral restriction. I also describe how genetic association studies have been used more broadly to understand the biology underlying both infectious and non-communicable diseases.

Evidence from *in vitro*, and *in vivo* work has demonstrated the *IFITMs* role as potent antiviral factors, however, no genome-wide association study has reported any significant associations to genetic variant in or around these genes. In Chapter 2, I explore reasons why this may be the case and calculate the coverage of *IFITM* genes by commercially available genotyping arrays. I show that *IFITM2* and *IFITM3* are amongst the 7% of all protein coding genes with less than 25% common variant (minor allele frequency > 5%) coverage across all arrays. Poor coverage of genetic variation is therefore one explanation for the lack of *IFITM* associations in GWAS.

The lack of coverage in the genotyping arrays led me to explore other tools to capture variation in the *IFITM* region. I employ a targeted sequencing method using two different sequencing technologies: short-read sequencing (Illumina MiSeq)

and single molecule, real-time sequencing (PacBio *RS*). Conventional pulldown protocols for targeted sequencing have not been designed for single molecule, real-time sequencing at the time, thus in Chapter 3, I provide some details of the optimisation work required to adapt the targeted method for PacBio sequencing. I then assess the performance of the method for both Illumina and PacBio sequencing. Although both platforms successfully capture variation in the region, cost constraints and the capacity for scalability of short-read sequencing guided the decision to use the standard Illumina short-read sequencing for future targeted sequencing studies of the region.

In Chapter 4, I apply the targeted sequencing method described in Chapter 3 to test genetic variants in and around *IFITM1, IFITM2* and *IFITM3* for association with rapid disease progression in HIV. I also explore the contribution of rare genetic variants (MAF < 1%) to this phenotype by testing for a differential enrichment between cases and controls across each of the three genes.

Studies *in vitro* have also reported that IFITM proteins are potent restrictors of dengue virus infection. In Chapter 5, I use genotype data across a cohort of 2,008 Vietnamese children diagnosed with dengue haemorrhagic fever (DHF) and 2,018 cord blood controls to test if common variants are associated with the disease. In order to boost the number of variants available for the association testing, I construct an *IFITM* imputation panel by deep-sequencing the locus in 100 Vietnamese individuals from the 1000 Genomes Consortium. I evaluate the use of these haplotypes for imputation versus those from the Human Reference Consortium (HRC) and the 1000 Genomes Phase 3 (1KP).

Finally, In Chapter 6, I provide an overview of the work from previous Chapters and reflect on the lessons learnt from this work. I also discuss some of the issues highlighted in my work and suggest some study design improvements that would be most relevant for testing genetic variants in *IFITM* for association to infectious diseases.

Table of contents

List of figures

List of tables

1. Introduction

1.1. The biology of interferon transmembrane (*IFITM*) genes

1.1.1. Activation of immune defenses upon pathogen infection

The interactions between pathogens and their hosts are under constant flux. Whilst it is the task of the immune system to protect the host against invading pathogens (bacteria, fungi, parasites or helminths), pathogens have evolved to circumvent such attacks and exploit their host to their advantage. Conversely, hosts have evolved to combat and limit such infections by activating 'non-specific' (innate) and 'specific' (adaptive) immune responses. Upon infection, immune cells such as macrophages and dendritic cells play an important role in the early phase of infection by inducing the production of cytokines to help eradicate pathogens(1). If the host fails to eliminate the pathogen, the adaptive immune response is activated. Crucially, the adaptive immune response develops 'memory' and it is this 'memory' that enables the host to mount a rapid response via antigen-specific effector cells, if the relevant antigen is encountered again. B-cells and T-cells recognise antigens in different manners. Antibodies expressed on the B-cell surface and later on secreted by plasma cells, bind the antigen directly and help neutralise the pathogens or the toxins produced by such pathogens, before they enter the cell. In contrast, T cells recognise antigens only if they are presented by MHC molecules on the surface of antigen-presented cells (APC)(2). The adaptive immune response is regarded as antigen-specific due to the clonal distribution of antigen receptors, exemplified by the surface Ig on antibody producing B-cells, and

T-cell receptors (TCR) on the surface of T-lymphocytes. On the other hand, the role of the innate immune response involves the activation of granulocytes, dendritic cells and macrophages and it is regarded as relatively unspecific(3).

Responses of the innate system to pathogenic challenge rely on the recognition of conserved structures on pathogens which are commonly referred to as pathogen-associated molecular patterns (PAMPs)(3). Four main pattern recognition receptors (PRR) families have been identified, the transmembrane toll-like receptors (TLRs), C-type lectin receptors (CLRs), nod-like receptors and retinoic acid-inducible gene (RIG)-I-like receptors (RLRs) (1, 4) (See Figure 1).

Toll-like receptor signalling:

TLRs are transmembrane glycoproteins located in the cell surface or the endosomes. TLRs contain an extracellular domain, a transmembrane domain and an intracellular domain (TIR) domain(5). All TLRs extracellular domains have a characteristic horseshoe-like structure as a result of tandem copies of a motif known as Leucine-rich repeats (LRR)(4, 6).

The transcriptional outcome of TLRs is dependent on the ligand and adaptor proteins recruited. For example, TLR4 signals from the plasma membrane as well as the endosomes upon recognition of a number of ligands of which LPS is the best described. Signalling from the plasma membrane requires the translocation of TLR4 to lipid rafts rich in TIR-containing adaptor protein (TIRAP). This translocation allows the interaction with MyD88 and the formation of the myddosome composed of MyD88, TIRAP and Interleukin-1 receptor associated kinase (IRAK) 2, 1 and 4. These IRAK proteins then recruit TRAF6 (E3 ubiquitin ligase) that interacts with TAK1-binding protein (TAB) 1, 2, 3, TAK1 and IκB kinase α, β and γ leading to the activation of NF-κB(7).

TLR4 translocation to the endosomes is controlled by CD14 via the activation of adaptor proteins ITAM, Syk and PLCγ2. Once in the endosome, TLR4 interacts with the sorting protein TRAM and signalling adaptor TRIF to induce NF-κB activation through RIPK1, TRADD, and the caspase 8 complex. Activation of type 1 interferon occurs via TRAF3 recruitment. TRAF3 then interacts with TRAF family member-associated NF-κB activator (TANK), IκB kinase γ, ε and TBK1 to induce IRF3 induction of type 1 interferon(7).

TLRs are also crucial for the recognition of viral nucleic acids. For example, endosomal pattern recognition receptor TLR3 recognises double stranded RNA which may be indicative of virus replication or viral genomes. Examples of RNA viruses that have been shown to be restricted by TLR3 include positively stranded RNA viruses in the *Picornaviridae* family such as coxsackie virus group B serotype 3(8), encephalomyocarditis virus(9) and poliovirus (PV)(10).

TLR3 activation results in the production IFN-α/β and cytokines through the Toll-interleukin-1 (IL-1) receptor domain-containing adaptor molecule-1 (TICAM-1, also known as TRIF)(11). Although TLR3 is expressed in several cell types such as fibroblasts and epithelial cells, it is highly expressed in myeloid dendritic cells (DCs), especially in antigen-presenting human CD141^{+} dendritic cells(12). Thus, TLR3 localisation and activation is tuned to detect intracellular virus-derived nucleic acid molecules.

Despite TLR3 antiviral activities, some studies have demonstrated that TLR3 mediated signalling can also exacerbate infection of RNA viruses such as Punta Toro virus (PTV) (13) and Influenza virus(14). TLR3-/- mice exposed to Influenza A virus had a reduced number of inflammatory mediators such as RANTES (regulated upon activation, normal T cell expressed and secreted), interleukin-6, and interleukin-12p40/p70) compared to wild type mice(14). More importantly, infected TLR3-/- mice had a survival advantage over their wild type counterparts.

Other important Toll-like receptors required to limit virus replication include TLR7, 8 and 9. TLR7 and 8 are sensors of ssRNA species produced by vesicular stomatitis viruses and influenza A virus. TLR9, on the other hand, recognises unmethylated CpG islands in DNA viruses such as murine cytomegalovirus(15).

C-Type lectin receptor signalling

C-type lectin receptors are vital for viral recognition. CLRs are expressed by antigen presenting cells such as dendritic cells and macrophages, thereby eliciting a rapid innate response upon virus infection. CLRs can be classified into four main groups: immunoreceptor tyrosine-based activating motif (ITAM) CLRs, hemi-ITAM (hemITAM) CLRs, immunoreceptor tyrosine-based inhibitory motif (ITIM)- CLRs, and a group of CLRs lacking typical signalling motif such as DC-SIGN(16).

Despite the C-type lectin role as antiviral factors, they are exploited by viruses such as HIV-1 in order to gain entry into host cells and to inhibit APCs function(17). For example, upon DC-SIGN recognition of HIV, the virions are transported to the proteasome where lysosomal degradation takes place. Recognition of HIV-1 also triggers Raf-1 activation and the modulation of cytokine responses through NF-κB activation. HIV binding and recruitment of Raf-1 dependent phosphorylation of NF-κB also leads to the recruitment of transcription elongation factor pTEF-b to nascent transcripts leading to transcription elongation and generation of full-length viral transcripts(18). Therefore, although CLRs are important players in innate immunity they are also susceptible to exploitation by viruses such as HIV.

Nod-like receptors

Nucleotide-binding oligomerization domain-containing (NOD)-like receptors are cytosolic proteins that contain C-terminal Leucine rich repeats and a single N-terminal CARD domain (NOD1) or a tandem N-terminal CARD domain (NOD2)(19).

NOD-like receptor family, CARD-containing 2 (NLRC2) appears to directly recognise ssRNA species derived from respiratory syncytial virus (RSV), influenza A virus (IAV) (20). Upon recognition of viral species, NLRC2 associates with IPS-1. This initiates the IPS-1-dependent pathway to induce type I IFN and proinflammatory cytokine release(20).

Amongst the different NOD-like receptors, NLRP3 (NOD-like receptor family, pyrin domain-containing 3) appears to act as an indirect sensor of viral invasion. NLRP3 recognises adenovirus (dsDNA virus)(21), Sendai virus (ssRNA virus)(22), and Influenza A virus (ssRNA virus)(23). Upon activation, NLRP3 oligomerises and recruits ASC and procaspase-1 to form the inflammasome complex. This complex activates caspase-1 which in turn leads the conversion of IL-1β and IL-18 precursors to fully functional IL-1β and IL-18(19).

RIG-1, MDA5 and LGP2 signalling

Another family of pathogen-associated molecular patterns receptors are the retinoic acid-inducible gene (RIG)-I-like receptors (RLRs). These RLRs consist of retinoic acid-inducible gene 1 (RIG-1), melanoma differentiation gene 5 (MDA5) and laboratory of genetics and physiology 2 (LGP2)(24). They possess an RNA helicase binding domain that enables the recognition of RNA. RIG-1 and MDA5 also possess two CARD domains which interact with the adapter protein mitochondrial antiviral signalling (MAVS)(25).

To distinguish between cytosolic endogenous RNA and viral RNA, these receptors have evolved to recognise features specific to viral genomes such as 5' triphosphate RNA and long double stranded RNA. RIG-1 recognises a variety of virus families such as paramyxoviruses and flaviviruses. MDA5 and LGP2, on the other hand, recognise picornaviruses(26). Viral mRNA may contain 7-methyl-guanosine cap at the 5' ends and a polyadenylate tails at the 3' ends. Many positively stranded RNA viruses, however, start with an uncapped 5′-triphosphate, and members of all of these viruses are recognised by RIG-I(26). Negatively stranded viruses with a non-segmental genome such as paramyxoviruses, initiate both replication and transcription *de novo* leading to 5′-triphosphate RNA in the cytosol and are also recognised by RIG-1. In contrast, picornaviruses use an RNA-dependent RNA Polymerase that uses a protein as a primer for positive- and negative-strand RNA production; as a result, during the life cycle of picornaviruses, uncapped triphosphorylated 5′ ends are not present. These species are therefore detected by

MDA5/LGP2 (27). Regardless of the pathway of stimulations by the different PRRs, one common factor following activation of these receptors is the induction of interferon and interferon induced genes to combat infections(28).

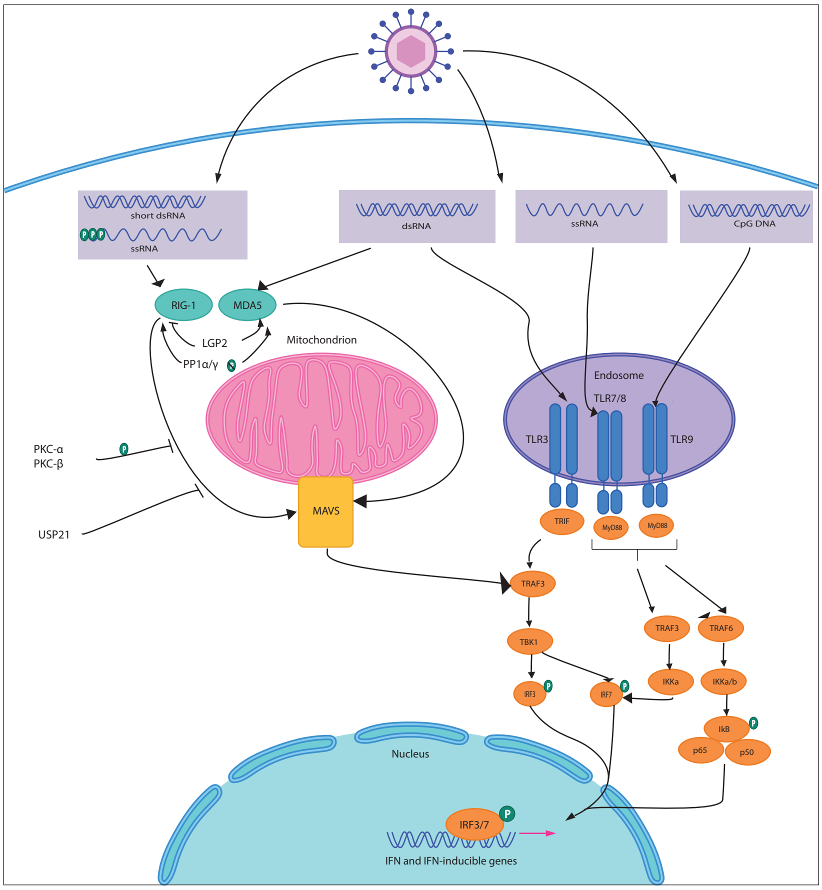


Figure 1. **Viruses are recognised by Intracellular Pattern Recognition Receptors (PRR).** In their inactive state, RIG-1, MDA5 are phosphorylated. Upon recognition of viral RNA, they are dephosphorylated by phosphoprotein phosphatase 1 α and γ (PP1α/γ). Their phosphorylation and consequent activation results in conformational changes in both receptors. RIG-1 C-terminal domain binds to 5'ppp RNA and wraps around RNA molecules through non-specific phosphate sugar interactions in the RNA backbone. MDA5 binds to long dsRNA (>1000bp) and assembles into filaments with the RNA molecules. LGP2 has similar CTD structure to RIG1 and MDA5 and binds to the termini of dsRNA molecules. The juxtapositions of these molecules activate MAVS signalling culminating with the induction of interferon genes. Protein kinases Cα/β act as regulators of RIG-1 signalling by phosphorylation. On the other hand, USP21 regulates RIG-I signalling by de-ubiquitination. Upon binding of RNA or DNA in the endosomal lumen, TLRs dimerise and undergo conformational changes. These conformational changes allow protein kinases to phosphorylate two tyrosine residues in TLRs receptors that trigger TLR activation and recruitment of adapter proteins. Similar to other PRR pathways, recognition of viral species leads to the induction of the interferon genes. (Figure was constructed using information from Mogensen, *et al.*, 2008 (3) and Fernstel, *et al.*, 2015 (29).

1.1.2. Activation of interferon upon infection

Several studies have shown that interferons are crucial for the antiviral defence of the organism. One of the first reports of interferon activity was published back in 1957 by Isaacs and Lindenmann when they observed that the supernatant from chicken cells exposed to heat-inactivated influenza virus 'interfered' with the infection of other cells(30). Since their discovery, Type I interferon IFN-α/β have been found to be one of the first and most important cytokines produced to fight infections. For example, wild type adult 129 Sv/Ev mice challenged with Sindbis virus strain TR339 show only mild symptoms due to functional IFN-α/β receptors. In contrast, 129 Sv/Ev mice carrying non-functional IFN-α/β receptors genes die soon after being infected by the same strain(31). There are many Type I interferons in humans (IFN-α,β,**ε,κ,ω**) and they all bind to a common cell surface receptor: type I, interferon receptor (Figure 2).

Similarly, type II interferon gamma (IFN-γ) was originally discovered due to its capacity to 'interfere' with pathogen infections(32). IFN-γ role is reflected by the susceptibility of C57BL/9 mice with defective IFN-γ gene to *Mycobacterium bovis* bacillus Calmette–Guérin (BCG) infection(33). Typically, defects in the IFN-γ pathway due to mutations in IFN-γ receptor 1 are characterised by severe viral and bacterial infections. In a case study by Jouanguy and colleagues, it was described how inoculation of live BCG vaccine, the most widely used vaccine at the time, proved fatal for a two-and-a-half-month girl, homozygous for a mutation in IFN-γ receptor. In normal circumstances, the attenuated strain of *Mycobacterium bovis* (BCG) is harmless. In rare cases, such as this one, vaccination causes disseminated BCG infection, and is lethal(34). Further evidence was provided by studies of children in Malta with familial immunologic defects caused by mutations in IFN-γ receptor 1 that predisposed these children to mycobacterial infections, despite having access to treatment.

They observed that these children failed to produce TNF-α by macrophages and had defective antigen processing and presentation(35).

The more recently discovered type III interferon lambda family (*IFNλ* family) is composed of *IFNλ1* (*IL29*), *IFNλ2* (*IL28A*), *IFNλ3* (*IL28B*) and *IFNλ4* (*IFNAN*) genes(36). Several polymorphisms in the interferon lambda region have been associated with hepatitis C treatment response and spontaneous clearance of the virus(37, 38). For example, prior to the discovery of *IFNλ4*, a genome wide association study found that the non-coding single nucleotide polymorphism (SNP) rs12979860 was associated with response to pegylated interferon-α with ribavirin (pegIFN-α/RBV) treatment for chronic hepatitis C patients and with spontaneous viral clearance(37). Ge and colleagues reported that patients of European ancestries who carried the homozygous alleles TT in rs12979860 were less likely to respond favourable to pegIFN-α/RBV treatment. On the other hand, carriers of the CC homozygous reference alleles (i.e. those who did not carry the mutation) had a twofold greater rate of sustained virological response (SVR). Sustained virological response refers to the absence of detectable virus following post-treatment evaluations(37). The authors concluded that this variant accounted for half the differences in SVR not only within ethnically matched groups but also across groups of different ethnic backgrounds. Notably 53% of African-Americans with CC genotype had a favourable response rate to treatment compared to only 33% of individuals of European descent that carried the TT genotype.

The authors also tested for association between the rs12979860 variant and baseline viral load pre-treatment. Counterintuitively, Ge and colleagues reported that individuals with the TT alleles and poor HCV treatment response had in fact lower baseline viral load, whereas the opposite was true for patients carrying the CC genotype who had previously been found to have better treatment response(37).

Until recently, the mechanism by which the rs12979860 SNP influenced HCV treatment and spontaneous clearance remained unknown. The first breakthrough in the understanding of this mechanism occurred in 2013 when a group of scientists identified *IFNλ4* (*IFNAN*) near *IFNλ3* gene. IFN-λ4 protein is created by a frameshift mutation rs368234815-ΔG allele which is in high linkage disequilibrium with rs12979860 SNP (now known to be located in intron 1 of *IFNλ4*) (39). In individuals of African ancestry, the *IFNλ4* rs368234815-ΔG allele is in fact a better predictor of poor treatment response to pegIFN/αRBV than rs12979860 TT genotype(40). *IFNλ4*-ΔG/TT shows the strongest association to spontaneous viral clearance and has therefore been proposed as a causal variant underlying the genetic associations reported in HVC clearance and treatment response so far(40).

IFNλ4 protein induces STAT1 and STAT2 phosphorylation, thereby generating an antiviral response in hepatoma cells(39). Experiments *in vitro* have shown that IFN-λ4 binds to the IFN-λ receptor and activates the Janus kinase (JAK)-signal transducer and activator of transcription (STAT) signalling pathway(41), inducing the expression of ISGs (42). As expected, the levels of ISGs in HCV-infected livers is associated with IFN-λ4 expression(43). It is now known that HCV-induced IFN-λ4 expression attenuates the response to exogenous IFN-α treatment by increasing the expression of USP18 and ISG15(44). USP18 is an ISG with the important role of establishing and maintaining long-term desensitisation to type I IFN signalling. It is also reported that overexpression of USP18 also leads to a decrease in the responsiveness to exogenous IFN-α(45). In humans, ISG15 ensures the stability of the USP18 by preventing its ubiquitination thereby enabling USP18-dependent regulation of IFN-α/β(46). In this context, the data supports early findings by Ge and colleagues(37) where it was reported that patients with favourable response to treatment paradoxically had higher viral load pre-treatment. It is now apparent that HCV infection induces the expression of IFNλ4 in patients carrying rs368234815-ΔG allele leading to an increased in the level of ISGs in the liver. Although this induction is not sufficient for virus clearance, it does contribute to the lower levels of HVC viral load observed in some patients. High levels of ISGs upon IFNλ4 expression, including USP18 and ISG15 will act through a negative feedback mechanism to block endogenous IFNα pathway and will contribute to the desensitisation of liver cells to administered IFNα(44) resulting in poor treatment response.

Ultimately, IFNλs together with IFNα and IFNβ can induce over 100 interferon stimulated genes (ISGs) following virus infection(47).

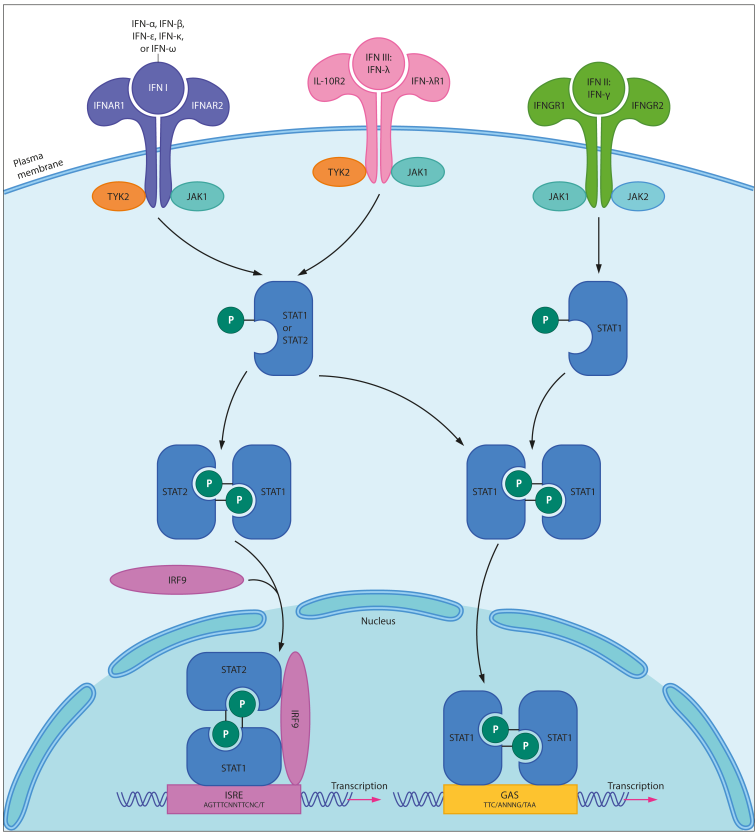


Figure 2. **The interferon (IFN)-signalling cascade.** There are three classes of Interferon (IFN) signals: type I IFNs act through IFN-α receptor 1 (IFNAR1) and IFN-β receptor 2 (IFNAR2) heterodimers; type II IFN act through dimers of heterodimers consisting of IFN-γ receptors 1 (IFNGR1) and 2 (IFNGR2). Finally, type III IFN act through interleukin-10 receptor 2 (IL-10R2) and IFN-λ receptor 1 (IFNLR1) heterodimers. Binding of both type I and type III IFNs triggers a number of signalling pathways that lead to the recruitment and phosphorylation of signal transducers and activators of transcription 1 and 2 (STAT1 and 2). STAT1 and STAT2 associate to form a heterodimer, which in turn recruits the IFN-regulatory factor 9 (IRF9) to form the IFN-stimulated gene factor 3 (ISGF3). Binding of type II IFN dimers to the IFNGR1/2 complex leads to the recruitment of STAT1. Phosphorylated STAT1 homodimers form the IFN-γ activation factor (GAF). Both ISGF3 and GAF translocate to the nucleus to induce IFN-stimulated genes via response elements (ISRE) and gamma-activated sequence (GAS) promoter elements, respectively.

1.1.3. Activation of interferon inducible genes

One classic example of an IFN-α/β stimulated gene is 2’,5’-oligoadenylate synthetase 1 (*OAS*). Variation within this gene was established as an important virus susceptibility factor as a result of mouse experiments in inbred mice. A study by Mashimo and colleagues observed that six unrelated inbred mouse strains from wild type ancestors of *Mus m. domesticus* (WMP/Pas), *Mus musculus* (MAI/Pas, MBT/Pas, PWK/Pas), and *Mus spretus* (SEG/Pas, STF/Pas) were resistant to a highly virulent strain of West Nile virus (strain IS-98-ST1) that normally causes 100% mortality rate in mice(48). Further investigations revealed that these mice lacked a point mutation in the exon 4 of isoform1 of 2’-5’-oligo(A)synthetase (2′-5′-OA) protein(48). Further work on the possible mechanism of restriction for this protein revealed that OAS becomes activated after coming into contact with double stranded RNA. This process triggers a set of reactions where ATP is polymerised into 2’-5’-linked oligoadenylates (2′-5′-OA) resulting in the activation of RNaseL and degradation of the invading pathogen’s nucleic acid(49).

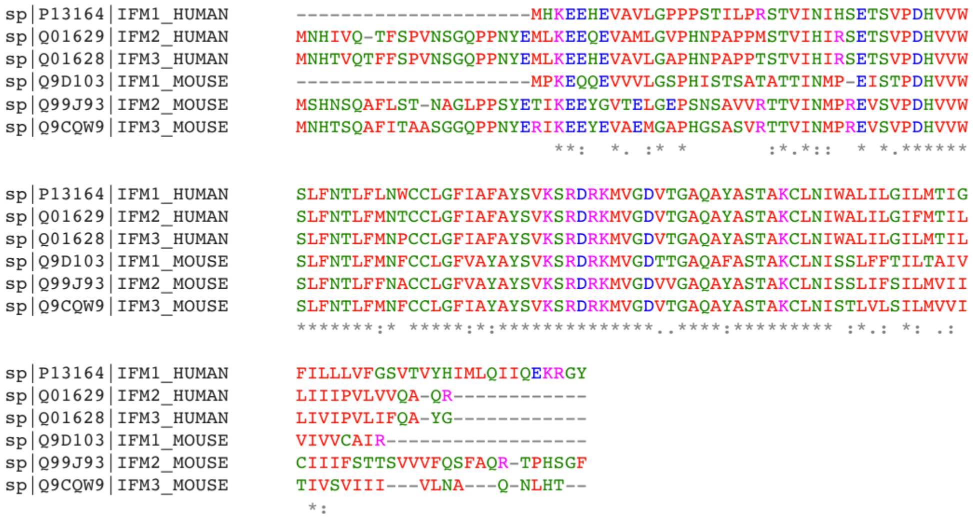
Since the antiviral function of OAS/RNase L system was established in mice, overexpression assays have provided some insight into the antiviral function of a similar system in humans, *in vitro*. Human RNase L is a 741 amino- acid polypeptide that contains nine ankyrin repeats, a kinase-like motifs and an RNase domain(50). 2′-5′-OA has been found to bind to ankyrin repeats 2 and 4 causing the inactivated RNase L monomers to form activated dimers with strong RNase activity(50). Using mammalian A549 lung carcinoma cell lines that stably overexpress wild-type RNase L, Lin and colleagues observed that these cells were resistant to DENV2 infection. By contrast, transduction of A549 cells with a lentivirus-based shRNA targeting human RNase L resulted in up to 39-fold increases in DENV-2 production(51). Similarly, A549 cells also transduced with human OAS1 p42/p46 and OAS p100 also triggered RNase L activity following DENV-2 infection(51). So far, no significant genetic associations have been reported for *OAS* or *RNase L* genes in human genetic studies (GWAS Central, http://www.gwascentral.org/, last accessed on February, 2017) of infectious diseases. A candidate gene study in a cohort of hospitalised patients infected with West Nile virus (WNV) in the USA reported that a polymorphism (rs3213545) in *OAS* was more frequently found in cases compared to controls (*P*=0.004)(52). However, the *P*-value was not genome-wide significant (5x10^{-8}) and the low number of cases (n=33) and controls (n=60) included in the study, suggests that the signal would need to be tested in other cohorts.

Other examples of interferon inducible genes with important roles in virus restriction, are the Interferon transmembrane (*IFITM*) genes. These genes *IFITM1, IFITM2* and *IFITM3* were discovered over a decade ago but interest in their role as antiviral factors re-emerged following an RNAi screen that identified over 100 genes involved in influenza A/B and dengue 2 (New Guinea C strain) restriction(53). In this RNAi screen it was discovered that depletion of *IFITM3* in osteosarcoma cells (U2OS cells) caused an increase in influenza A (A/Puerto Rico/8/34 H1N1-PR8 strain) infection whereas overexpression of all three proteins (IFITM1,2 and 3) resulted in the restriction of influenza A, dengue 2 (New Guinea C strain) and West Nile virus (strain 2741) infections(53).

In vivo studies have also established IFITMs role as antiviral restriction factors. Two groups set out to characterise the susceptibility of *ifitm3*$^{-/-}$ mice to influenza A infections using low pathogenicity virus (A/X-31)(54) and 2009 H1N1 pandemic strain (A/09 Eng/195)(55). Everitt and colleagues found that *ifitm3*$^{-/-}$ mutant mice exhibited rapid loss of body weight (>25%) by day 6 and had to be euthenised(54). By contrast, wild type mice lost <25% of their body weight and fully recovered(54). Bailey and colleagues also reported weight loss >20% in both *ifimt3*$^{-/-}$ and *ifitmDel*$^{-/-}$ (whole locus deletion) with a pathogenic H1N1 strain (A/PR/8/34). Although, unlike Everitt and colleagues, they observed that 60% wild type mice exposed to H1N1 strain (A/PR/8/34) loss around 20% of their body weight by day 7-8 whilst the rest recovered(55). Although, there is some variability in both studies, this suggests that *IFITMs* have important roles modulators of restriction.

1.1.4. Localisation and expression patterns of *IFITM* genes

Members of the *IFITM* gene family in humans are located within 30kb genomic stretch in chromosome 11 at position 11:298,205-327,846. Only IFITM1,2 and 3 have been reported as important restriction factors in viral infections. The other two members of the family, *IFITM5* and *IFITM10* have no reported roles as viral restrictors and their functions will not be addressed in this work. There is a high level of amino acid homology between some of the members of the IFITM protein family in different species and high levels of amino acid similarity (>90%), between human IFITM2 and IFITM3 (Figure 3). In mice, the *IFITM* family include *IFITM*1,2,3,5,6 (*fragilis* 2-6) on chromosome 7(56). Other paralogous and orthologous genes have been reported in mammals including marsupials(57).



```
sp|P13164|IFM1_HUMAN      ---------------------MHKEEHEVAVLGPPPSTILPRSTVINIHSETSVPDHVVW
sp|Q01629|IFM2_HUMAN      MNHIVQ-TFSPVNSGQPPNYEMLKEEQEVAMLGVPHNPAPPMSTVIHIRSETSVPDHVVW
sp|Q01628|IFM3_HUMAN      MNHTVQTFFSPVNSGQPPNYEMLKEEHEVAVLGAPHNPAPPTSTVIHIRSETSVPDHVVW
sp|Q9D103|IFM1_MOUSE      ---------------------MPKEQQEVVVLGSPHISTSATATTINMP-EISTPDHVVW
sp|Q99J93|IFM2_MOUSE      MSHNSQAFLST-NAGLPPSYETIKEEYGVTELGEPSNSAVVRTTVINMPREVSVPDHVVW
sp|Q9CQW9|IFM3_MOUSE      MNHTSQAFITAASGGQPPNYERIKEEYEVAEMGAPHGSASVRTTVINMPREVSVPDHVVW
                                                 **:  *. :* *       :*.*::  * *.******

sp|P13164|IFM1_HUMAN      SLFNTLFLNWCCLGFIAFAYSVKSRDRKMVGDVTGAQAYASTAKCLNIWALILGILMTIG
sp|Q01629|IFM2_HUMAN      SLFNTLFMNTCCLGFIAFAYSVKSRDRKMVGDVTGAQAYASTAKCLNIWALILGIFMTIL
sp|Q01628|IFM3_HUMAN      SLFNTLFMNPCCLGFIAFAYSVKSRDRKMVGDVTGAQAYASTAKCLNIWALILGILMTIL
sp|Q9D103|IFM1_MOUSE      SLFNTLFMNFCCLGFVAYAYSVKSRDRKMVGDTTGAQAFASTAKCLNISSLFFTILTAIV
sp|Q99J93|IFM2_MOUSE      SLFNTLFFNACCLGFVAYAYSVKSRDRKMVGDVVGAQAYASTAKCLNISSLIFSILMVII
sp|Q9CQW9|IFM3_MOUSE      SLFNTLFMNFCCLGFIAYAYSVKSRDRKMVGDVTGAQAYASTAKCLNISTLVLSILMVVI
                          *******:* *****:*:**************..****:********* :*.: *: .:

sp|P13164|IFM1_HUMAN      FILLLVFGSVTVYHIMLQIIQEKRGY
sp|Q01629|IFM2_HUMAN      LIIIPVLVVQA-QR------------
sp|Q01628|IFM3_HUMAN      LIVIPVLIFQA-YG------------
sp|Q9D103|IFM1_MOUSE      VIVVCAIR------------------
sp|Q99J93|IFM2_MOUSE      CIIIFSTTSVVVFQSFAQR-TPHSGF
sp|Q9CQW9|IFM3_MOUSE      TIVSVIII---VLNA---Q-NLHT--
                           *:
```

Figure 3. **Amino acid alignment for IFITM proteins in human and mouse using CLUSTAL OMEGA.** Alignment of human and mouse IFITM1, 2, 3 protein sequences using Clustal Omega. Core protein amino acids are the most conserved and their colours reflect the physiochemical properties: red=hydrophobic amino acids; green=polar and basic amino acids; blue=acidic amino acids, magenta=basic amino acids. An asterisk (*) represents positions that have a single conserved amino acid. A colon (:) indicates conservation between groups of strongly similar properties. A single dot (.) represents conservation between groups of weakly similar properties.

IFITM1, 2 and 3 also share a common CD225 domain and although there is no resolved structure for the IFITM proteins, NMR studies have proposed a model where the N-terminal domain (NTD) is located in the cytoplasm and the C-terminal domain is in the extracellular space (Figure 4).

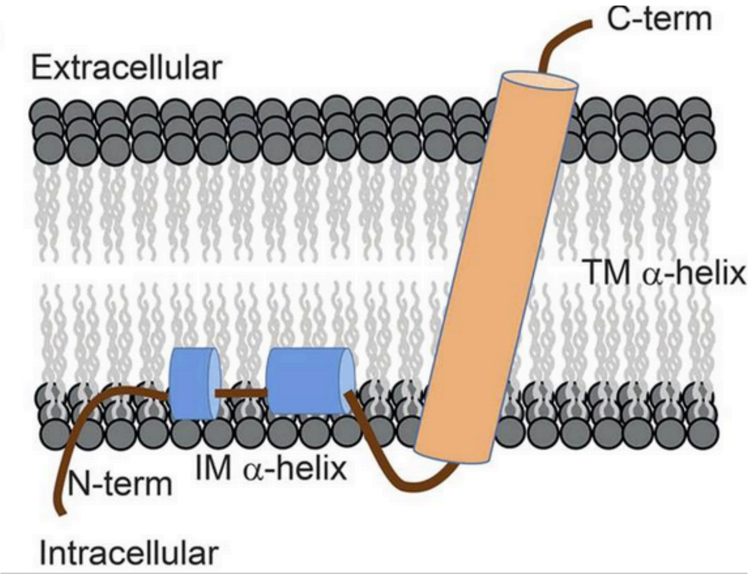


Figure 4. **IFITM3 schematic model.** The latest schematic model of IFITM proteins in the membrane showing a C-terminal transmembrane α-helix and two short intramembrane α-helices. The structure is derived from solution NMR analysis by Ling and colleagues(58).

The Human Protein Atlas (http://www.proteinatlas.org) and GTEx (http://www.gtexportal.org) have provided some insights into IFITM protein expression in human tissue (Figure 5). IFITM1 expression concentrates mainly in the muscle, whole blood, ovaries and lung. The expression patterns of IFITM2 and 3 proteins are similar, their expression concentrated in the fallopian tubes, lung and whole blood (Figure 5).

In mice (C57BL/6) immunohistochemistry studies show that IFITM3 is constitutively expressed in many respiratory tissues and induced in lower airway epithelium upon influenza infection(55). These tissues act as physical barriers between the host and the environment; thus, the expression pattern is consistent with the antiviral roles of IFITM proteins as a potent antiviral factor(54, 55).

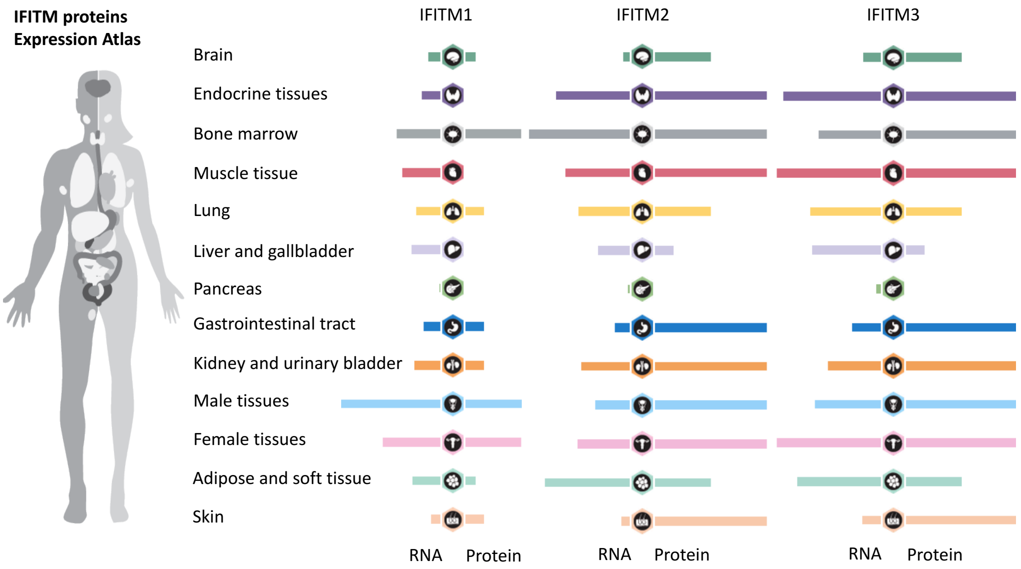


Figure 5. **Snapshot from the Human Protein Atlas**. Overview of the expression of IFITM proteins in 13 human tissues and organs analysed by RNA-seq by the Human Protein Atlas Consortium (HPA).

In terms of the subcellular localisation of IFITM proteins, the general consensus in the field is that IFITM1 localises mainly to the plasma membrane and IFITM2 and 3 localise to intracellular compartments such as the lysosomes and endosomes. For example, in BEL-7404 and Chang liver cells, IFITM1 was found to co-localised with caveolin-1 (CAV-1) in the plasma membrane(59). Other studies, however, report localisation of IFITM1 to the endoplasmic reticulum (ER)(60). IFITM2 and 3 proteins have been reported to localise in lysosomes and endosomes(61). This is indicated by co-localisation with lysosomal and endosomal markers LAMP1(62), and Ras-related protein 7 (Rab7) or CD63(63), respectively.

1.1.5. Broad Spectrum antiviral function of IFITM proteins

The differences in cellular localisation are reflected in the distinct antiviral functions reported for IFITM proteins, *in vitro*. Restriction of Hepatitis C virus is greater in cells that overexpress IFITM1, compared to IFITM2 or IFITM3(64-66). IFITM1 protein is expressed in the plasma membrane where it interacts with the cell surface protein CD81(67). Previous studies have established that both IFITM1 (previously known as Leu 13) and CD81 (previously known as TAPA-1) are associated noncovalently in the plasma membrane(67). This is relevant for IFITM1 HCV antiviral function because CD81 is a cell surface tetraspanin that directly interacts with Hepatitis C virus E2 protein during infection(65, 68). Wilkins and colleagues observed that IFITM1 is localised in hepatic tight junctions where they disrupt the interactions between HCV and cell plasma membrane proteins such as occluding and CD81(65). Although, IFITM2 and 3 have also been reported to restrict infection, it seems that IFITM1 restricts the virus with greater potency compared to its counterparts(66). By contrast, there is no observed inhibition of infection by IFITM1 of other RNA viruses such as Sindbis (SINV) or Semliki Forest virus (SFV)(69). At least for Semliki Forest virus, the lack of IFITM1 restriction may be explained by the endocytic uptake and fusion of SFV with early endosomes which enables it to escape restriction(69).

The patterns of restriction of IFITM2 and 3 are similar; although most studies report a more potent inhibition of infection by IFITM3(53, 54, 70). Studies *in vitro,* show that depletion of *IFITM2/3* by RNAi causes an increase in Influenza A/B and dengue infections whilst overexpression of all three proteins inhibits viral replication(53). Restriction patterns for IFITM2/3 have been more widely studied (Table 1) and have established that, unlike IFITM1, IFITM2/3 can restrict Semliki Forest virus (SFV) and Rift Valley virus (RVFV) and none have been observed to restrict Crimean-Congo haemorrhagic fever virus (CCHFV).

Overall, several labs have reported a broad spectrum of viral restriction for all IFITM proteins (Table 1). However, some of the claims of IFITM differential restrictions are not backed up by strong functional evidence. For example, Mudhasani and colleagues reported restriction (in Vero cells) of two Rift Valley pseudotyped viruses by IFITM2 and 3. Although the difference between the percentage of infected control cells and cells where IFITM proteins were overexpressed was approximately 40% for RVFV-M12 strain and 20% for RVFV-ZH501, the authors still claimed that the proteins restricted both viruses(71). In most infection assays that demonstrate the restriction role of IFITM proteins, they observed 50-70% less infectivity in cells overexpressing IFITM2 and around 70-80% less infectivity in cells expressing IFITM3(53, 72, 73).

Table 1. Table representing some of the most relevant infection assays for IFITM proteins. This table has been adapted from Smith, *et al.*, 2014 to include the type of experiment that was carried out (overexpression or depletion) and the level of restriction reported by these assays.

Family	Virus	pH dependent	Restricts infectivity	Prevents cell–cell fusion	Pseudotyped virions (P) or live virus (L)	IFITM protein	Cell Line	Model	Level of restriction	Reference
Enveloped										
Orthomyxoviridae	Influenza A virus	✓✓	✓	✓	P L	M1–3	A549	Overexpression	M1 - 85% M2 - 75% M3 - 80%	Brass *et al.* 2019
Flaviviridae	West Nile virus	✓	✓		P	M1–3	Vero E6	Overexpression	M1 - 85% M2 - 70% M3 - 90%	Brass *et al.* 2019
	Yellow Fever virus	✓	✓		P	M1–3	Vero E6	Overexpression	M1 - 75% M2 - 55% M3 - 80%	
	Hepatitis C virus	✓	✓		P	M1	Huh7	Overexpression	M1 - 99%	Wilkins et al*., 2013*
	Dengue virus	✓✓	x		P	M3	Hela	Depletion (siRNA)	No restriction	Brass *et al.* 2019
Rhabdoviridae	Vesicular stomatitis virus	✓	✓	✓	P L	M2–3	HEK293	Overexpression	M2 - 5 fold reduction in VSV yields M3 - 15 fold reduction in VSV yields	Weidner et al., 2010
Filoviridae	Marburg virus	Δ	✓		P L	M1–3	A549, Vero E6	Overexpression	M1 - 90% M2 - 80% M3 - 90%	Huang et al., 2011
	Ebola virus	Δ	✓		P L	M1–3			M1 - 90% M2 - 80% M3 - 90%	
Coronaviridae	SARS coronavirus	Δ	✓		P L	M1–3	Vero E6	Overexpression	M1 - 90% M2 - 80% M3 - 70%	Huang et al., 2011
Retroviridae	HIV-1 (CCR5 user)	x	✓		P	M1	U87 neuroblas toma	Overexpression	M1- 60%	Foster *et al.,* 2016
	HIV-1 (CxCR4 user)	x	✓		P	M2	U87 neuroblas toma	Overexpression	M2- 60%	
	HIV-1	x	x		P	M3	Hela-CD4 cells	Overexpression	No restriction	Brass *et al.* 2019
	Jaagsiekte sheep retrovirus	✓	✓	✓	P	M1	HTX	Overexpression	M1 - 60%	Li *et al*., 2013
	Moloney leukaemia virus	×	×		P L	No	Vero E6	Overexpression	No restriction	Brass *et al.* 2019
Arenaviridae	Lassa virus	✓	×		P	No	Vero E6	Overexpression	No restriction	Brass *et al.* 2019
	Machupo virus	✓	×		P	No	Vero E6	Overexpression	No restriction	
	Lymphocytic choriomeningitis virus	✓	×		P	No	Vero E6	Overexpression	No restriction	
Alphaviridae	Semliki Forest virus	✓	✓	✓	L	M2 and M3 best	A549	Overexpression	M2- 60% M3- 90%	Weston *et al*., 2016
Bunyaviridae	Hantaan virus	✓✓	✓		L	M1-3	Vero E6	Overexpression	M1 - 30% M2 - 30% M3 - 30%	Mudhasani *et al*., 2013
	Andes virus	✓✓	✓		L	M1-3	Vero E6	Overexpression	M1 - 30% M2 - 30% M3 - 30%	
	Rift Valley fever virus	✓✓	✓		L-attenuated	M2 and M3	Vero E6	Overexpression	M2- 60% M2 - 60%	
	Crimean–Congo haemorrhagic fever virus	✓✓	×		L	No	Vero E6	Overexpression	No restriction	
Non-enveloped *Reoviridae*	Reovirus	✓✓	✓		L	M3	Hela	Overexpression	M3 - 50%	Anafu *et al.*, 2013

✓= fuses at pH >6; ✓✓ = fuses at pH <6; x = does not require fusion; Δ = requires cathespin L in lysosome.

1.1.6. Mechanism of IFITM restriction

The restriction mechanism of IFITM proteins is still under study. Early investigation of the restriction patterns of IFITMs showed that retroviruses pseudotyped with influenza A haemagglutinin (HA) were restricted in a similar fashion to influenza A pseudotyped viruses. By contrast, retroviruses pseudotyped with murine leukaemia virus, Lassa virus or Machupo virus were not affected by the presence or absence of IFITM proteins, indicating that HA-dependent mechanism of viral entry is targeted by these proteins(70). *In vitro* studies indicate that restriction occurs after the viral particles have been endocytosed but before there is membrane fusion and virions are released into the cytoplasm. Studies using fluorescent microscopy in cell lines that overexpress IFITM3 have found that labelled influenza A virions are internalised and trafficked to the endocytic compartments where they accumulate(70, 74) but then fail to be released. Similarly, recent studies of Semliki Forest virus (SFV) have found that binding, internalisation and endocytosis of the virus is observed in cells that express IFITM3 proteins as well as in non-expressing cells. However, the release of the viral capsid protein to the cytosol is inhibited in IFITM3-expressing cells only(69) (Figure 6). As a consequence, two models of restriction have been proposed. In the first model, IFITM proteins are thought to make an adverse environment in endosomes, so that the viruses cannot fuse with the vesicle membrane. It is thought this is accomplished by interfering with the activity of V-type proton ATPase, which is responsible for the acidification of endocytic compartments(75, 76). Further evidence that amphotericin B (AmphoB) overcomes IFITM2,3 (not IFITM1) from inhibiting influenza A, also supports this hypothesis. AmphoB is a known antifungal drug that is known to interact with sterols present in the plasma membrane; thus compromising its physical properties(77). Clinical preparation and dosage of AmBisome (AmphoB) *in vivo*, found that, similar to *Ifitm3*$^{-/-}$ mice, wild type littermates treated with AmBisome developed severe illness upon low pathogenicity influenza A infection(77). Other studies have also shown that when amphotericin B is added with increasing doses to IFITM2 or IFITM3-expressing

TZM-bl cells, the levels of infection observed are equal to infection in control cells that do not express these proteins(78). Others have suggested that the restriction observed in IFITM3-expressing cells may also be due to a disruption in the interaction of vesicle membrane associated protein A (VAPA) and oxysterol binding protein (OSBP) in endosomal membranes(79). However, attempts to replicate these interactions between VAPA and IFITM3 have failed. For example, overexpression of VAPA had a modest effect on reducing IFITM3-mediated restriction of influenza A in A549 cells (same cells used in the VAPA study). Lin and colleagues also found that modulation of cholesterol levels had no effect on IFITM3-mediated restriction, suggesting that cholesterol mislocalisation is not a contributing factor for VAPA's antagonism of IFITM3(77). In addition, cells expressing either IFITM1,2 or 3 that contain mutations that span the intermembrane domain 2 (IM2), the reported interaction domain of IFITM proteins with VAPA, exhibited levels of restriction similar to cell expressing wild type IFITM. A second model of restriction states that expression of IFITM proteins induces the formation of large vacuoles that are thought to interfere with trafficking and fusion of the virions(70). However, this mechanism has been called into question due to lack of correlation between the size of the vacuoles and the restriction efficiency observed(76).

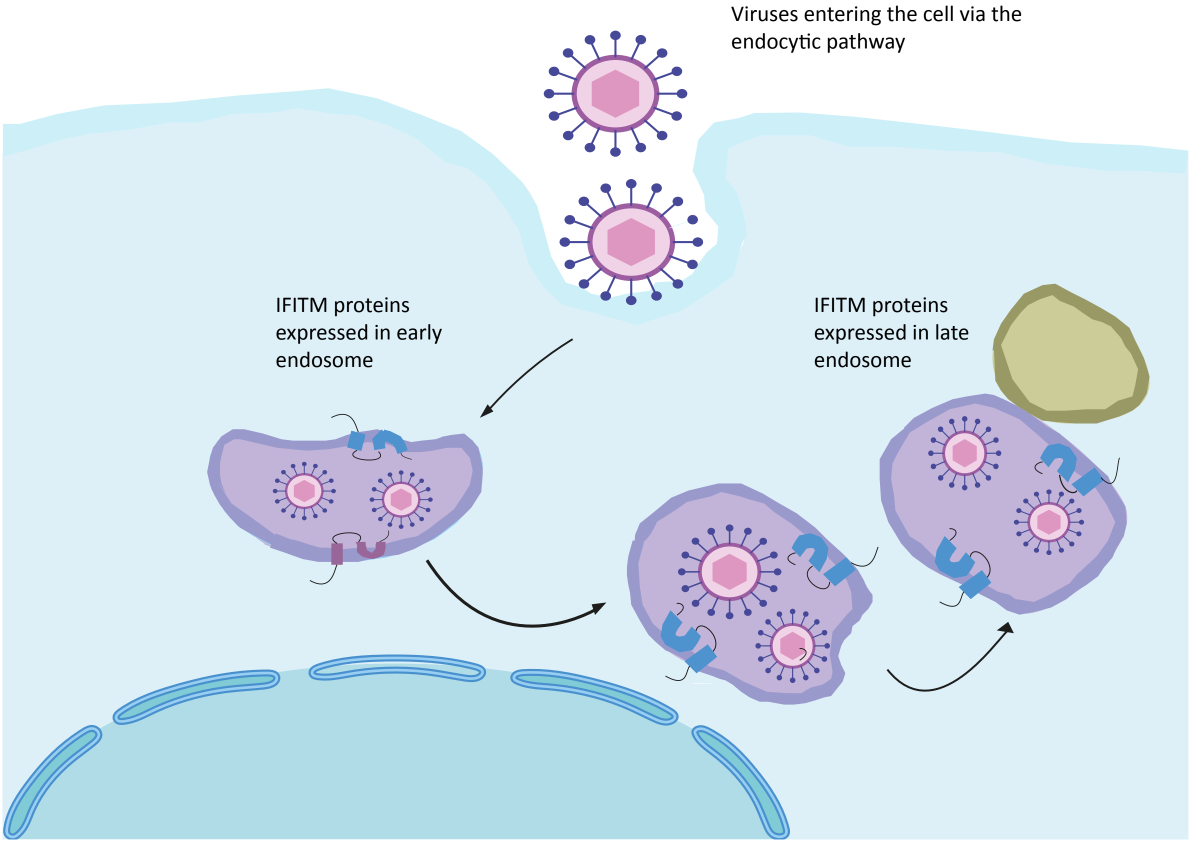

Viruses entering the cell via the endocytic pathway
IFITM proteins expressed in early endosome
IFITM proteins expressed in late endosome

Figure 6. **Mechanism of restriction of IFITM proteins**. IFITM2 and IFITM3 proteins are expressed in the endosomes (vesicles in purple) and restrict a number of viruses entering the cell via the endosomal pathway. The green vesicle = lysosome.

Importantly, the research into the mechanisms of restriction provide some indication of the therapeutic potential of IFITM proteins. These proteins act throughout the early restriction steps and this suggests that the escape mechanisms generally employed by viruses will not be as effective upon their expression. For instance, viral proteins generated after viral entry and replication such as HIV-1 viral infectivity factor (Vif) and viral protein U (Vpu), allow the virus to evade host responses by degrading restriction factors such as Apolipoprotein B MRNA Editing Enzyme Catalytic Subunit 3G (APOBEC3G)(80) and Tetherin(81), respectively. In contrast, IFITM-mediated restriction precedes viral replication, thus, there is little opportunity for the synthesis of *de novo* viral inhibitors. This suggests that unless the virion carries a mutation that counteracts IFITM-mediated restriction, it will be challenging for the virus to evade restriction(76). As a direct result of these observations, several groups have attempted to demonstrate the important role of IFITM proteins as modulators of disease susceptibility in the context of infectious diseases and in the clinic. In the following sections, I will give an overall overview of the genetics of infectious diseases and will address some of the issues associated with genetic studies of *IFITM3* in particular.

1.2. Infectious diseases have a genetic component

1.2.1. Study of genetic susceptibility: twin studies.

Disease susceptibility to infection arises from the intricate interaction of environmental and host factors. One important host factor that is now known to contribute to susceptibility and disease outcome is genetic variation.

Twin studies enable the estimation of the relative contributions of shared genetic and environment effects to variation in a particular disease or trait. The study design was based on comparisons of the phenotypic concordance for a particular trait in genetically identical monozygote twins to that in dizygotic twins, who share on average 50% of their genes. Although there are reservations with regards to the appropriate determination of disease phenotypes and zygosity on some of the early twin studies, there are a number of examples that show there is substantial concordance in susceptibility to infectious diseases in monozygotic compared to dizygotic twins(82), thus suggesting a significant genetic component in disease susceptibility. For example, Hernon and Jenning, 1950, chose to study 46 families of monozygotic (MZ) and dizygotic (DZ) twins suffering from poliomyelitis. They showed the poliomyelitis disease concordance rate in MZ twins was 35.71% compared to 6.06% in DZ siblings thereby highlighting the importance of genetic predisposition to the disease(83).

1.2.2. Candidate gene studies in infectious diseases

Once we have established that genetic effects are important, it is crucial to identify the particular regions of the genome responsible. Finding associated genetic variation and the genes through which they have an effect, can give us important insights into the biology of the disease. For this reason, candidate gene studies have been extremely important for genetic research. In its simplest form, genetic associations studies correlate differences in allele frequencies between cases and controls or within specific continuous traits such as antibody responses to a virus. One important assumption of these types of analysis is therefore that any observed differences in allele frequencies are not the result of unobserved confounding effects such as population stratification but the result of true differences between study groups(84). Due to poor quality controls that include failure to account for population admixture and stratification, and poor choice of candidate genes, candidate gene studies have often been known to report a large number of spurious associations(85, 86).

1.2.3. Candidate gene studies for *IFITM3*

Everitt and colleagues, were first to report an association between the C allele in the *IFITM3* non-coding, splice region variant SNP rs12252 (T → C) and an increase in susceptibility to pandemic influenza (H1N1) infection. Using 55 hospitalised cases of severe flu and 360 European controls from the 1000 Genomes, they found an enrichment in the number of patients (13.2%) that carried the minority allele C for rs12252 (P = 0.00006, no Odds Ratio provided). In particular, they discovered that 5.7% of hospitalised cases of European descent were homozygous for rs12252 and carried the CC genotype compared to only 0.3% of their control population (n=360 Europeans from the 1000 Genomes Project). They hypothesised that a consequence of carrying the minority C allele is the expression of a truncated IFITM3 protein lacking the first 21 amino acids due to the use of an alternative start codon(54).

Other genetic studies soon followed (Table 2) confirming similar findings in Asian populations where the frequency of rs12252 is much higher (MAF = 0.53). These groups reported associations to not only influenza H1N1(87, 88) infections (n=83, OR=6.4) but also for HIV(89) (n=178, OR=3.8) and Hantaan(90) (n=69, OR=2.1) virus infections. For example, Zhang and colleagues, analysed the association between the rs12252 SNP and H1N1 influenza in a Chinese cohort (n=35 cases of severe influenza and n=48 control patients displaying symptoms of mild flu)(88). This study reported an association between homozygotes for the C allele and severe influenza when compared to the mild control population (*P*=0.0002, OR=6.4). Specifically, they found that 69% of hospitalised patients suffering from severe flu carried the CC alleles compared to 25% of patients suffering mild symptoms. As a consequence, they concluded that rs12252 associates with severe H1N1 influenza.

Table 2. Reported associations for *IFITM3* rs12252

Phenotype	**Population**	**rs12252 AF**	**Cases**	**Controls**	***P* value**	**Odd ratio (95% CI)**	**Model**	**Genotyping method**	**Reference**
Pandemic or seasonal H1N1/09 influenza	European	0.04	53	360*	6x10^{-5}	not stated	not stated	PCR	Everitt, *et al.*, 2012
Pandemic H1N1/09 influenza	Asian	0.53	35	48	2x10^{-4}	6.4 (2.4–17.1)	Recessive	PCR	Zhang, *et al.*, 2013
H1N1 influenza	European	0.04	87$^{\triangle}$	2,623	1x10^{-2}	23.4 (5.2–106.1)	Recessive	PCR	Mills, *et al.*, 2014
H7N9 influenza	Asian	0.53	16	197*	3x10^{-2}	not stated	not stated	PCR	Wang, *et al.* 2014
Pandemic H1N1/09 influenza	European	0.04	84	184	4x10^{-1}	0.7 (0.3-1.5)	Dominant	PCR	Gaio, *et al.* 2016
Pandemic H1N1/09 influenza	European	0.04	118	353	4x10^{-2}	1.9 (0.9–3.9)	not stated	PCR	Lopez-Rodriguez, *et al.*, 2016
HIV progression	Asian	0.53	74	104	4x10^{-3}	3.8 (1.5–9.7)	Dominant	PCR	Zhang, *et al.*, 2015
Hantaan virus	Asian	0.53	69	197*	8x10^{-3}	2.1 (1.1–4.2)	not stated	PCR	Xu-yang, *et al.* 2017

(*) represents controls for that same population, in the 1000 Genomes Phase 3 release

($^{\triangle}$) represents combined cases from Everitt, *et al.*, 2012 and Mills, *et al.*, 2014

Since these studies were published, there have been almost an equal number of reports contradicting these findings(91-93). Mills and colleagues examined two separate cohorts to test for associations to viral infections just focusing on *IFITM3*. Their cohorts included patients with severe H1N1 influenza that required hospitalisation due to pneumonia (n = 34); patients with lower respiratory tract infection (LRTI) (n = 2,730), and healthy controls matched to the patients with LRTI (n = 2,623). Even when combining data from severe influenza cases used in other candidate gene studies, they did not detect association between rs12252 and severe influenza(91). More recently, Lopez-Rodriguez reported no associations between rs12252 and severe influenza(93) (P=0.048). Further analysis of combined genotype data from their study and that included in Mills, *et al*., resulted in a marginal association of this SNP with non-severe influenza infections. They therefore concluded that at least in European populations, rs12252 is not associated with severe influenza(93). These findings are in line with another recent study which did not find association between rs12252 and severe influenza cases(92).

Notwithstanding all the conflicting evidence of association for rs12252, there is extensive functional data from *in vitro* and mouse studies which have demonstrated the important role of IFITM proteins as modulators of disease susceptibility. This suggests that better study designs, with larger sample sizes, will be necessary in the future in order to ascertain if these genes contribute to variation in infectious disease risk.

1.2.4. Other examples of candidate gene studies in infectious diseases

Despite the known shortcomings of candidate gene studies, especially with regards to IFITM research, there are some examples where candidate gene designs have been successfully employed. One classic example is the 32bp deletion in CCR5 (CCR5Δ32), a major receptor for HIV virus. A lack of CCR5 receptors in the plasma membrane as a result of this deletion, confers individuals homozygous for the mutation almost complete protection from HIV-1 infection(94-96). Even for heterozygous individuals, carrying the 32 base pair deletion is associated with delayed disease progression after HIV-1 infection(94, 97). In a case-control study of 364 homosexual men (long term survivors versus rapid progressors) with HIV-1 infection in the Netherlands, it was found that although seroconvertion occurred in both groups, 48% of long-term survivors were heterozygous for CCR5Δ32 compared with 9% of progressors (odds ratio, 6.9 [95% CI, 1.9 to 24.8])(97).

Other candidate gene studies have successfully identified other host factors that influence HIV-1 progression. One example is the human leukocyte antigen type (HLA) that is strongly associated with HIV-1 infection. Comparison of frequencies of HLA B*57 allele between long term non-progressors (LTNP) and progressors revealed that 85% of non-progressors (11 out of 13 patients) had an enrichment in HLA B*5701 class I allele compared to 9.5% (19 of 200) in progressors ($P < 0.001$)(98). Both CCR5Δ32 and HLA B*5701 associations have been replicated in genome-wide association studies. Interestingly however, the association of CCR5Δ32 in a large genome-wide scan have only recently been reported(99).

1.2.5. General principles of genome-wide association studies

Genome-wide association studies (GWAS) allow researchers to scan the genome for association without any *a priori* knowledge about the role of specific genes in disease susceptibility. Specific technological advances in the production of high throughput genome wide arrays(100) coupled with efforts by the HapMap Project(101) to map the correlation between common genetic variation in the human genome, contributed towards their advent.

The HapMap project identified more than 3 million SNPs across 269 individuals from three distinct populations: 30 parent and child trios with European ancestries (referred to as CEU samples); 30 trios from the Yoruba population in Nigeria (YRI); 44 unrelated Chinese individuals (CHB) and 44 unrelated Japanese individuals(101). The HapMap project also established that the rate of mutation in the human genome is in the order of 10^{-8} per site per generation, which is very low compared to the number of generations (10^{-4}) since our most recent common ancestors(101). For this reason, when a mutation occurs, it does so on a specific genetic background; thus, the newly generated allele remains associated with other nearby alleles that were present when the mutation occurred. This order of alleles across a chromosome is known as 'haplotype' whereas linkage disequilibrium (LD) refers to the correlation of these alleles. There are practical implications to the strong associations found between SNPs; typically, with only a few carefully chosen SNPs (tag SNPs), it is possible to determine a large percentage of the genetic variation genome-wide.

There are two main classes of phenotypes that are employed in GWAS designs: the binary case-control designs (e.g. HIV rapid progressors versus HIV non-progressors); and the quantitative design (e.g. viral load of asymptomatic HIV individuals or IgG response following an HIV infection). Regardless of the design in question, GWAS studies look for statistically significant differences in the frequency of a particular allele between cases and controls or across a quantitative trait such as viral load. Typically, genome-wide studies test for a deviation from the null hypothesis and quantify this deviation by providing a *P* value. When the marker reaches a certain threshold ($5x10^{-8}$) for genome-wide significance, then the marker is said to be associated. This specific threshold for the *P* value accounts for a 5% type I error rate (5% chance of a false positive finding due to chance) in 1 million independent tests. When performing one test, *P*=0.05 is a reasonable number, however when performing 1,000,000 tests (approximate number of independent common (>5%) variable regions in European genomes), a *P*=0.05 would result in a large number of false positives ($1x10^{6}$*0.05=50,000), thus a more stringent value is required. Generally, association studies also report the odd ratios (ORs) or beta-coefficients (β) to provide an estimation of risk.

1.2.6. Genome-wide association studies for infectious diseases

To date, no genome-wide significant associations have been reported for any of the *IFITM* genes but a number of significant findings have been made around other genes for infectious diseases such as HIV-1, tuberculosis and malaria. Typically, many GWAS studies of HIV use the set-point viral load (spVL) to define cases during the asymptomatic phase of HIV infection. This phase follows on from the initial acute HIV infection and it is characterised by relatively stable viral replication. This study design was successfully employed by Fellay and colleagues to confirm the central role of genetic variants in the MHC region for HIV-1 infection and progression using genome-wide association studies. Using data from 486 patients from the Euro-CHAVI (Centre for HIV/AIDS Vaccine Immunology) cohort, they reported two genome-wide significant associations. One association for rs2395029 ($P = 9.36 \times 10^{-12}$), accounted for 9.6% of the total variation in the viral set point with patients heterozygous for this mutation displaying a marked reduction in viral load(102). This SNP is situated near the HLA complex 5 gene which is in complete LD with HLA-B*57, already known to be protective from reports by a previous candidate gene study(98). Another association was reported for rs9264942 (P = 3.77 × 10^{-9}), located 35kb form HLA-C gene(102). Replication studies in a larger number of individuals confirmed these findings and reported new nominally significant associations for rs9468692 (P = 3.6 × 10^{-5}) located at 3' region of *TRIM10* and the non-synonymous SNPs rs8192591 (P = 5.5 × 10^{-5}) located in coding region of *NOTCH4* gene(103, 104). A recent GWAS on 6,315 HIV positive European individuals also confirmed the central role of the MHC region in HIV susceptibility and progression(99). In this study, they mapped MHC association signals to a peptide-binding groove of HLA-B and HLA-A regions(99). Interestingly, the authors also observed a strong association between CCR5Δ32 with reduced set-point viral load ($P=1.6x10^{-16}$) and reported that seven other markers overlapping the *CCR5* region appear to tag variants distinct from CCR5Δ32, suggesting the presence of other causal variants that are yet to be discovered(99).

Further successes of GWA studies have been reported for other infectious diseases such as tuberculosis (TB). In one of the largest genetic studies of TB so far, Curtis and colleagues genotyped 5,530 individuals with pulmonary TB and 5,607 healthy controls(105). They found one association in *ASAP1*, a gene that encodes an Arf GTPase-activating protein (Arf GAP). Meta-analysis using published data from Ghanaian (971 TB cases and 988 controls) and Gambian (1,306 TB cases and 1,372 controls) cohorts validated these findings. Furthermore, functional analysis of dendritic cells showed that when these cells were infected with *Mycobacterium bovis* BCG, this led to the reduction of *ASAP1* expression. In addition, homozygotes of the allele A at rs10956514, which was associated with higher TB risk in this study, displayed a stronger reduction of *ASAP1* expression following *Mycobacterium bovis* BCG infection; whereas the opposite was observed for homozygotes of allele G, which was associated with lower TB risk(105).

Other genome-wide studies have served to validate previous associations. For example, in a GWAS of severe malaria (n=1,060 cases and n=1,500 controls) from The Gambia, Jallow and colleagues(106) reported that their strongest signal of association was near the haemoglobin beta (*HBB*), the gene that contains the sickle haemoglobin variant haemoglobin S (*HbS*) polymorphism rs334. This polymorphism results in a non-synonymous protein change where a glutamic acid in the β-globin chain is replaced by the amino acid valine. Homozygotes for this mutation experience life-threatening disease due to sickle cell anaemia, whereas heterozygotes have a tenfold reduced risk of severe malaria(107). Unfortunately, the authors did not find any new loci that reached genome-wide significance, although they did report nominal associations for rs6503319 (trend test OR = 1.21, $P = 7.2 \times 10^{-7}$) close to the *SCO1* gene that encodes a protein involved in cytochrome oxidase function; and rs1451375 (dominant model OR = 0.75, $P = 6 \times 10^{-6}$; and rs7803788, OR = 0.76, $P = 2.4 \times 10^{-6}$) intronic to *DDC*, a gene that encodes dopa decarboxylase, which is involved in dopamine and serotonin synthesis(106).

1.2.7. Challenges for GWAS of infectious diseases

Despite the successes of GWAS in infectious disease(99, 102, 105, 108), especially for HIV/AIDS and TB susceptibility, there are limitations that can hinder the advancement in the field. GWAS studies are underpowered to detect any burden of rare variants (MAF < 0.01) and often require very large cohorts to detect association. For infectious diseases such as HIV, even when the influence of host and viral genetics is taken into consideration, over 60% of the variability remains unaccounted for(99). Indeed, the modest contributions of some of the genetic variability and the low frequency of causal variants could explain this missing heritability. Future work will therefore need to employ a more combinatorial approach where genotyping as well whole genome sequencing data is used to detect new associations with lower effect frequencies. For example, in a recent paper exploring the genetic architecture of inflammatory disease (IBD), Luo and colleagues used genotyped data from 27,176 samples (cases and controls) to detect associations with relatively modest OR (1.2 to 2.1)(109). Parallel to this, they also used low-coverage whole genome sequencing across 7,932 individuals to test for a burden of rare variants associated with IBD. They reported several associations including a low-frequency missense variant in *ADCY7* which affects the production of cAMP, a predisposing factor to IBD(110). Although the recruitment of such large cohorts of individuals, especially in low-income countries, will be difficult, it is possible that the gap in recruitment can be reduced through worldwide collaborations. Generally, the greatest burden of disease caused by infectious diseases is in developing countries that lack the necessary resources, or the will to invest in such large-scale programmes. Any programmes that will further our understanding of these diseases and help reduce their burden will be a start.

2. Assessing the coverage of variation in the IFITM locus using commercially available genotype arrays

2.1. Introduction

Genetic studies of susceptibility to infectious disorders aim to provide a greater understanding of disease to reduce mortality and morbidity associated with these conditions(111). Various genome-wide studies have indeed contributed to this aim by reporting new associations between specific genes and particular infectious diseases(112, 113) or replicating previous findings(103). For example, variants in the NOD2 locus have been associated with leprosy risk and a 32bp deletion in CCR5 (CCR5Δ32) has been shown to slow the progression of HIV (99, 112). Despite these successes, however, our understanding of the number of host factors influencing disease outcome to infectious diseases is limited. Notwithstanding functional studies reporting IFITM1,2,3 as important restriction factors against a wide number of viruses, so far no genome-wide association (GWA) studies have reported any significant genetic association in or around these genes. One hypothesis that could explain the lack of GWAS signals is that the variation for the IFITM region is poorly captured by existing genotyping arrays(106). I have set out to test this hypothesis by estimating the tagging efficiency of several commercially available chips, including ones used in previous GWAS of infectious diseases(38, 103, 113). Although earlier studies have highlighted limitations of genotyping arrays at capturing variation at specific loci(114-116), this is the first time a comprehensive analysis of coverage for the IFITM region has been carried out.

2.1.1. Principles of genome-wide association studies

First envisioned by Risch and Merikangas(117) genome-wide association studies are a direct result of two crucial international projects: the Human Genome Project (HGP)(85) and the HapMap project(101, 118). Their work, coupled with technical advances in the chemistry of probe-target hybridisation and amplification techniques in commercial microarray companies(119), made GWAS possible. Initial data from the draft genome constructed by the Human Genome Project identified over 1.4 million single nucleotide polymorphisms (SNPs) and provided the quantification of the extent of linkage disequilibrium (LD) between SNPs in close proximity to each other(85). In such cases, LD can be measured in terms of the squared correlation coefficient (r^2) between the two SNPs. Thus, r^2 is 1 when two SNPs are in complete LD and are not disrupted by recombination. The value becomes less than 1 when correlation between SNPs has been disrupted by crossing over(101).

The HapMap project was a natural continuation of the Human Genome Project. It provided details of correlations between SNPs by studying variation in 270 individuals from West Africa (YRI), Asia (CHB+JPT) and Europe (CEU)(101). Crucially, the HapMap data provided a genome-wide linkage disequilibrium map that contained population-specific haplotype structural patterns(101). Understanding the LD structure enabled scientist to assay genome-wide variation using only a fraction of the total number of variants. Indeed, one important finding from the HapMap project that facilitated genome-wide scans was that 500K SNPs could 'capture' or 'tag' around 80% of common variation (MAF $\geq$ 0.05) in the HapMap Phase II in CEU and CHB+JPT populations, with $r^2 \geq 0.8$; and twice that number (1.09 million SNPs) could capture the same level of variation in YRI(101). The ability to tag SNPs as a consequence of LD, enables scientist to test for association and detect causal SNPs even when these SNPs are not directly genotyped (Figure 7).

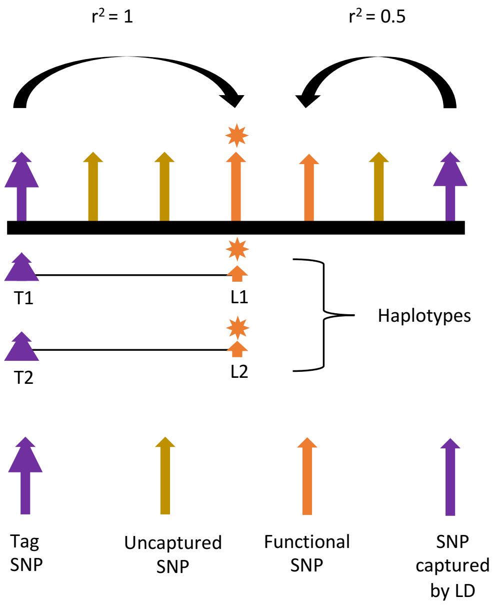


Figure 7. **Principle of microarray tagging**. The schema represents a genomic region that contains 7 SNPs. The 2 SNPs in purple with double triangles are genotyped directly and represent the tag SNPs on the chips. The 2 SNPs in orange are captured through linkage disequilibrium (LD) with the tag SNPs (as denoted by arrows). The 3 SNPs in dark yellow are neither genotyped nor captured by tag SNPs. The orange star represents a SNP associated with disease. It has 2 alleles (L1 and L2) which are in LD with a tag SNP that has 2 alleles (T1 and T2). There is perfect LD between T1-L1 and T2-L2 as measured by the square of the correlation coefficient r^2=1.

The HapMap study constituted a powerful resource for the last 15 years not only for the scientific community, but also for private companies interested in the automation of genome-wide scans. It became common for companies to use this universal reference panel to select marker SNPs for their microarrays. Despite very useful applications, however, one limitation of this dataset was that it only contained approximately 3.5 million variants with MAF ≥ 0.05 which represented 25-35% of common variation in the populations surveyed(101). Recently, more comprehensive reference panels have been constructed that capture over 80% of common genetic variation in the human genome. For example, the release of Phase1(120) and Phase3(121) 1000 Genomes Project (1KGP), UK10K(122) and Haplotype Reference Consortium(86) reference panels, mean that the catalogue of human variation has expanded greatly, with over 8 million variants with MAF ≥ 0.05 already reported, out of a total 9-10 million variants predicted to exist(101).

The availability of reference panels has also contributed to advances in statistical tests such as genotype imputation, that are now commonly used in genome-wide studies(123). Typically, imputation algorithms rely on the identification of haplotypes (how SNPs are arranged along the chromosome) using typed SNPs in the study individuals that can then be used to scan similar haplotypes in individuals in the reference panel. Algorithms use this sharing to predict missing alleles in the study individuals that are not directly genotyped by the microarray(106) but exist in the reference panel within a similar haplotype context. These *in silico* imputed SNPs can boost the power of the GWAS by increasing the number of SNPs that can be tested for association. Therefore, an association signal can result from directly genotyped or imputed SNPs.

2.1.2. Design and coverage of commercial genotyping chips

There are currently two companies, Affymetrix and Illumina that dominate production of genotyping arrays. Typically, commercial companies exploit existing LD information in reference panels to develop their products, geared towards targeting specific populations or functional information and phenotypes. For instance, the Axiom® Genome-Wide PanAFR (Affymetrix, CA, USA) is the first array by Affymetrix to offer genomic coverage (>80%) in admixed populations of African ancestry. Whilst the Metabochip(124) (Illumina, CA, USA) is a custom array design by the Cardio-Metabochip Consortium that targets cardiovascular, metabolic and anthropometric traits. Similarly, the Immunochip consisted of variants selected primarily from the GWAS-associated regions of eleven immune-mediated phenotypes(125, 126) aimed to replicate the top 2000 independent associations found from each of the autoimmune and inflammatory diseases included. All these arrays have been designed to maximise coverage in populations or functional pathway of interest.

To compare genotyping arrays, it is common practice to consider what proportion of SNPs can be directly genotyped by the array and how many variants can be 'captured' by markers in the chip. It has been demonstrated that in theory, 500K maximally efficient tag SNPs could capture nearly 80% of common variation in CEU and JPN + CHB populations and 70% of variation in YRI population at a correlation coefficient ($r^2 \geq 0.8$)(115). It is this proportion of total variants 'captured' or 'tagged' at a given correlation threshold ($r^2 \geq 0.8$) by SNPs in the array which is referred to as *global coverage* of the chip and constitutes one important metric for chip selection and study design.

Typically, estimations of coverage set an arbitrary threshold (usually at $r^2 \geq 0.8$) to find correlations between markers in the genotype array and variants in a specific reference panel (HapMap or 1KGP) using the following formula:

$$cov = \frac{\left(\frac{L}{R-T}\right)(G-T)+T}{G}$$

where (R) represents the number of common SNPs used in the reference panel dataset, (T) the number of SNPs included in the genotyping chips, (L) the number of SNPs not on the chip but tagged at $r^2 \geq 0.8$ by at least one SNP in the chip and (G) the number of common SNPs estimated to be present in the human genome. Other groups have expanded this method(116) to include the extra parameter '*m*' to represent SNPs in the chip not found in the reference dataset. This had advantages for calculations that used the HapMap reference dataset, which contained only a proportion of tag SNPs(116). Their updated formula is represented by $R_1 = R + m$, $T_1 = T + m$ and $L1 = \left(\frac{T1}{T}\right) \times L$ where cov is defined as:

$$cov = \frac{\left(\frac{L}{R1-T1}\right)(G-T1)+T1}{G}$$

Although lack of coverage can be bridged by imputation analysis[32], this is not always possible[33,34]. For example, it has been extensively documented that rare

SNPs (MAF ≤ 1%) are more difficult to impute than SNPs with MAF ≥ 1%. Differences in the genetic make-up of the study populations and the reference panel can also influence the quality of the imputation(106). For example, genetic studies in African populations have been limited not only by the lack of well-designed genotyping arrays but also by the lack of population specific reference panels for accurate imputation(106).

2.2. Aims

To achieve a full understanding of the representation of *IFITM* genes in commercially available arrays, I assessed a subset of Illumina (San Diego, CA, USA) and Affymetrix (Santa Clara, CA, USA) chips for coverage in the region. I also estimated coverage for over 15,000 protein coding genes to have a better understanding of how my estimates of coverage for the *IFITM* genes compared to those obtained for the rest of the genes in the genome. Because imputation is a common technique in current GWAS analysis, I also assessed the quality of imputation for my region and compared it to the imputation quality genome-wide.

2.3. Materials and Methods

2.3.1. Choosing genotyping arrays

I chose to analyse genotyping arrays used in previous genetic studies of infectious diseases (Table 3) for HIV(127, 128), chronic hepatitis C(38) and dengue(113), listed in the GWAS Central http://www.gwascentral.org/ and GWAS catalogue https://www.genome.gov/26525384 on (December, 2015). Because the majority of these arrays have been retired from the market, I also chose to calculate coverage of more recent genotyping chips. Table 3 lists all chips analysed and the number of markers included in each.

The first step of my analysis involved making a list of chromosomal positions of all markers included in the array from http://www.well.ox.ac.uk/~wrayner/strand/, the Wellcome Trust Sanger 'in-house' repositories and http://www.affymetrix.com/catalog/prod350001/AFFY/. All annotation files were updated to Version 3 NCBI Build 37 of the human genome.

Table 3. List of all the genotyping arrays analysed

	Number of SNPs	Targeted MAF	Based on
Illumina			
Illumina 550	547,327	5.0%	HapMap
Human670-QuadCustom_v1_A	654952	5.0%	HapMap
Human660W-Quad_v1	657,366	5.0%	HapMap
Human OmniExpress-24	713,014	5.0%	HapMap
HumanHap 1M-Duo_v3	1,199,187	5.0%	HapMap
Human Omni1S	1,185,076	2.50%	1KGP*, HapMap
Human Omni2.5S-8	2,015,318	1.0%	1KGP
Infinitum Human Omni5-4 v1.1	4,284,426	1.0%	HapMap, 1KGP
Affymetrix			
Affymetrix 500K **	500,568	5.0%	
Affymetrix 6.0	906,600	5.0%	HapMap and previous Mapping 500K and SNP 5.0 Arrays
Axiom® Genome-Wide Pan-African	2,217,402	2.0-5.0%	HapMap, 1KGP, and Southern African Genomes Projects

* 1KGP; 1000 Genome Project

** Tag SNPs on this microarray are randomly distributed across the genome.

2.3.2. Reference panel used in coverage calculations

In this study, I used Phase 1, 1000 Genomes Project reference panel(85) to calculate coverage. This dataset includes the low coverage whole genome sequences of 1,092 individuals from 14 populations across Europe, Asia, Africa and the Americas. Specifically, 286 individuals from the IBS, GBR, TSI, and CEU 1000 Genomes Project formed the European reference population; 286 individuals from the CHS, CHB, and JPT 1000 Genomes Project formed the Asian reference population and 246 individuals from the YRI, ASW and LKW1000 Genomes Project formed the African reference population.

These reference panels provide a haplotype map that includes 38 million SNPs and captures approximately 98% of SNPs at MAF ≥ 1%. I used this reference panel for two steps: to obtain the number of SNPs per population and to find SNPs in LD with markers in the array.

2.3.2.1. Estimating global coverage

Global coverage was estimated using the Barret and Cardon(115) formula and the reference panel previously described. I did not include the extra '*m*' parameter proposed by Li, *et al*, 2008 because the number of SNPs in the chip not found in the reference dataset was very low (< 1%).

$$\frac{(\frac{L}{R-T})(G-T)+T}{G}$$

Where:

T represents the number of common SNPs (MAF ≥ 1%) per population on the genotyping array. I made a list of these positions from the information provided in the annotation files for each chip

L denotes the number of SNPs not on the microarray but which are tagged at r^2≥ 0.8 by at least one marker on the chip within a 1000kb window. To find all proxy SNPs, I used the LD options in PLINK (v1.9). Output SNPs were filtered by MAF (≥ 1%). I also excluded all SNPs included in microarrays as 'tag' SNPs.

R represents the number of autosomal SNPs (MAF ≥ 1%) identified in Version 3 NCBI Build 37 of 1000 Genomes phase 1 project. To calculate the number of variants per population, I used PLINK (v1.9) with the following command: plink --allow-no-sex --write-snplist --geno 0.1 –exclude INDELS.

G represents the predicted number of SNPs. The current number of single nucleotide variants in the NCBI database with MAF ≥ 1% is nineteen million, of which approximately eight million have MAF ≥ 5%(121). This provides an estimate of two SNPs on average per 300bp with MAF ≥ 1% and half that for SNPs with MAF ≥ 5%.

2.3.2.2. Coverage calculation methodology for IFITM2 and IFITM3 and over 15,000 protein-coding genes

Although useful for global coverage, I found that the formula by Barret and Cardon was not adequate for my gene coverage estimations (results not shown) because of the *G* parameter. The G parameter represents the predicted number of SNPs genome wide. For small gene regions, these values can be inflated and can result in values of over 100% for coverage values. For this reason, I decided to use a simplified version of formula as shown below:

Simplified version of Barret and Cardon formula used to calculate gene coverage:

$$\frac{T + L}{R}$$

In this instance, values for G can be ignored because the 1000 Genomes phase 1 panel is thought to contain 98% of SNPs at MAF ≥ 1%(120). I obtained a list of transcription start and end positions for known protein-coding genes from Ensembl Biomart GRCh37.p13 (http://grch37.ensembl.org/biomart/martview/75576048dab692fe6e30bf79259fe775) (February 2015). I filtered this list to include only genes that carried more than five SNPs (MAF ≥ 1%) within their start and end region across all populations. Previous studies had reported that values lower than five tended to affect the coverage calculations(116). I obtained a total of 15,637 protein-coding genes after filtering. Due to low number of SNPs within *IFITM1* gene, I was unable to calculate coverage for this gene. Although my interest is primarily on the *IFITM* gene family, having an estimate for other genes enables coverage estimates for the *IFITM* locus to be compared to other genes in the rest of the genome.

2.3.3. Imputation quality for the IFITM region

The imputation quality metrics reported here were extracted from 'in-house' data generated by the imputation software IMPUTE2 for the Illumina HumanOmni2.5-8 BeadChip, using Phase3 1000 Genomes Project (2,504 individuals, 85 million sites) and UK10K reference panels (3,781, 24 million sites)(129). We looked at two imputation quality metrics: 'INFO score' and 'r2'. The INFO score is a quality score that the imputation algorithm IMPUTE2 generates for each imputed genotype. For directly genotyped SNPs, the 'r2' measurement is the correlation between the directly observed genotype and an imputed genotype at this SNP. Both INFO scores and r2 are used as quality control metrics for genome-wide association studies. The genotype imputation was carried out by Dr. Jimmy Liu at the Wellcome Trust Sanger Institute.

2.4. Results

2.4.1. Estimating global coverage for genotyping arrays

My estimates of global coverage demonstrate that on average, genotyping arrays cover approximately 65% of the variation (MAF ≥ 1%) in European and Asian populations and 30% of the variation in populations of African ancestry (Table 2). The differences in genome coverage between European, Asian and African populations have been reported in previous coverage studies(114-116). Low values of coverage observed for Africans stem for the greater levels of ethnic diversity and complex variation of haplotype structures between ethnic groups(106) in the African population.

I also observe a lower coverage rate for some genotyping arrays than the coverage reported by manufacturers'. For example, I estimated the coverage rate for Axiom Pan-African to be ~40%, half the value reported by Affymetrix. These differences may be the result of various definitions of coverage and may also occur as a consequence of using the HapMap panel instead of a more comprehensive 1000 Genomes Panel to estimate these values. Furthermore, for genotyping arrays with more than one million SNP markers, the Illumina Omni5.4v1.1 and Affymetrix Pan African offer the greatest coverage for all populations analysed.

Table 4. Global coverage estimations for over 15,000 protein coding genes

Populations	Illumina 550	Illumina 660W-Quad_v1	Human670-QuadCustom_v1_A	Human OmniExpress-24	HumanHap 1M-Duo_v3	Human Omni1S_H	Human Omni2.5S-8_B	Infinitum Human Omni5-4 v1.1	Affymetrix 500	Affymerix SNP6.0	Axiom PanAfrican
European	0.56	0.66	0.59	0.64	0.68	0.64	0.61	0.84	0.50	0.64	0.73
Asian	0.62	0.70	0.63	0.67	0.69	0.63	0.55	0.75	0.52	0.65	0.73
African	0.21	0.25	0.22	0.25	0.30	0.26	0.28	0.56	0.19	0.28	0.47

2.4.2. A map of genome-wide gene coverage

I estimated the coverage of *IFITM2* and *IFITM3* in eight Illumina genotype arrays and three Affymetrix chips (Table 5). In order to ascertain how the coverage for the *IFITM* genes compared to other genes genome-wide, I also calculated the coverage of a further 15,635 protein-coding genes. The motivation behind these calculations was to assess how the coverage for *IFITM2* and *IFITM3* compared to other genes genome-wide.

I found large differences in coverage across populations with most genotyping arrays performing poorly in African admixed populations. For example, less than 25% of common SNPs located within the gene regions analysed, can be tagged by Affymetrix 500K array ($r^2 \geq 0.8$). These values improve in denser genotyping arrays such as the HumanOmni 5.4 (~5 million SNPs). This array captures 100% of the variation identified by the 1000 Genomes Project in around 90% of protein-coding genes (Figure 8).

In contrast, most of the variation in Asian and European populations is captured by all genotyping arrays. For the Asian and European samples, only 10% of genes have coverage <25%. As expected, denser genotyping arrays such as the HumanOmni 2.5 (2.5 million SNPs) and HumanOmni 5.4 (~5 million SNPs) captured ~100% of the variation in 80-90% of protein-coding genes (Figure 9 and Figure 10).

This analysis demonstrates that *IFITM2* and *IFITM3* are in bottom 7% of all protein-coding genes analysed. Both Illumina and Affymetrix arrays captured only 25% of common variation within these genes (Table 5). Coverage was particularly poor (6-12%) on the Illumina 550K, 660 and 670, as well as Affymetrix 500K genotyping arrays. Interestingly, all of these arrays have been used in previous genome-wide association studies of HIV(103), dengue(113) and Hepatitis C(130). Despite the potential important role of *IFITM* genes in these diseases, none of the genotyping arrays tag more than 50% of SNPs in *IFITM2* or *IFITM3*.

Coverage analysis shows that denser genotyping arrays such as HumanOmni2_5, HumanOmni5_4 and Affymetrix Axiom Pan African provide better coverage across populations. Close examination of tagging SNPs included in these arrays, shows that the increase in coverage is the result of introducing a greater number of tag SNPs in the *IFITM* region. This suggests that in order to capture variation in the *IFITM* region, SNPs would need to be genotyped directly; undoubtedly limiting the usefulness of using tagging strategies for this locus.

It is common practice as part of the GWAS analysis to use imputation to boost statistical power. Simulations studies have shown that imputation can improve the performance of genotyping arrays even when the coverage of such arrays is low (<50%). Imputation methods such as IMPUTE2 provide a probabilistic prediction at each imputed genotype given by the 'INFO' score(123) that allows researchers to filter out poorly imputed sites. Another imputation quality assessment commonly used, involves measuring the squared correlation between the best-guess genotype and the true genotype(131) to give a single measure r^2. I used both these metrics to assess the imputation quality in the *IFITM* region using 'in-house' imputation data for the Omni2.5 array. I found that imputed SNPs in the region are indeed of lower quality compared with the genome average throughout the allele-frequency spectrum (Figure 11).

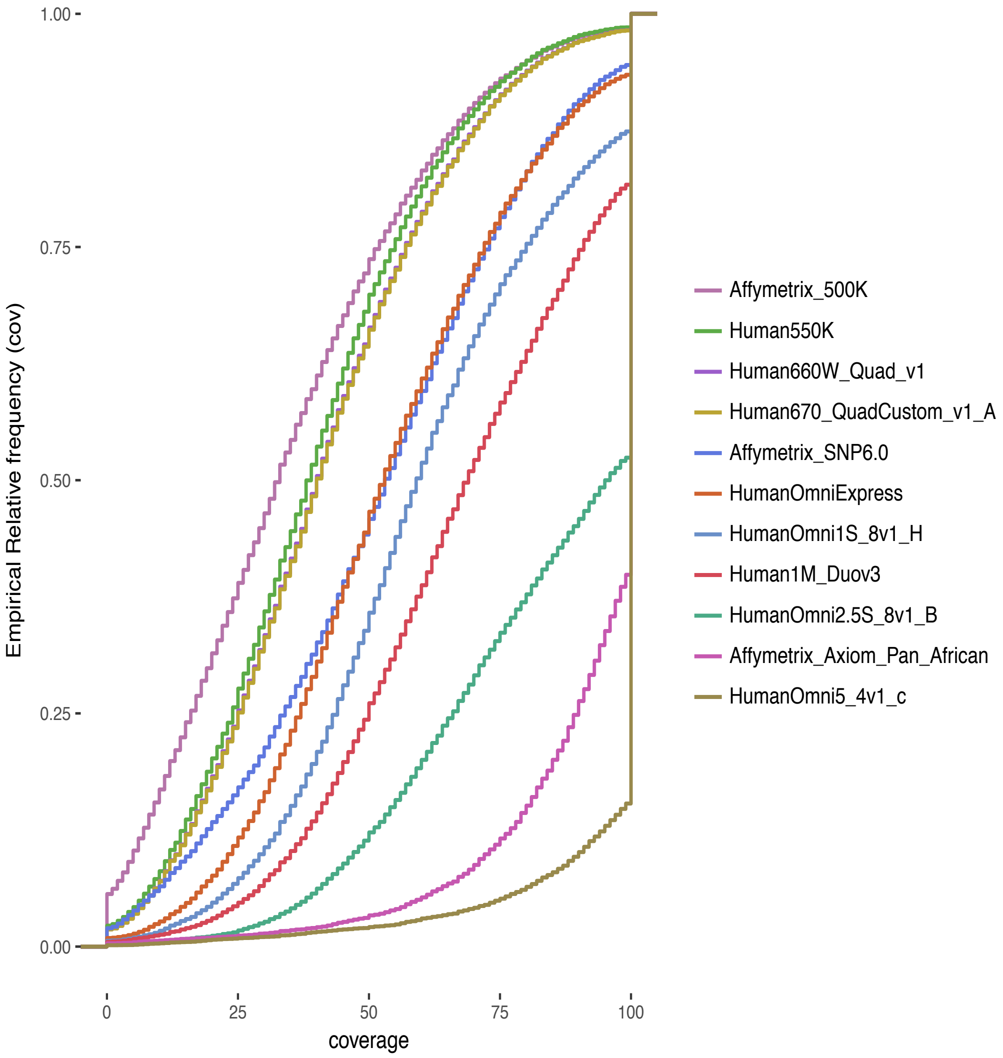


Figure 8. **Distribution of coverage for 15,637 protein-coding genes for 11 genotyping arrays in 246 African individuals from the YRI, ASW and LKW 1000 Genomes Project populations.** The empirical distribution is plotted in the y-axis and the coverage (%) is plotted on the x-axis. The names of the genotyping arrays are shown to the right ordered by levels of coverage. The array with the lowest coverage at the top and the array with the highest coverage at the bottom. Affymetrix 500K shows the lowest coverage with 50% of genes having <27% coverage. HumanOmni5_4 array shows the highest coverage, with 90% of genes having 100% coverage.

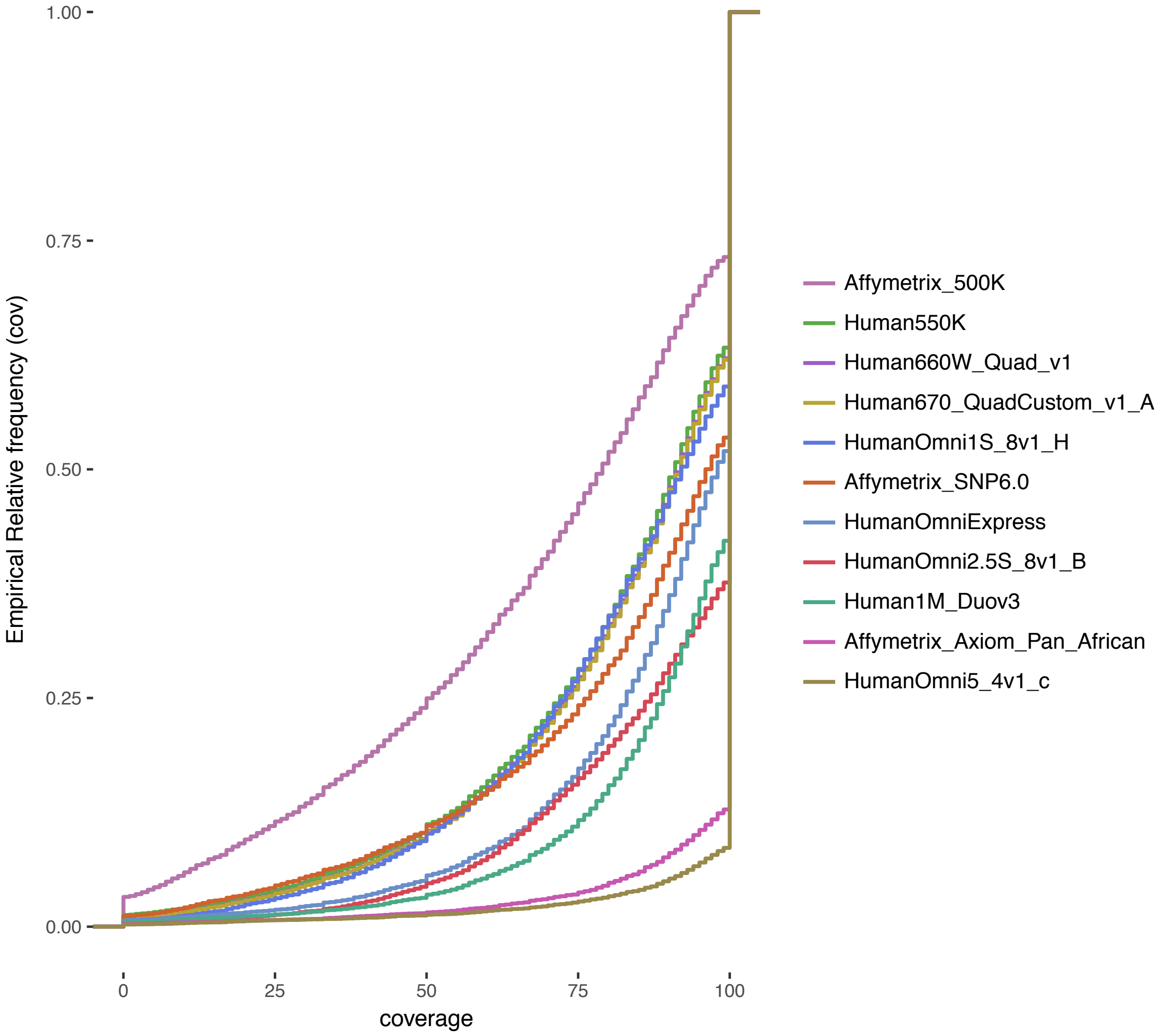


Figure 9. **Distribution of coverage for 15,637 protein-coding genes for 11 genotyping arrays in 286 Asian individuals from the CHS, CHB and JPT 1000 Genomes Project populations.** The empirical distribution is plotted in the y-axis and the coverage (%) is plotted on the x-axis. The names of the genotyping arrays are shown to the right ordered by levels of coverage. The array with the lowest coverage at the top and the array with the highest coverage at the bottom. Affymetrix 500K shows the lowest coverage, although it performs substantially better than in African populations with only 10% of genes having <27% coverage. HumanOmni5_4 array shows the highest coverage with 90% of genes having 100% coverage.

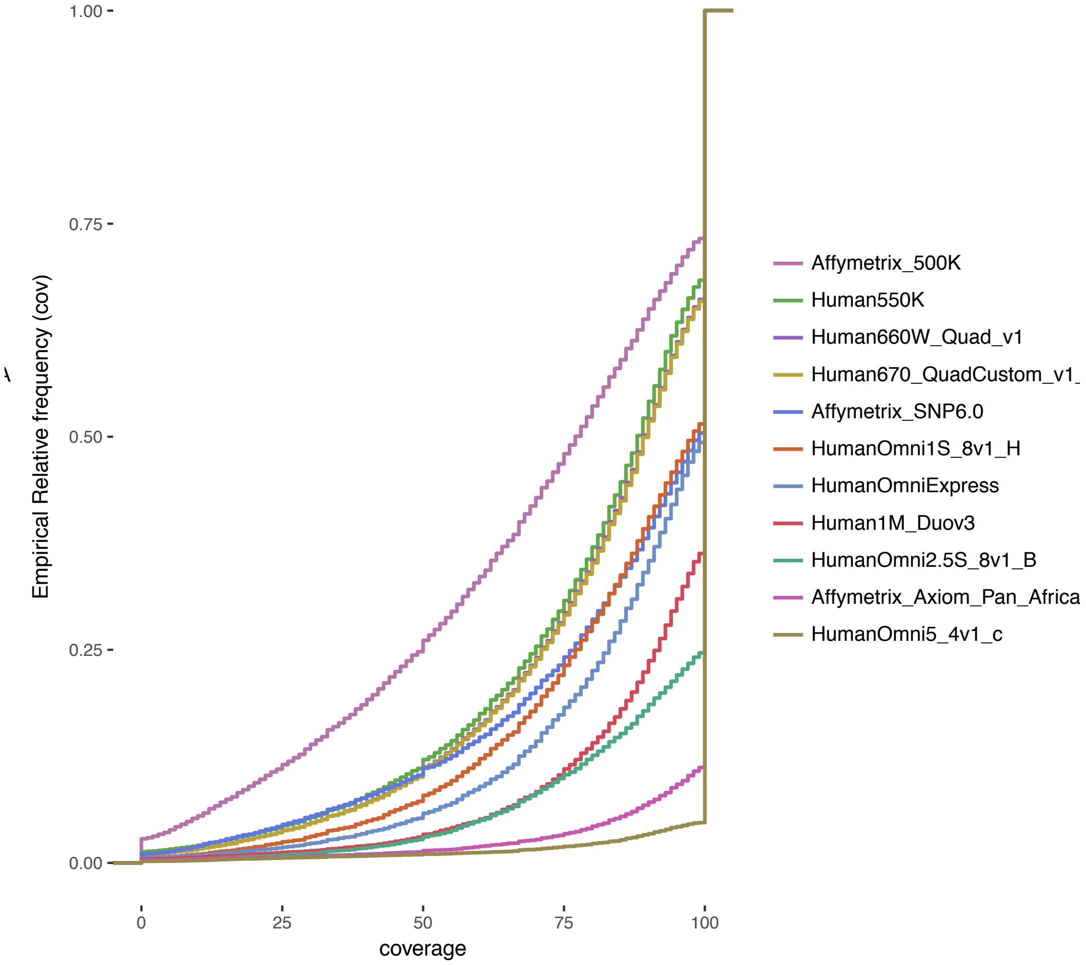


Figure 10. **Distribution of coverage for 15,637 protein-coding genes for 11 genotyping arrays in 286 European individuals from the IBS, GBR, TSI and CEU 1000 Genomes Project populations.** The empirical distribution is plotted in the y-axis and the coverage (%) is plotted on the x-axis. The names of the genotyping arrays are shown to the right ordered by levels of coverage. The array with the lowest coverage at the top and the array with the highest coverage at the bottom. Similar to African and Asian populations, Affymetrix 500K shows the lowest coverage, although it performs substantially better than in African populations with only 10% of genes having <27% coverage. HumanOmni5_4 array shows the highest coverage, with 95% of genes having 100% coverage.

Table 5. Coverage calculations for *IFITM2* and *IFITM3*

Genotype Platform	African	Asian	European	African	Asian	European
Illumina 550	6	12	9	0	0	5
Illumina 660W-Quad_v1	6	12	9	2	2	8
Human670-QuadCustom_v1_A	6	12	9	2	2	8
Human OmniExpress-24	6	12	12	8	14	27
Human Omni1S_H	6	22	18	10	20	15
HumanHap 1M-Duo_v3	20	34	26	18	22	39
Human Omni2.5S-8_B	16	34	24	5	6	15
Infinitum Human Omni5-4 v1.1	16	34	29	18	24	40
Affymetrix 500K	0	0	0	0	0	0
Affymetrix SNP6.0	30	44	32	12	8	24
Axiom_Pan_African	8	28	18	6	20	10

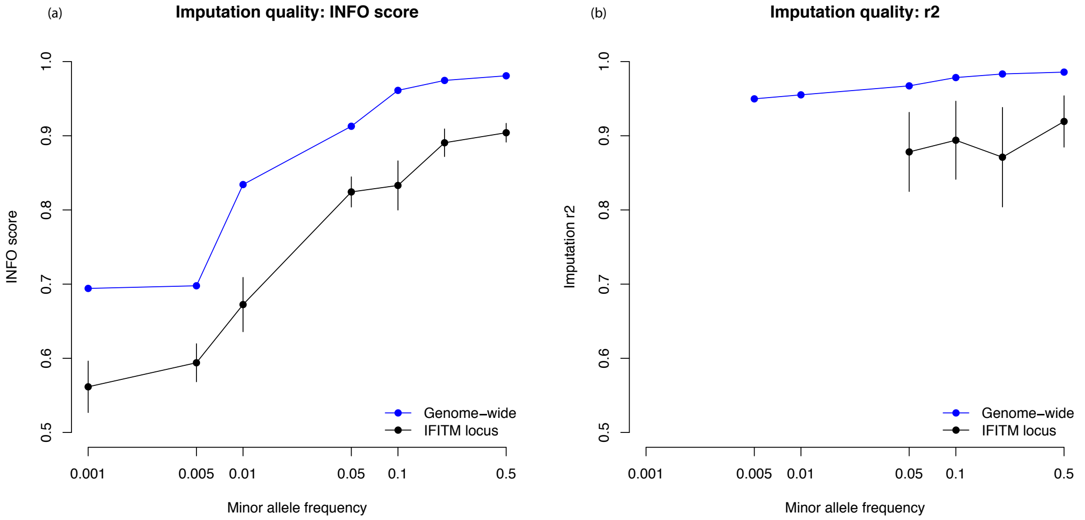


Figure 11. **INFO scores and r^2 imputation quality metrics for the *IFITM* region.** (a)The black line represents the average INFO scores for different allele frequency bins in 6,000 European individuals genotyped using Omni2.5 array from a recent GWAS of Primary Sclerosis Cholangitis(129). The INFO scores for these individuals is 1.5X less than the predicted genome-wide INFO scores for SNPs at the specified allele frequency bins. (b) The black line represents the average r^2 scores for different allele frequency bins in the same 6000 individuals. Similar to the INFO scores, the r^2 is 1.5X less than the predicted genome-wide r^2 scores for SNPs at the specified allele frequency bins. Error bars represent the standard error of the mean for the INFO and r^2 values.

2.5. Discussion

The main conclusion from this study is that *IFITM2* and *IFITM3* genes are poorly covered by all existing genotype arrays. As a consequence, common variant associations with infectious disease phenotypes within these genes may have been missed via GWAS.

I screened several commercially available genotyping arrays to estimate the coverage across African, Asian and European populations and found that the average coverage for *IFITM2* and *IFITM3* is ≤ 25% across all populations, placing both these genes in the bottom 7% of protein-coding genes analysed. This reflects the fact that a significant proportion of tag SNPs do not have strong correlation with other single nucleotide polymorphisms in the region and therefore are unable to tag most of the *IFITM2* and *IFITM3* common SNPs.

In order to calculate coverage, one makes the assumption that all SNPs in the array will pass the different QC steps required for GWAS analysis(132). Although QC thresholds are subjective, markers can fail QC due to a variety of reasons including differences in genotype call rates between cases and controls or excessive missing data rate(132). A consequence of this is that values reported in this study may constitute an overestimation of true coverage. Most importantly, having good coverage does not necessarily mean that there is good power to detect associations. In cases where causal SNPs have large effect sizes, an association signal can be detected even when the correlation $r^2 < 0.8$(133).

Although coverage is boosted in genotyping arrays that carry increased number of tag SNPs, the implication of this is that important genetic signals may be missed if the variant is not directly genotyped. The hemoglobin S (*HbS*) locus is a classic example where a similar situation has been observed. *HbS* is a well-known determinant of risk for Malaria with protective effects conferred by the causal SNP rs344 . This variant is located in chr11, in the coding region of the beta globin gene

(*HBB).* In a landmark study of Malaria in The Gambia, it was observed that association signals with rs344 barely reached genome-wide significance (P value=3.9x10^{-7}) when genotyped with Affymetrix 500K. Close examination of the region revealed that there were no strong correlations between tag SNPs in the array and the rs344 variant. The SNP rs344 was not directly genotyped in that study (i.e. not a tag SNP in the array), therefore the power to detect any associations with it decreased substantially (123).

Imputation methods are commonly used in GWA analysis to boost power to detect association(106, 123, 133). They rely on reference panels to 'fill in' gaps for variants not included in the genotyping arrays. Strict quality controls are therefore required in order to reduce the false positive signals from these *in silico* genotypes. One such quality measure is the Info score that is automatically generated by some imputation softwares (134). By looking at the info score for my region of interest, I found that the quality of imputation given by the IMPUTE2 INFO scores is 1.5X lower for the *IFITM* locus compared to the rest of the genome. This trend is observed across the full spectrum of allele frequencies (0.001 ≤ MAF ≤ 0.5). Although the quality of the imputation is dependent on factors such as the quality of the phasing and regional haplotype structures(123, 134), our results highlight the difficulties of capturing the full extent of variation for the region even after imputation.

Lastly, this study provides proof of principle that the lack of GWA signals can at least in part be explained by lack of tagging efficiency of the genotyping arrays. This work also emphasises the importance of understanding local variation in haplotype structures of regions of interest. For the *IFITM* locus, the low coverage of genotyping arrays means that the full variation of the region cannot be analysed, thus signals of association in the region could be missed.

Given that a significant proportion of variants have no strong marker SNPs, imputation analysis would be crucial for association analysis. Although the data to assess imputation quality in this study is limited to only one genotyping array, I show that imputation quality is lower for the region. As a consequence, genotyping is not the best experimental design to test for association in this locus unless better imputation panels become available or more SNPs are included in the genotyping arrays. Other approaches such as direct sequencing could possibly overcome the problems of this region.

3. Targeted sequencing of the *IFITM* locus

3.1. Introduction

In the previous chapter, I established that *IFITM2* and *IFITM3* are amongst 7% of all protein coding genes with less than 25% common variant (minor allele frequency > 0.05) coverage across all arrays. Furthermore, all attempts to characterise *IFITM* variation by sequencing have focused on sanger sequencing one SNP (rs12252), located at 5' end of *IFITM3*.

In this chapter, I employ a targeted sequencing method using two different sequencing technologies: Illumina MiSeq and PacBio sequencing to characterise *IFITM* variation. I test this method on nine lymphoblastoid cell lines (LCLs) that had been previously screened for *IFITM3* SNP rs12252. In the following sections, I present some background information on two of the main target enrichment methods; and explain the technical aspects of the sequencing platforms used in this project. The ultimate goal of this study is to develop the right framework for future targeted sequencing of the *IFITM* locus.

3.1.1. Hybrid enrichment methods for targeted sequencing

Over the past decade, next-generation sequencing has revolutionised the field of genetics and facilitated the discovery of loci associated with a number of complex diseases(126, 129, 135). There are a number of sequencing techniques currently available to scientists and the decision to employ a particular technique over another depends on several factors such as the number of samples, the length of the region of interest and the overall aims of the study. For example, sanger sequencing (capillary electrophoresis) is commonly used for the screening and validation of a small number variants in relatively few regions (<20)(136, 137). In contrast, whole exome and genome sequencing are employed in larger sequencing projects when scientists require a genome-wide understanding of variation(138-141). Although sequencing costs have substantially fell in the past few years, it is still relatively expensive to sequence and analyse whole genomes. As a consequence, targeted sequencing provides scientists with the opportunity to scan particular regions of the genome at a relatively low cost. Most importantly, it has the advantage that it can provide depths of 1000x and even higher for the targeted regions(138, 142-145).

There are two main target enrichment technologies, the 'array-capture' and 'in-solution' based methods. In the array-based methods, probes immobilised on a microarray chip hybridise to a fragment library, non-specific fragments are washed away and the targeted DNA is eluted. Roche NimblGen first adapted the technology to work in high-throughput sequencing studies(146) back in 2007. Their high-density microarrays carrying more than 250,000 oligonucleotides >60bp in length spaced between 1-10 bases apart, were tailored to target over 600 genes dispersed throughout the human genome. Their technology was originally designed for the Roche 454 sequencer but many groups worked to adapt it to other technologies such as Illumina sequencers(147). Although the method was undoubtedly an improvement on other enrichment protocols at the time such as

PCR-based enrichment, there were disadvantages. For example, in array-capture experiments, all enriched library DNA had to be eluted at the same time, limiting the number of arrays a person could physically carry out in a day. Furthermore, the excess of DNA libraries over the number of probes, meant that in order for the hybridisation reaction to occur, the starting material had to be in excess of 8μg.

The solution-based method, on the other hand, was developed by Agilent to overcome some practical difficulties of microarray-based enrichment. For instance, Agilent reduced the amount of starting material by a factor of 16 by having an excess of probes over the template, driving the hybridisation reaction to completion. In addition, they substantially reduced the number of oligonucleotides required for the enrichment, by designing sets of probes that tiled across the region of interest. In their initial pilot study they used 22,000 bait sequences of 170bp in length to target a total of 1,900 genes in the human genome(148) . This represented approximately 11x less probe sequences than the number used by NimblGen at the time. In addition, performance comparisons between the array and solution-based methods established that the former provided a more uniform depth of coverage(143).

Currently, there are number of solution-based capture methods including the Agilent SureSelect and NimbleGen SeqCap EZ. Although recent studies show that Agilent has the highest target enrichment efficiency and highest accuracy for SNP detection(140, 149), most studies have highlighted that both technologies deliver similar sensitivity for single nucleotide polymorphism detection(138, 142, 150). The main differences between both methods lie on the type and length of probes. For example, Agilent uses 120bp RNA probes whereas NimbleGen uses 60-90bp DNA probes. Both technologies however, rely on hybridisation reactions to enrich for a region of interest (Figure 12). The 'enriched' DNA is then isolated from the rest of the genomic sequences, using streptavidin beads. The DNA is PCR-amplified and sent for sequencing

Importantly, both array and solution-base captures have been adapted for technologies that required PCR amplification steps for their library preparation (Illumina MiSeq). Only recently, however, several laboratories have started to adapt the targeted sequencing method to make it compatible with other technologies such as PacBio *RS* (151, 152) to define structural variation in regions of the genome or to use for *de novo* assemblies of repetitive plant genomes. Typically, PacBio has been used successfully to characterise regions with high levels of sequence similarity that cannot be accurately mapped with short-read technologies (153, 154)

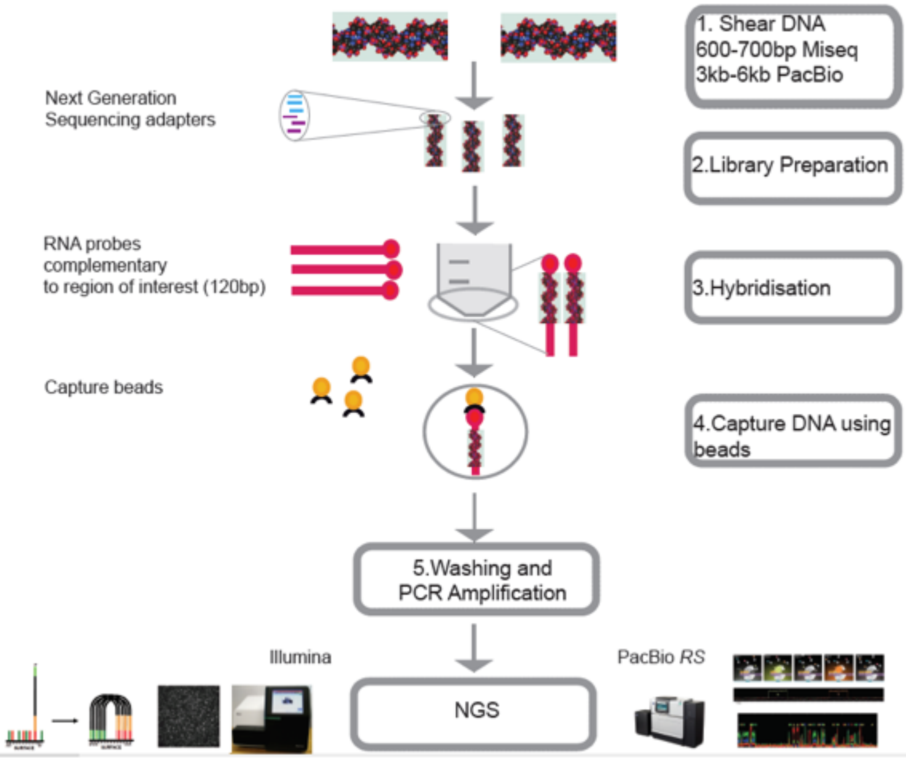


Figure 12. **A schematic representation of a target enrichment method.** (1) Genomic DNA is sheared to 600-700bp (Illumina) and 3kb (PacBio). (2) For the library preparation, the DNA fragments are end-repaired, extended with 'A' bases to the 3' end of the DNA fragments, ligated with paired-end adaptors and PCR amplified. (3) The adaptor-ligated libraries are hybridised to oligo RNA probes for 24hrs and enriched by 'pulling down' with Streptavidin-coupled Dynabeads (4) Libraries are PCR amplified and (5) enriched libraries are processed and sent to appropriate sequencing platforms.

3.1.2. Next generation sequencing approaches: sequencing by synthesis

Currently, technologies can be divided into two main groups: those that use a sequencing by synthesis strategy (Illumina) and others that specialise in single molecule sequencing (PacBio). Sequencing by synthesis is a term that describes various sequencing methods that are dependent on DNA-polymerase activity. First, fragmented PCR-amplified DNA molecules, carrying sequencing adapters, are bound to immobilised primers in a flowcell. The bound DNA template contains a free end; therefore, it can interact with other immobilised primers nearby, forming a bridge structure. PCR is then used to create a second strand from these templates and start a process known as solid-phase amplification (Figure 13). For Illumina sequencing, the technology relies on a cyclic reversible termination (CRT) method that resembles the principles of Sanger sequencing(155). During each cycle, a single fluorescently labelled nucleotide is incorporated to the new strand. Following this incorporation, the fluorophore is imaged and then cleaved to prevent other nucleotides from occupying the same position (Figure 13). It is this CRT method that enables Illumina to have such high accuracy rate (>99%)(156). Nonetheless, this platform displays AT(136, 157) and GC(158, 159) bias and can underperform in repetitive regions. For example, sequencing analysis of the AT-rich *P. falciparum* genome revealed that Illumina provided 10x less coverage through AT-rich regions compared to PacBio whilst in the GC-rich *R. sphaeroides* genome, Illumina provided 54x less coverage than PacBio(160). As a consequence, many groups interesting in sequencing the genomes of organisms with very low or very high GC content, tend to use sequencing platforms such as PacBio *RS* to overcome these limitations.

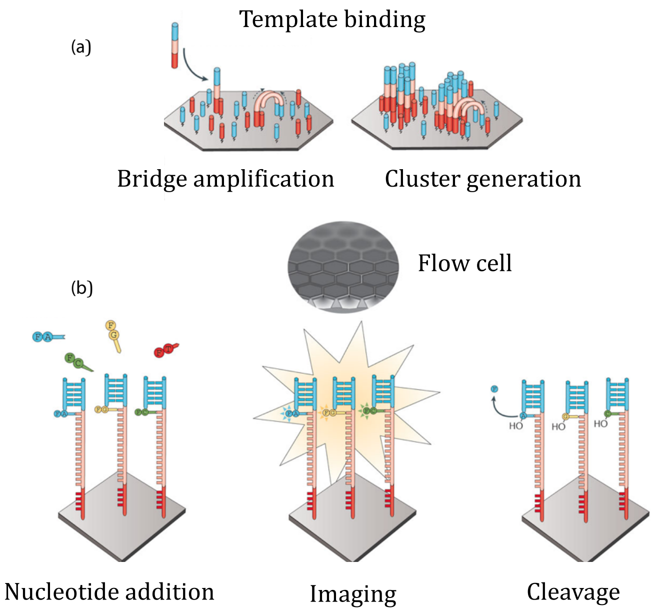


Figure 13. **Next-generation sequencing by solid based amplification** (a) For Illumina sequencing fragmented DNA templates are ligated to immobilised primers on a flow cell. The bound fragments are amplified using nearby primers. (b) Soon after amplification, a mixture of DNA primers, polymerase and four fluorophores-labelled nucleotides (F-A, F-T, F-G, F-C) are added to the reaction. All bases contain a specific cleavage fluorophore (F) that is imaged once it hybridises to the newly form strand. Figure adapted from Goodwin, *et al.*, 2016 (155).

3.1.3. Next generation sequencing approaches: single molecule sequencing

Long-read sequencers such as PacBio, use a single-molecule real-time (SMRT) sequencing approach that does not require PCR-amplified DNA libraries. Instead, fragmented DNA is capped by hairpin loops at either end, and hairpin adapters are ligated to these loops to provide the binding site for DNA polymerase (Figure 14). Unlike the Illumina methodology, where the DNA polymerase travels along the template, PacBio polymerase has a fixed position at the bottom of specialised flowcell wells known as zero-mode waveguides (ZMW). The stationary position of the DNA polymerase means that the system focuses on a single molecule at a time and circular topology of the SMRT-bell template also allows the polymerase to have multiple passes along the DNA molecule (Figure 15). These multiple passes are used to generate high quality consensus reads of insert (ROI) with up to 1% error rate(155).

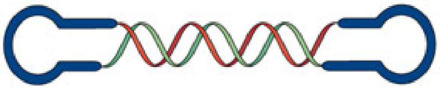


Figure 14. **SMRTBell template**. Hairpin adaptors (blue) are ligated to the end of a double-stranded DNA molecule (red and green). The DNA polymerase is bound to the blue adaptors and anchored at the bottom of the zero-mode waveguide (ZMW)in the SMRT cell. Figure adapted from Goodwin, *et al.*, 2016(155).

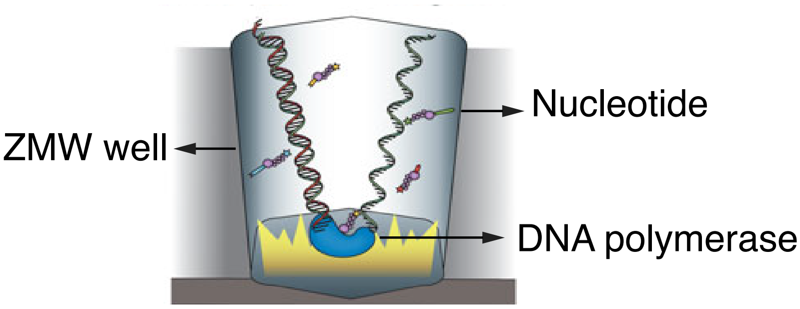


Figure 15. **Single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio**). SMRTbell libraries are loaded to a SMRT cell that contains tens of thousands of zero-mode waveguides (ZMW). Each ZMW well is illuminated from below, but the light wavelength is too large to go all the way through the well. Thus, attenuated light penetrates the lower 20-30nm of the well and creates a detection volume. A DNA template-polymerase complex is immobilized at the bottom of this ZMW well, in the detection volume. A mixture of labelled nucleotides is added to the SMRT cell and as each nucleotide is held at the detection volume by the enzyme's active site, a light pulse is produced. During the nucleotide incorporation, the phosphate group is cleaved and the fluorophore attached to the nucleotide diffuses away. This process occurs in parallel within each of the tens of thousands of ZMW wells that make up the SMRTcell. Figure adapted from Goodwin, *et al.*, 2016(155).

The PacBio *RS* has the advantage of producing long sequencing reads. For example, it can generate average polymerase reads of 10-15kb in length which are ideal for *de novo* assemblies of genomes that are generally difficult to assemble with shorter reads. For example, PacBio sequencing data of AT-rich genome of the *P. malariae* species, enabled the successful assembly of this genome using only 63 scaffolds with average length of 2.3Mb. In contrast, previous draft genomes for this same species using Illumina short reads required 7,270 scaffold with average length of 6.4kb(161).

In addition, long PacBio reads have helped solve gaps and complex long-range genomic structures in the human genome. For example, using *de novo* assembly of a haploid cell line, Chaisson, *et al.*, 2015 closed 55% of sequencing gaps in the human GRCh37 reference genome. The majority of these gaps represented repetitive elements in GC-rich regions of the genome that had remained inaccessible to short-read sequencing technologies(154). This assembly also identified 47,238 breakpoint positions that resolved over 25,000 euchromatic structural variations (SVs), including inversions, large deletions and repetitive regions(162).

Despite these advantages, however, there are notable limitations of the PacBio *RS* technology. For instance, single-pass PacBio reads contain approximately 12% insertion errors, 2% deletions errors and 1% mismatch errors(163). Notwithstanding, Carneiro, *et al.*, showed that these errors were randomly distributed within each read and that read length and base position did not have an effect on base quality, demonstrating that with a high enough coverage (10X), it is possible to overcome such disadvantages(163).

3.1.4. Why use longer reads to sequence the *IFITM* region

Over the past few years, several *in vivo* and *in vitro* studies have demonstrated the important role of the *IFITM* genes as restriction factors of a number of infectious diseases(55, 73, 74, 77, 164). However, no genome-wide association studies have reported any significant associations in or around these genes. Close examination of my region of interest, using publicly available whole genome dataset (75bp paired-end reads), shows that there are regions (>130bp) within the *IFITM* locus that have no depth of coverage (Figure 16). In addition, using whole genome PacBio sequencing data (Sample ID NA12878), I observed that PacBio reads were able to span those regions. As a result, I hypothesised that long sequencing reads could help resolve some of the observed sequencing gaps in the *IFITM* locus.

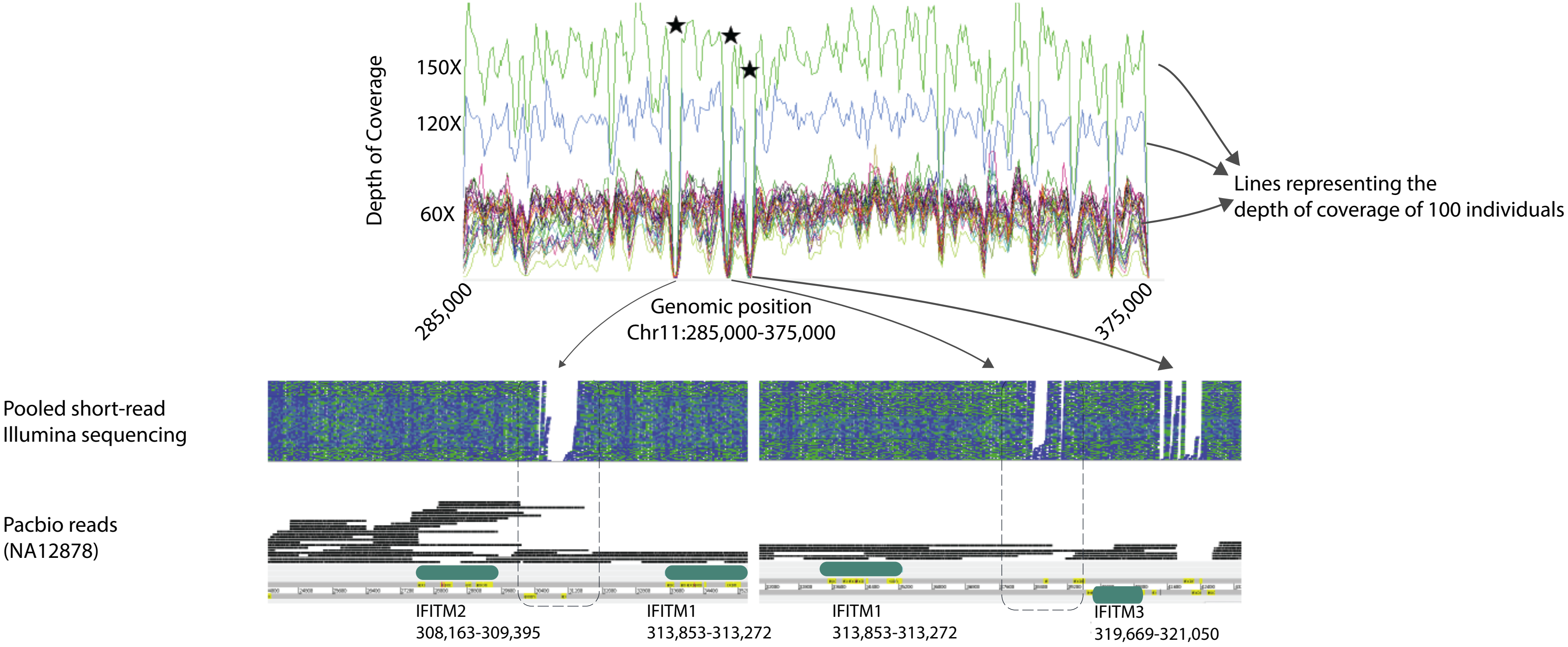


Figure 16. **Depth of coverage for the *IFITM* region.** Each coloured line on the top panel represents coverage of 100 cancer individuals that had been whole genome sequence at a coverage of 60x. Two samples (represented by a blue and green lines) had higher genome coverage of 120X and 150X, respectively. The stars highlight three regions near the IFITM genes where coverage is zero. The middle panel shows pooled reads from these samples (green and blue) showing the gaps of coverage near *IFITM* genes. The bottom panel shows PacBio reads for a sample NA12878. Whole genome sequencing of NA12878 with PacBio *RS* demonstrates that long PacBio reads can span regions difficult to sequencing using Illumina technology.

3.2. Aims

The aim of this pilot study is to conduct, for the first time, a comprehensive screening of *IFITM* genetic variation in nine healthy population samples. As part of this work, I will assess the performance of two sequencing platforms and their ability to capture variants in the region.

3.3. Statement of work

All the work presented in this chapter is my own work unless otherwise stated. This research was also carried out under the supervision and guidance of Dr. Thomas Dan Otto and Dr Paul Coupland at the Wellcome Trust Institute (WTSI)

3.4. Methods

3.4.1. Cell culture and DNA extraction

Lymphoblastoid cell lines (LCL) from nine unrelated individuals (Table 6) included in the 1000 Genomes Project and distributed by Coriell Cell Repository were selected for this study. These cells were maintained in RPMI 1640 plus Glutamax medium (61870-010, Invitrogen) supplemented with 20% v/v Fetal Bovine Serum (FBS Biosera) and incubated at 37^{o}C in 5 % CO2. LCLs were passaged twice a week (1:10 split). Genomic DNA was extracted from these LCLs using Blood and Cell culture kits (Qiagen, Germany) following the manufacturer's instructions. Because the purified DNA was subsequently used for targeted sequencing of the *IFITM* locus with PacBio *RS,* harsh vortexing was avoided throughout the DNA extraction.

Table 6. Table representing details of the DNA samples used in this study

Samples	Population code	Description
NA11994	CEU	Utah residents with Northern and Western European ancestry
NA12154	CEU	Utah residents with Northern and Western European ancestry
NA12155	CEU	Utah residents with Northern and Western European ancestry
HG00524	CHS	Han Chinese South
HG00478	CHS	Han Chinese South
HG00530	CHS	Han Chinese South
HG00533	CHS	Han Chinese South
HG00557	CHS	Han Chinese South
HG01108	PUR	Puerto Rican in Puerto Rico

3.4.2. Probe design

This study was carried using SureSelect biotinylated 120bp RNA probes. In the first phase, I designed probes (ELID 0604211) that covered approximately 96% of the *IFITM* locus from Chr11:280,000-380,000, including repetitive regions. I wanted to capture most of the locus (including repeat regions) and so I purposely used the least stringent repeat masking in my designs. To refine the probe library, I excluded probes with 95% and greater sequence similarity with off-target sequences in the genome. In this first study, I only used three of the nine selected samples (Table 7).

Unsurprisingly, the low pulldown efficiency shown in Table 7 is the result of including probe sequences that span repetitive regions. Following this first pilot, I established that 100% similarity of a continuous 25bp region was sufficient for off-target hybridisation (Personal communication with Agilent representatives).

For the second phase of the study, I used the same probe library as a backbone and excluded a number of probes with 'hits' on several chromosomal regions of the genome. The final probe design for the region (ELID 0695421), still contained probe sequences spanning repetitive regions but in less numbers. This left a total of 3,198 probe sequences, covering around 81.6% of my region of interest (Chr11:280,000-380,000). All the data shown in this chapter is the result of this last pulldown pilot experiment.

Table 7. Table representing the pulldown efficiency of three initial samples used in the study

Samples	Pulldown efficiency (%)	
	Pacbio	Illumina MiSeq
NA12154	4.4	5.3
NA12155	6.0	1.3
HG00524	5.1	1.7

3.4.3. Target enrichment of the IFITM region to sequence with Illumina

Library preparation and SureSelect targeted sequencing for Illumina was performed by Sara Widaa at the Sanger Institute. Briefly, genomic DNA was sheared using a Covaris S2 (Covaris, Inc., Massachusetts, USA) to obtain 700bp fragment libraries. Illumina library preparation and probe hybridisation (capture baits ELID number 0695421) was carried out following Agilent SureSelect ^{XT} Target Enrichment System for Illumina Paired-End Sequencing Library. Sequencing data generated from these runs were released in the form of fastq files by the Pathogen Sequencing Informatics team at the Sanger Institute.

3.4.4. Target enrichment of the IFITM region to sequence with PacBio

The steps below describe the method I developed to prepare the samples for enrichment and subsequent sequencing with PacBio *RS.* My method is based on the Agilent SureSelect^{XT} Target Enrichment System for Illumina Paired-End Sequencing Library, Illumina HiSeq and Miseq Multiplexed Sequencing Platforms. Protocol: Version 1.6, October 2013. There are four main workflows in the method:

1. Preparing the genomic DNA for hybridisation
2. Hybridising the genomic DNA to the custom probes
3. Post-amplifying the hybridised libraries
4. Constructing SMRTbell libraries.

1. Preparing the genomic DNA for hybridisation

Shearing

I sheared the genomic DNA using a Covaris S2 (Covaris, Inc., Massachusetts, USA) at a 20% duty cycle, level 5 intensity and 200 cycles per burst for 600s to obtain 3kb average fragment lengths (Figure 17).

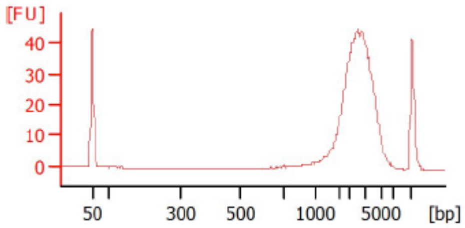


Figure 17. **Electopherogram of fragmented DNA**. This Electopherogram shows the size distribution of DNA fragments following shearing with Covaris S2 ultrasonicator. The peaks at 50bp and 10,000bp represent the DNA ladder. The wider peak with sizes ranging from 2,000-6,000bp represent the DNA fragment of interest.

Preparation of samples for hybridisation

I purified and size selected the DNA by adding 0.6X Agencourt AMPure beads (Beckman Coulter) to 120μl of sheared genomic DNA. The quality of the sheared libraries was assessed by running samples on the Agilent 2100 Analyser (Agilent Technology, California, USA). This was followed by end repair steps, dA-tailing of 3' ends of the genomic DNA and ligation of sequencing adaptors following the SureSelect ^{XT} Target Enrichment System instructions (Agilent Technology, California, USA).

Amplification of DNA libraries

The PCR conditions for library amplification were optimised to fit the fragment sizes of the genomic DNA (Table 8) to enable the amplification of fragments ≥ 3kb with the lowest number of cycles. The low number of cycles ensures the minimal number of PCR type errors being introduced. For the amplification reaction, I used Herculase II Fusion DNA Polymerase enzyme with the following conditions:

Table 8. Optimised PCR conditions for PCR amplification of libraries

Steps	Temperature [^{o}C]	Time
1	98	2 min
2	98	30 secs
3	65	30 secs
4	72	3 min + 10 secs
5		(repeat step 2-4 for 6 cycles)
6	72	10 min
7	10	Hold

2. Hybridisation of genomic DNA and probe libraries

I prepared each DNA library to a concentration of 221ng/µl, in a final volume of 3.4µl. I then added 5.6µl of SureSelect block mix to avoid hybridisation of probes to library indexes. This was followed by the addition of 40µl of hybridisation buffer and 2µl of SureSelect RNA probes (capture baits ELID number 0695421). The hybridisation reaction was left on a thermal cycler at 65°C. After 18hrs, I added 200µl of a mixture of Dynabeads MyOne Streptavidin T1 and SureSelect Binding Buffer to the hybridisation reaction. With my captured DNA retained on the streptavidin beads, I then proceeded to purify and concentrate the hybridised library to a final 30µl volume.

3. Amplification of hybridised libraries

I used 14µl of the Biotinylated RNA library hybrids for the amplification reaction (Table 9). This meant that for every pulldown experiment, I was able to carry out four amplification reactions if the DNA concentration remained low. I proceeded to amplify the DNA using Herculase II Fusion DNA Polymerase under the following PCR conditions:

Table 9. Optimised PCR conditions for the PCR amplification steps

Step	Temperature	Time
Step 1	98°C	2 minutes
Step 2	98°C	30 seconds
Step 3	57°C	30 seconds
Step 4	72°C	3 minutes
Step 5		Repeat step 2 through to 4 for 16 cycles
Step 6	72°C	10 minutes
Step 7	10°C	Hold

4. Pacbio SMRTbell library preparation

The hybridised libraries were quantified via fluorescence using Qubit. This was followed by end repair of non-phosphorylated 5' ends PCR products and sample purification. I then ligated blunt hairpin adapters following manufacturer's instructions, included in the PacBio template kit for 2kb fragments. I carried out the SMRTbell quality assessment using Agilent 2100 Analyser (Agilent Technology, California, USA) and quantified the library using Qubit. Libraries were sequenced using two SMRT cells per sample.

Due to the number steps involved (>100) in manually constructing libraries to sequence with PacBio *RS*, the typical DNA yield is very low. For the pulldown hybridisation step, I had a yield of approximately 8-10% of my starting material (Table 10).

Table 10. DNA sample concentration for different library preparation steps

Samples	Total starting concentration (μg)	Pre-pulldown concentration amplification (μg)	Post-pulldown concentration (μg)	Final concentration after PacBio library (μg)
NA11994	3.0	3.8	0.18	0.14
NA12154	3.0	4.8	0.20	0.14
NA12155	3.0	4.2	0.12	0.08
HG00524	3.0	3.9	0.12	0.08
HG00478	3.0	4.6	0.22	0.17
HG00530	3.0	4.1	0.19	0.13
HG00533	3.0	4.0	0.21	0.17
HG00557	3.0	3.7	0.15	0.11
HG01108	3.0	3.8	0.09	0.06

3.4.5. Analysis of PacBio and Illumina sequencing data

I generated ROI with a minimum of 1 pass along the insert and 90% accuracy using the SMRT portal (Table 11). The average quality for the ROI >98% and the average length = 1,500bp (Figure 18). Due to the nature of the library preparation, ROI contained the insert of interest as well as PCR amplification primers and indexes. The PCR primer sequences are not publicly available. However, by definition, PCR primers and indexes should be situated at the start and end of each ROI. I therefore removed 70bp from each end of ROI to guarantee exclusion of these sequences from downstream data analysis (Figure 19).

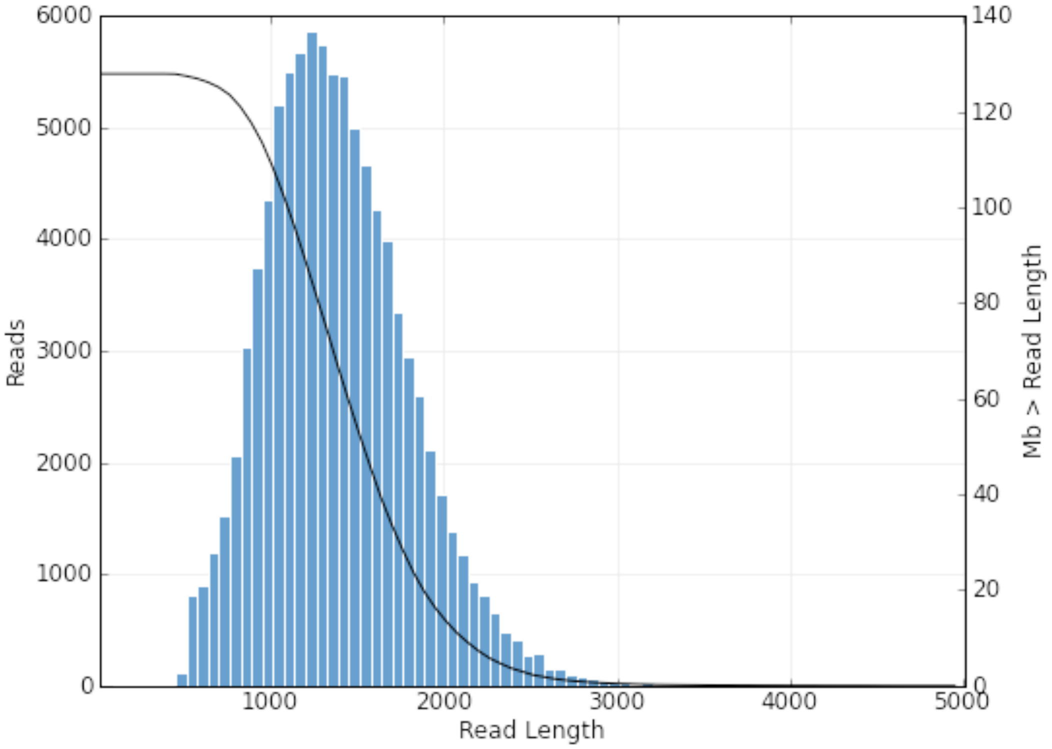
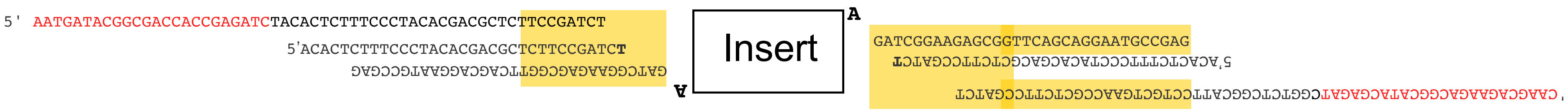


Figure 18. **Reads of Insert for PacBio**. Reads of Insert (ROI) were generated using the SMRT portal website. Reads are approximately 1,500bp in length.

Table 11. Statistic measures for PacBio Reads of Insert generated in the SMRT Portal.

Samples	Length of insert (bp)	Quality of insert (%)	Number of passes
NA11994	1,358	98.5	10.0
NA12154	1,270	98.8	12.0
NA12155	1,383	98.8	12.0
HG00524	1,486	97.7	8.0
HG00478	1,416	98.0	8.0
HG00530	1,500	98.7	11.0
HG00533	1,465	98.4	9.0
HG00557	1,465	98.4	9.0
HG01108	1,591	98.9	10.0


5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
5'ACACTCTTTCCCTACACGACGCTCTTCCGATCT
A
Insert
A
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

```
>m140525_090102_00127_c100646712550000001823121809101487_s1_p0/1579/ccs
AATGATACGGCGACCACCGAGATCTACACTCTTTCCTACACGACGCTCTTCCGATCTCGCACTACCATGC
CCAGCTAATTTTTTGTATTTTTAGTAGAGACAGGGTTTCACCATGTTGGCCAGGGATGGTCTCGATCTCT
TGACCTCGTGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATTACAGGTGAGCCACCGCGCCCGGCCT
GTCTTTGATCTTTCTGATAAGTGCGTAGAAGAAAACGCTGACGTATGCTGCCTTCCCTCTCTGTTTCGGC
TACCTATAAGGGAAGGGCCCCTGTCCTATGATCATGTGACTTGCTTGACCTTATCAATCATTTGGATGAC
TCACCCTCCTTACCCTGCCCCCTTGTTTTGTATACAATAAATATCAGTGCACCCAGCCATTCAGGGCCAC
TACCGGTCTCCGCGTCTTGGTGGTAGTAGTCCCCCCGGGCCCAGCTGTTTTTCTCTTTATCTCTTTGTCC
TGTGTCTTTATTTCGTACGATCTCTCTCTCGTCTCCACACACGGGGAGAACACCCGTCAAGCCCCGTAGG
GCGGGACCCTACATTTTCACGCCTTGTAATCCCAGCACTTTGGGAGGCCACGGCGGGTGGATCACCTGAG
GTTAGGGGTTCGAAACCAGCCTGACCAACATGGTGAAACCCTGTCTCTACTAAAAACACGAAAATTAGCT
GGGCATGGTGGCGGGTGCCTATAATCCCAGTTAGTCGGGAGGCTGAGGCAGGAGAATTGCTTGAACCTAG
GAGGCAGAGGTTGCAGTGAGCTGAGATCACGCCACTGCACTCCAGCCTGGGCAGCAGAGCAAGACTCCGT
GTCAAAAAAAAGAAAAAAAAAAAAGCTGGATTATCATGTAAGATCTTCTGATCTTTCAATGTGGGCAACTA
ATTTGAAAAATTGGCTTTTAAAAAATGTTAGAGTAATGTCTGCATATGGATTAAAAAGACCAACAGTAGT
AGGACTGAAAGGCTCATAGTGAAAAACAACCCAAAGTATGTGTTAATTAAATGAATGAGCTGTGTTCCGA
ATTTTAACTTTATTTTAAACTGTAAATTGATAAACTCCAACCTCTTAGAAGATAGTTTTTGTTTTTTTGT
CTCCGAAAGCAAATAAAGGTTTTTAACCTTGGGAATATTTTTACTGTTAAATAAAAATGTTAAAATTAGA
AAGTGTAGCTAGATGTAGTTTGCTAAAAATTATCTAGATGTAAATTCTTAAAATTAGAAAGTGTATCTAG
ATGTAAAAAGTCAGGAGAGACCATGAGTTGCTCCCGTTTGTTTAAAAAGCAGGCTTTATGTAGATGGCAT
TCTCGAACTTCCATTATTTAATTTTTTAAGGATTTTAAATCACATGAACATGAATTCCTAATAACAATGT
ACATGTATGGCCTATGGCCTATGGCTTTATCTTTAAAATCAGAACTAAAATAGGCAAATAATTTTTCTGT
ATGTTTGACATTTTTCAAAATAGATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTAT
GCCGTCTTCTGCTTG
```

Figure 19. **PacBio Read of Insert carrying Illumina library adapters.** This is a true representation of a Read of Insert (ROI) with 99% accuracy. Highlighted in red are the Illumina adapters commonly used during the library preparation to amplify the DNA libraries. Highlighted in yellow are the regions targeted by sequencing primers in Illumina. These sequencing primers enable the insert or region of interest to be sequenced with little Illumina adapters or primers' contamination. In this pilot study, a combination of Illumina library preparation and PacBio sequencing was applied. A SMRTbell with the insert as well as the Illumina primers and adapters was sequenced. As a result, 70bp of DNA sequenced had to be trimmed from both ends of the Reads of Inserts to ensure that there was no primer contamination in the final sequence.

I aligned the ROI for PacBio data and Illumina paired-ends reads to the Human Reference Genome hs37d5 with bwa-mem(165). In order to increase the quality of downstream variant calling, I followed the data processing guidelines set out by the GATK pipeline. First, I marked duplicates and added read group information using Picard Tools (http://picard.sourceforge.net). Using GATK version 3.6, I re-calibrated each base in the bam file using a process known as Base Quality Score Recalibration (BQSR). Variant calling algorithms rely on the quality score of each base in the sequencing reads. These quality scores are generated by the sequencing machines and are therefore subjected to technical errors. Base Quality Score Recalibration (BQSR) is a pre-processing step developed by the GATK team in which they adjust the quality scores per base using external information for known variants in dbSNP or 1000 Genomes. Following recalibration, I performed variant calling using HaplotypeCaller(166) with the following parameters: -ERC GVCF --allowNonUniqueKmersInRef to create single sample gVCFs. I then ran GenotypeGVCF to combine files into one VCF.

To QC the raw variants for both PacBio and Illumina, I followed the hardfiltering recommendations stated by GATK (Figure 20). Due to the low number of samples, I could not apply the Variant Quality Score Recalibrator (VQSR) which uses sophisticated machine learning processes for the QC steps.(167). The hardfiltering parameters included were:

QualByDepth (QD) metric that represents the quality of the variant normalised by the allele depth (AD). This way variants in high depth regions do not have artificially inflated QUAL scores. The generic recommendation suggests that variants with QD <2 should be filtered out.

Fisher Strand (FS) which represents the Phred scale probability that there is strand bias at that site. A value equal to 0 means that there 100% confidence there is no strand bias at that site.

StrandOddsRatio (SOR) is an updated form of the Fisher Strand test better suited to analyse large amounts of data in high coverage regions. The name in this case does not represent the annotation because SOR is not an odds ratio, it was changed by the GATK throughout the development of the test.

RMSMapping Quality (MQ) represents the mean mapping qualities plus the standard deviation of the mapping qualities

MappingQualityRankSumTest (MQRankSum) compares the mapping qualities of the reads supporting the reference alleles and the alternate alleles. A negative value means that the mapping qualities of the reads supporting the reference alleles are higher than the reads supporting the alternative alleles.

ReadPosRankSum test measures whether a variant call is positioned at the end of reads more often compared to the reference allele.

Lastly, I also measured the ratio of transition to transversion (Ts/Tv) to detect if there was any evidence of a poorly sequenced samples (Table 12).

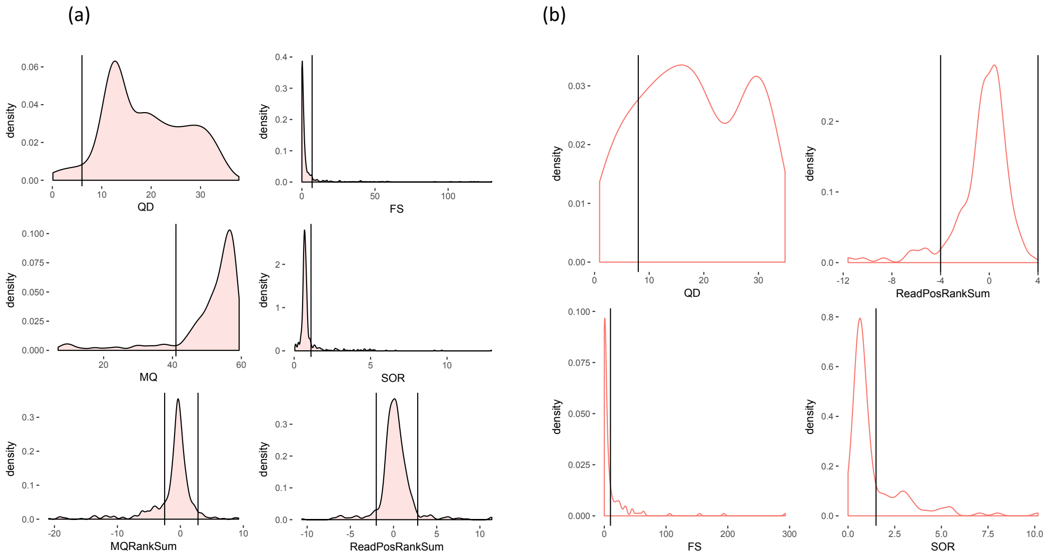


Figure 20. **Density plots of QC parameters used to filter out bad quality SNPs and Indels.** On the x-axis are the annotation values used to assess the quality of the dataset, and on the y-axis, are the density values. The black line represents the threshold values used to hardfilter variants according to GATK best practices **(a)** Distribution of several QC parameters on single nucleotide polymorphisms (SNPs). The distribution of Quality by Depth (QD) values range from 0-30, with two peaks representing heterozygous reads (QD =12) and homozygous calls with approximately double the number of reads (QD = 18). Variants with QD values < 6 represent low quality calls and were excluded from the dataset. SNPs have Fisher Strand (FS) values close to zero showing no strand bias. The black lines show the cut off value FS >10 used to exclude variants that deviate from zero. SNPs Mapping Quality (MQ) values range from 40-60. Any variants with MQ <40 were excluded from the analysis. The majority of SNPs had Strand Odds Ratio (SOR) values close to zero. To exclude variants with some degree of strand bias, SOR values > 3 were used as cut off values. Variants with MQRankSum < -3 (x3 more reads supporting reference alleles) and MQRankSum > +3 (3x more reads supporting alternate alleles) were also excluded in the analysis. Finally, the dataset showed that the distribution of the Read Position Rank Sum (ReadPosRankSum) is close to zero indicating that there is little or no difference between the SNP positions within the reads. **(b)** Distribution of several QC parameters on indels. Similar to SNPs QC steps, Quality by Depth (QD) values range from 0-30, with two peaks representing heterozygous reads (QD =12) and homozygous calls with approximately double the number of reads (QD = 30). Variants with QD values < 9 represent low quality indels and were excluded from the dataset. Indels that exhibited any positional bias (ReadPosRankSum) because they were always at the start or end of reads were excluded from the dataset. Likewise, indels that displayed some level of strand bias were also excluded (FS and SOR).

Table 12. Statistics showing the improvement of Ts/Tv ratios following variant filtering

Callset	Number of SNPs before QC	**Ts/Tv** before QC	Number of SNPs after QC	**Ts/Tv** after QC	Number of indels before QC	Number of indels after QC
Illumina	634.0	2.5	531.0	2.6	157.0	43.0
Illumina	571.0	2.6	521.0	2.7	204.0	65.0

3.4.6. Sanger sequencing validation of variants

I designed the primers and set the PCR conditions to sequence a 684bp region at the 5' end of the *IFITM3* gene from chr11:320,543-321,227. I used primers listed in Table 13 and Kapa HiFi (KK2600, Kapa Biosystems, Roche) polymerase for the PCR reaction to amplify the DNA using conditions listed in Table 14.

Table 13. PCR primers used to target a region at the 5' end of *IFITM3* gene

Name	Sequence 5'-3'	Usage	Manufacturer
IFITM3_F	CATTCCCTGGGCCATACG	target region chr11:320,543-321,227	Metabion, Germany
Uni-Alexa555	CATTCCCTGGGCCATACG-AGAGGTGAGGGCTTTGGGG	target region chr11:320,543-321,227	Metabion, Germany

Table 14. PCR conditions for the amplification of DNA prior sanger sequencing

Step	**Temperature**	**Time**
Step 1	95°C	3 minutes
Step 2	95°C	20 seconds
Step 3	60°C	15 seconds
Step 4	72°C	2 minutes
Step 5		Repeat step 2 through to 4 for 30 cycles
Step 6	72°C	2 minutes
Step 7	10°C	Hold

The amplified DNA was gel extracted using QIAquick Gel Extraction Kit (Qiagen) and Sanger sequenced on an Applied Biosystems 3730xl DNA Analyzer (GATC Biotech). Single-nucleotide polymorphisms were identified by assembly to the human reference (chr11:320,543-321,227) using Lasergene (DNAStar). Homozygotes were called based on high, single base peaks, whereas heterozygotes were identified based on low, overlapping peaks of two bases.

3.5. Results

3.5.1. Capture robustness for two sequencing platforms

To gain a more thorough understanding of the variation in the *IFITM* locus, I evaluated a target enrichment capture method in two separate sequencing platforms: Illumina MiSeq and PacBio *RS*. At the time I started this project, conventional pulldown protocols for targeted sequencing had not been designed for single molecule, real-time sequencing technologies such as PacBio. In order to adapt this method, I carried out a number of optimisation steps, some of which I presented in the Methods section of this chapter.

I first assessed the performance of each method by computing the fraction of the whole *IFITM* region (Chr11:280,000-380,000) that was covered by more than one read in both platforms. I found that the mean coverage for Illumina was 1,700x with approximately with 97% of all bases covered at 100x or more (Figure 21). This value was substantially higher than the overall coverage for the rest of the genome (0.98%). In contrast, the depth of coverage for the *IFITM* locus for the PacBio data was 40X (versus 0.03% coverage for the rest of the genome) and 97% of the locus had coverage values equal of greater than 2X (Figure 22).

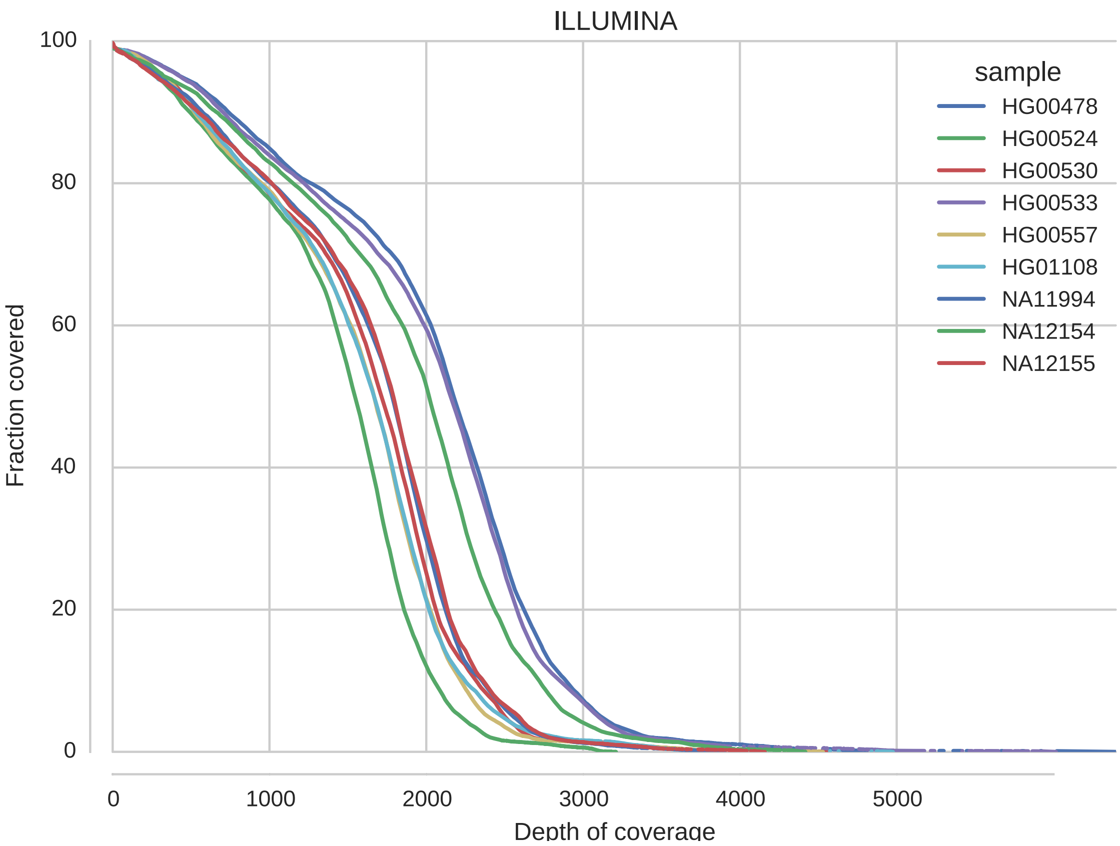


Figure 21. **Depth of coverage for the IFITM region by Illumina sequencing reads.** Approximately 97% of all bases are covered by at least 100 reads. Around 20% of the region contains coverage in excess of 2000X.

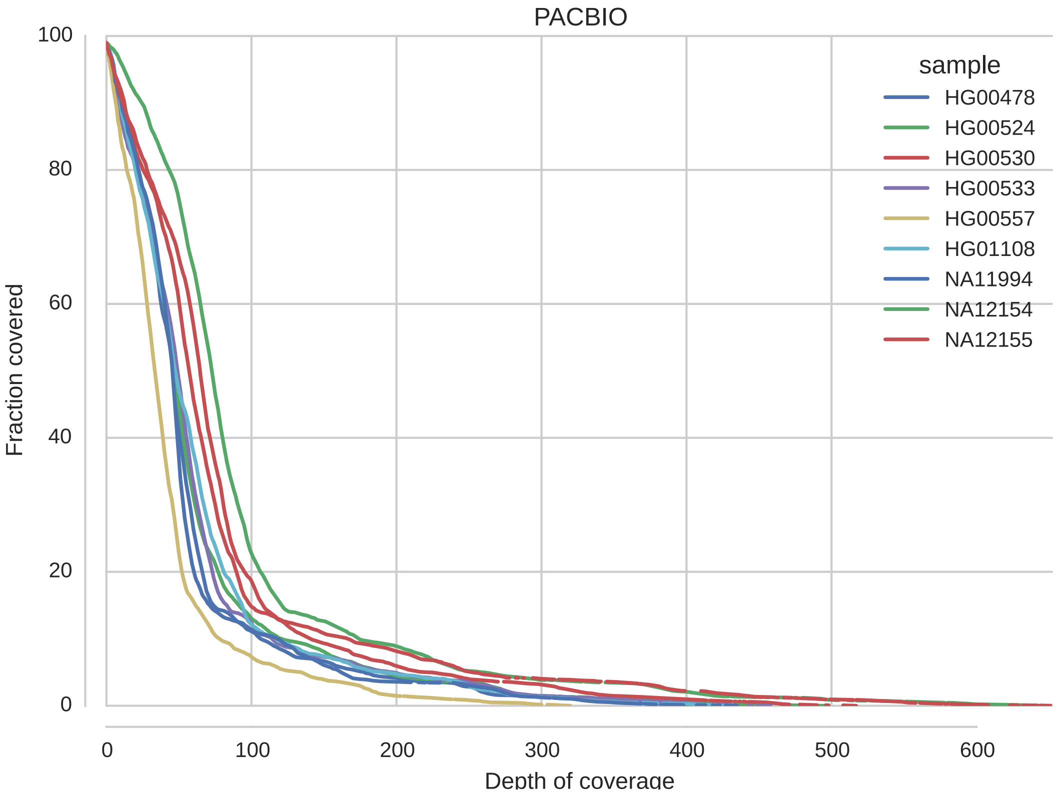


Figure 22. **Depth of coverage for the IFITM region by PacBio sequencing reads.** Approximately 97% of all bases are covered by at least 10 reads. Around 20% of the region contains coverage in excess of 50X.

In addition to the read coverage, I also assessed the efficacy of the pulldown in both Illumina and PacBio datasets. I obtained an average two million reads per sample in the Illumina dataset that mapped to the whole human genome. From these, a total of ~300,000 reads mapped to the targeted *IFITM* region, resulting in an efficiency of 13%. For long-read sequencing data, efficiency was substantially poorer, with an average of 3,000 reads of 1,500bp in length that mapped to the whole *IFITM* locus, for a final efficiency of 4%. The poor efficiency is the result of the inclusion of RNA probes within repetitive elements.

3.5.2. Assessing the influence of GC content on coverage

Several reports have highlighted that base composition can bias the sequencing efficiency(160). For example, regions with very high GC or high AT content vary in how well they are covered depending on the sequencing platform.(152, 168). There are two main technical reasons for this bias. It may be that the efficiency of amplification reactions required during library preparation is substantially reduced in these regions. Another reason is that the reduced efficiency is a direct consequence of the lower hybridisation efficiency of probe libraries to target regions in the genomic DNA with either too high or too low AT regions(140). To compare this sequencing bias for my region of interest, I explored the effect of GC content on coverage, as described by Clark and colleagues (169). I discovered that the average GC content for the whole locus is 60% and that none of the technologies displayed a fall in coverage. Indeed, it has been documented that GC content of this value does not have a substantial effect on sequencing reads(140, 160). I observed some regions with GC content >70% which are covered by PacBio only albeit with low sequencing depth (Figure 23).

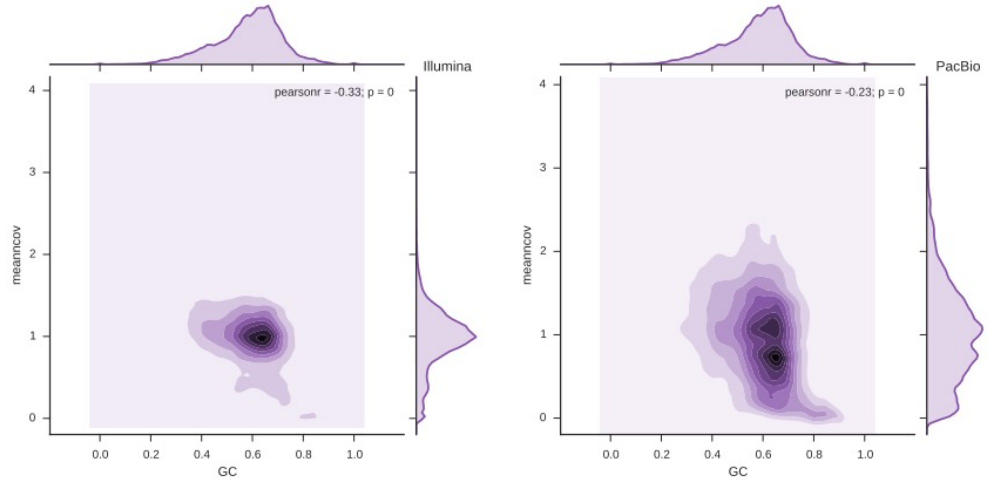


Figure 23. **Distribution of coverage as a function of GC content for the IFITM locus**. For each base pair in the target region, the mean coverage of its bases is calculated as well as the percentage of Gs and Cs it contains. The mean coverage for the entire region is plotted in the y-axis and it is represented with the value='1'. Any mean coverage values lower than 1 represents coverage values that are below the region's average. Values over 1, represents coverage above the average. The mean coverage is plotted against the GC content (x-axis), thus allowing to measure coverage bias as a result of nucleotide composition. (a) Distribution of coverage as a function of GC content for Illumina data. (b) Distribution of coverage as a function of GC content for PacBio data. (Graph made by Javier Diez from the Wellcome Trust Sanger Institute).

3.5.3. Variant analysis: comparison with Phase 3, 1000 Genomes Project dataset

The DNA samples used in this study are from individuals that took part in the 1000 Genomes Project, thus allowing for comparison between my sequencing data and the 1000 Genomes Project dataset. In the first comparison, I included all high-quality variants found in the *IFITM* locus. There were 531 SNPs and 46 indels in the Illumina pulldown callset, of which ~92% of SNPs and 65% of indels were found in the 1000 Genomes dataset. For the PacBio dataset, I found 521 SNPs and 65 indels of which 88% of SNPs and 45% of indels were found in 1000 Genome callset (Figure 24).

Only 15 SNPs were reported only by the Illumina pulldown and none were found to be novel variants. From the 35 variants reported in the PacBio dataset, only 6 were novel (Figure 24). For variants reported by both PacBio and Illumina not found in 1000 Genomes (28 SNPs in total), only 4 were novel (1 intronic and 3 intergenic). None of the novel variants reported by either Illumina or PacBio resulted in non-synonymous protein coding changes.

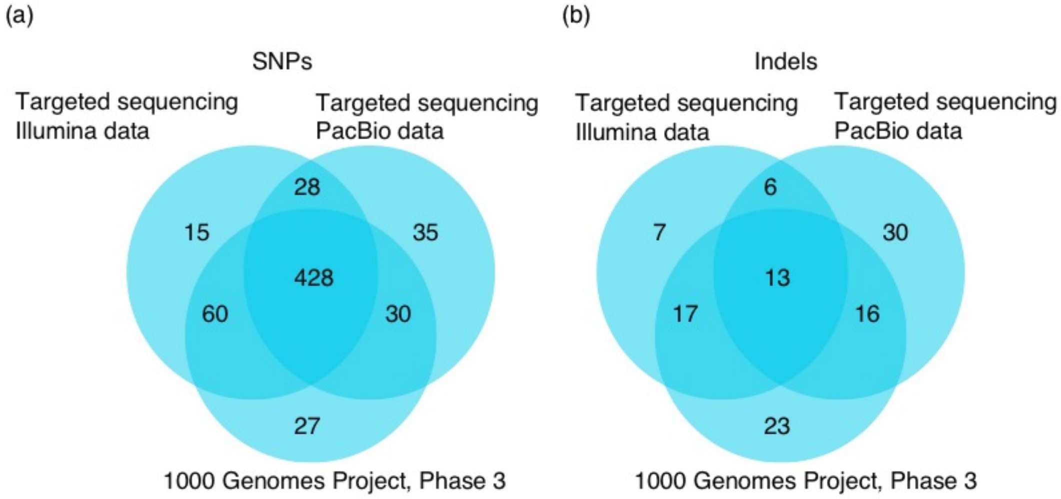


Figure 24. **Venn diagram showing site level evaluation for variants in the targeted pulldown study.** (a) Represents the comparison between all SNPs called in the targeted pulldown study (Illumina and PacBio) and the SNPs reported for the same samples in the 1000 Genomes dataset. (b) Represents the comparison between all Indels called in the targeted pulldown study (Illumina and PacBio) and the Indels reported in the 1000 Genomes dataset for the same samples.

In addition to site-level comparisons, I also carried out genotype concordance evaluations (Table 15) between variants in the pulldown dataset (Illumina and PacBio pulldown) and variants from the 1000 Genomes dataset (Gold standard dataset). In addition, SNPs used for this initial comparison were included regardless of whether they were located in repetitive and non-repetitive regions. I found moderate levels of sensitivity and precision between the pulldown dataset and 1000 Genomes dataset with values ranging from 86-92% for both metrics (Table 16). The low specificity (34-66%) between variants in the pulldown dataset and the 1000 Genome project suggests that the pulldown dataset may contain a high number of calls which may be false positives.

Table 15. Summary of variant filtration metrics used to evaluate the quality of *IFITM* locus pulldown data with 1000 Genome dataset

Sensitivity	$\frac{TP}{TP + FN}$	Proportion of variants found in the truth set (1KP) and in my callset.
Precision	$\frac{TP}{TP + FP}$	Fraction of calls that are true positives from all calls made.
Specificity	$\frac{TN}{TN + FP}$	Fraction of calls that represent the rate at which we make calls that are not true.

Table 16. Comparison table between PacBio and Illumina enrichment dataset and the 1000 genomes dataset for all single nucleotide variants in the entire IFITM locus, including repetitive regions.

	Samples	Precision (%)	Sensitivity (%)	Specificity (%)
PacBio	HG00478	88.0	91.0	34.9
	HG00524	91.0	89.0	48.6
	HG00530	90.0	88.0	41.0
	HG00533	89.0	85.0	43.5
	HG00557	88.0	91.0	35.9
	HG01108	85.0	84.0	46.3
	NA11994	90.0	88.0	45.2
	NA12154	90.0	88.0	52.1
	NA12155	87.0	83.0	51.7
	Average	88.7	87.4	44.4
Illumina	HG00478	93.0	92.0	48.4
	HG00524	93.0	92.0	53.8
	HG00530	92.0	87.0	50.0
	HG00533	90.0	84.0	53.3
	HG00557	94.0	87.0	66.7
	HG01108	91.0	86.0	53.7
	NA11994	91.0	90.0	38.2
	NA12154	93.0	90.0	52.6
	NA12155	92.0	91.0	51.3
	Average	92.1	88.8	52.0

For the next comparison analysis, I excluded variants from low-complexity regions as suggested by Li, 2014(170). I found that when this extra filtering was applied, the level of concordance increased dramatically (Figure 25). The average specificity also increased to 98% for the Illumina dataset and 96% for PacBio

(Table 17). Unsurprisingly, this suggests that erroneous alignments in low-complexity regions are a major source of erroneous calls(170).

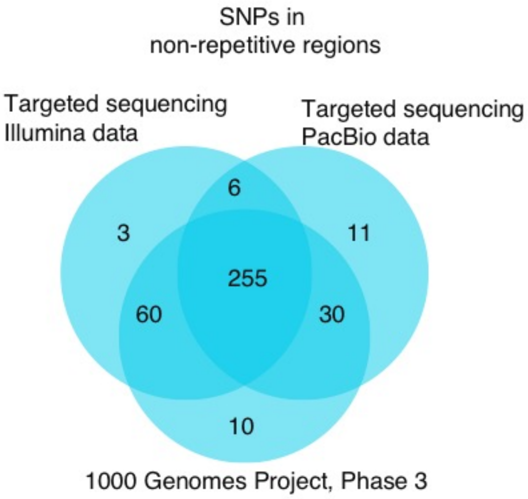


Figure 25. **Venn diagram to show site level evaluation of variants in non-repetitive regions.** Represents the comparison between all SNPs called in the targeted pulldown study (Illumina and PacBio) in non-repetitive regions and the SNPs reported for the same samples in the 1000 Genomes dataset.

Table 17. Comparison table between PacBio and Illumina enrichment dataset and the 1000 Genomes dataset for a subset of variants in non-repetitive target region.

	Samples	Precision (%)	Sensitivity (%)	Specificity (%)
PacBio	HG00478	96.0	99.0	95.0
	HG00524	98.2	97.0	98.5
	HG00530	98.1	96.0	97.5
	HG00533	98.6	95.0	97.0
	HG00557	97.2	99.0	96.0
	HG01108	98.0	97.0	98.0
	NA11994	98.0	96.0	97.0
	NA12154	99.0	98.0	96.3
	NA12155	98.0	98.6	97.0
	Average	97.9	97.3	96.9
Illumina	HG00478	98.0	99.0	98.0
	HG00524	98.5	98.0	98.0
	HG00530	97.4	97.9	98.4
	HG00533	97.8	98.7	97.9
	HG00557	99.0	95.9	98.0
	HG01108	99.0	98.0	98.7
	NA11994	99.3	98.0	97.0
	NA12154	99.0	98.3	97.9
	NA12155	98.0	99.0	97.8
	Average	98.4	98.1	98.0

3.5.4. Sanger validation of a subset of variants

For the next step of the project, I decided to validate eight SNPs within the region: Chr11:320,988-321,138 by Sanger sequencing. This region is located at the 5' end of the *IFITM3* gene and it is generally poorly represented by sequencing data (Figure 26). Figure 27 and Figure 28 show that all genotype calls from PacBio sequencing data were successfully validated by sanger sequencing. In contrast, some of these genotypes were incorrectly called or entirely missed by Illumina sequencing (25%). This suggests that long-read PacBio data offers greater precision compared to short-read sequencing, especially in regions that are repetitive or difficult to sequence. These results also support the exclusion of region embedded within low-complexity regions for comparison, as they often result in mapping artefacts and miscalls. In addition, the results also highlight the advantages of targeted sequencing over low-coverage whole genome sequencing data as it can offer greater resolution for regions of the genome that are not easily sequenced with short-read sequencing platforms.

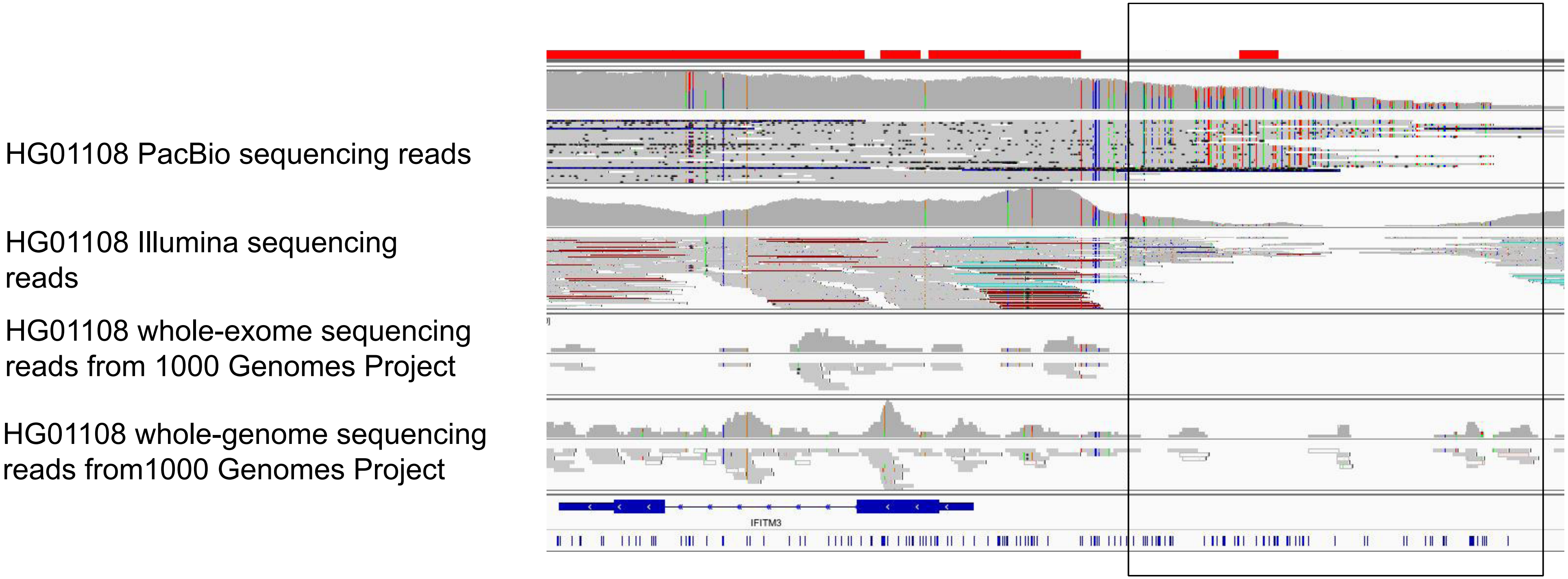
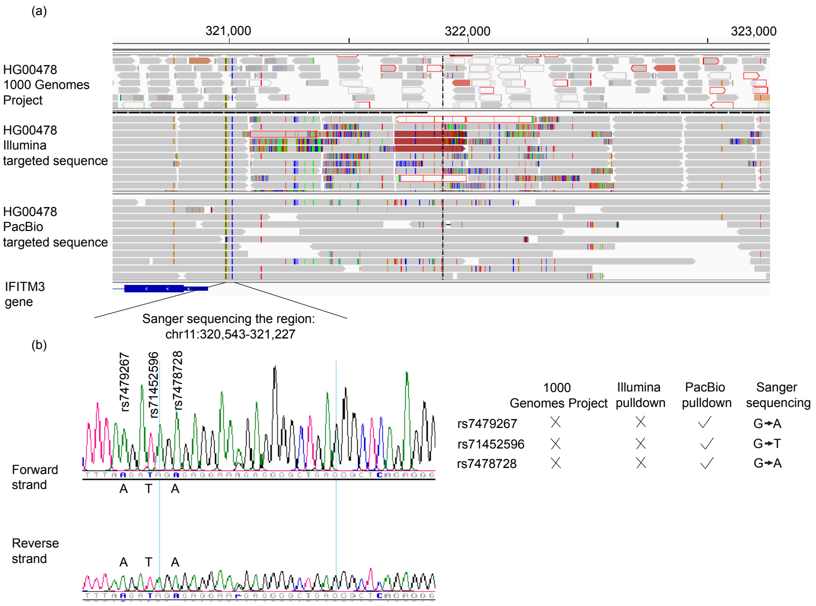

HG01108 PacBio sequencing reads
HG01108 Illumina sequencing reads
HG01108 whole-exome sequencing reads from 1000 Genomes Project
HG01108 whole-genome sequencing reads from1000 Genomes Project
IFITM3

Figure 26. **IGV view of Coverage for the *IFITM* region at chr11:320,988-321,138**. The top panel (in red) represents the probes used in the study. The square is highlighting a region that is poorly covered in most sequencing datasets. PacBio covers some percentage of the region. Paired-end Illumina reads also cover the region at a very low depth of coverage. No coverage is observed for whole exome 1000 Genomes data (as expected) and no sequencing reads are observed for whole genome sequencing from the same dataset. This figure shows one representative example (sample ID HG01108)



	1000 Genomes Project	Illumina pulldown	PacBio pulldown	Sanger sequencing
rs7479267	✗	✗	✓	G→A
rs71452596	✗	✗	✓	G→T
rs7478728	✗	✗	✓	G→A

Figure 27. **Integrative Genomic Viewer (IGV) Screenshot of a region at the 5' *IFITM3* genes showing the complexity of region in sample HG00478**. The blocks in grey and white represent reads that have either passed the mapping quality metric or have failed the mapping quality (MQ=0), respectively. Any mismatches of bases within the reads and the reference sequence are highlighted by colour changes, with green representing the nucleotide adenine "A"; orange, guanine "G"; red, thymine "T" and blue, cytosine "C". (a) Low coverage sequencing reads from the 1000 Genomes Project for HG00478 contains low number of reads, a number of which have also failed QC (white blocks). Targeted sequencing of the locus with Illumina reads displays a higher number of reads. However, some of these reads contain a number of mismatches with the reference genome, thus highlighting the disadvantages of using Illumina sequencing to target low complexity regions. PacBio sequencing appears to provide the best coverage for the region given the reduced number of mismatches found between the sequencing data and the reference genome. (b)Sanger sequencing of 680bp region (chr11:320543-321227 for sample HG00478) highlights the complexity of SNP calling in the region. The genotypes for three SNPs rs7479267, rs71452596, rs7478728 were incorrectly called as heterozygous in the 1000 Genomes Project and the Illumina targeted sequencing dataset. Sanger sequencing confirmed that these SNPs are in fact homozygous at those positions. These results support PacBio data results that correctly called these SNPs as homozygous.

PacBio *IFITM* Pulldown

rs ID	Functional Consequence	Position/Alleles	HG00478	HG00524	HG00530	HG00533	HG00557	HG01108	NA11994	NA12154	NA12155
rs7479267	Upstream of *IFITM3*	11 320988 G A	1/1	0/1	1/1	1/1	1/1	0/0	1/1	1/1	0/1
rs71452596	Upstream of *IFITM3*	11 320991 G T	1/1	0/1	1/1	1/1	1/1	0/0	1/1	1/1	0/1
rs7478728	Upstream of *IFITM3*	11 320994 G A	1/1	0/1	1/1	1/1	1/1	0/0	1/1	1/1	0/1
rs6598045	Upstream of *IFITM3*	11 321001 A G	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/1	0/0
rs3888188	Upstream of *IFITM3*	11 321017 A C	1/1	0/1	1/1	1/1	1/1	0/0	0/0	0/0	0/0
rs28602580	Upstream of *IFITM3*	11 321044 G A	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/1	0/1
rs35409983	Upstream of *IFITM3*	11 321055 G T	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/1	0/1
rs35218683	Upstream of *IFITM3*	11 321138 C T	0/1	0/0	0/0	0/0	0/0	0/0	1/1	0/1	0/1

Illumina *IFITM* Pulldown

rs7479267	Upstream of *IFITM3*	11 320988 G A	0/1	0/1	0/1	0/1	0/1	0/0	1/1	1/1	0/1
rs71452596	Upstream of *IFITM3*	11 320991 G T	0/1	0/1	0/1	0/1	0/1	0/0	1/1	1/1	0/1
rs7478728	Upstream of *IFITM3*	11 320994 G A	0/1	0/1	0/1	0/1	0/1	0/0	1/1	1/1	0/1
rs6598045	Upstream of *IFITM3*	11 321001 A G	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/1	0/0
rs3888188	Upstream of *IFITM3*	11 321017 A C	1/1	0/1	1/1	1/1	1/1	0/1	0/0	0/1	0/1
rs28602580	Upstream of *IFITM3*	11 321044 G A	0/0	0/0	0/0	0/0	0/0	0/0	0/1	0/1	0/1
rs35409983	Upstream of *IFITM3*	11 321055 G T	0/0	0/0	0/0	0/0	0/0	0/0	0/1	0/1	0/1
rs35218683	Upstream of *IFITM3*	11 321138 C T	0/1	0/0	0/0	0/1	0/0	0/0	1/1	0/1	0/1

Phase3, 1000 Genome Project

rs7479267	Upstream of *IFITM3*	11 320988 G A	0/1	0/0	0/1	0/1	0/1	0/0	0/1	0/1	0/1
rs71452596	Upstream of *IFITM3*	11 320991 G T	0/1	0/0	0/1	0/1	0/1	0/0	0/1	0/1	0/1
rs7478728	Upstream of *IFITM3*	11 320994 G A	0/1	0/0	0/1	0/1	0/1	0/0	0/1	0/1	0/1
rs6598045	Upstream of *IFITM3*	11 321001 A G	0/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
rs3888188	Upstream of *IFITM3*	11 321017 A C									
rs28602580	Upstream of *IFITM3*	11 321044 G A									
rs35409983	Upstream of *IFITM3*	11 321055 G T									
rs35218683	Upstream of *IFITM3*	11 321138 C T									

Confirmed by sanger sequencing	🟧
Unconfirmed by sanger sequencing	🟦
Not found in released dataset	⬜

Figure 28. **Panel to represent the genotype calls for nine samples sequenced in the study.** All the genotypes for the eight SNPs shown under rsID were validated by Sanger sequencing. Sanger sequencing results were therefore used to mark any given genotypes as "correct" in orange or "incorrect" in blue. Samples names are given as HG00478, HG00524, HG00530, HG00533, HG00557, HG01108, NA11994, NA12154, NA12155 whilst the rsID represents the positions that were validated by Sanger sequencing. Reference calls are represented as "0/0"; heterozygous calls are represented as "0/1" and homozygous calls as "1/1". The panel is further divided into three sections: PacBio *IFITM* Pulldown shows the genotypes of the eight SNPs called using the PacBio dataset; Illumina *IFITM* Pulldown shows the genotypes of these same eight SNPs called in the Illumina dataset and Phase 3, 1000 Genome Project shows the genotypes of the same eight SNPs in the 1000 Genome Project. The PacBio section shows all the genotypes called using this dataset are validated as correct by sanger sequencing (all in orange). Genotypes for six SNPs were incorrectly called across samples as heterozygous in the Illumina dataset (positions in blue). Four SNPs were either incorrectly called in the 1000 Genome Projects dataset (in blue) or not called at all (shown in grey).

3.6. Discussion

Target enrichment by hybridisation has shown rapid progress over the past few years. The methodology has become more widely accessible to the scientific community and constitutes a popular choice for many scientists because of cost and its potential for scalability.

In this sequencing study, I present a comparative study of a SureSelect target enrichment method in two different sequencing platforms (Illumina and PacBio). I looked at four main parameters for each technology: the number of bases covered in my region, the target efficiency, the impact GC bias on read coverage and the sensitivity of both technologies to capture variants in the *IFITM* region.

Although I was able to capture over 90% of the *IFITM* locus with both Illumina and PacBio technologies, the low efficiency of the pulldown resulted in a large portion of the data having to be discarded. I found that the amplification of the off-target hits was most severe for PacBio than Illumina, possibly due to low PCR amplification efficiency of longer reads (171). Many studies have observed that DNA repeat templates, if present in high copy numbers, can result in self-priming events during the genome amplification steps. As a result, there is an exponential increase in the number of repeat sequences in the final PCR amplicons(168). Target efficiency is heavily design-dependent (143, 150), therefore having probe sequences spanning repeats would certainly have a negative effect on the efficiency of the study.

Another reason for low efficiency could be due to the poor performance of the Agilent SureSelect capture method. However, efficiency values for pulldown studies that used the same targeted and sequencing methods rule out this hypothesis. For example, I applied the capture method described in this Chapter to enrich for Epstein-Barr virus (EBV) from a BCL37 cell line. I found that the capture efficiency in this experiment was more in line with what it is generally observed in pulldown studies (85% capture efficiency, Dr Anne Palser, personal communication)(143, 149, 150). I also carried out a pilot pulldown experiment in chicken lines where I observed a similar capture efficiency value (81%, Dr Irene Bassano, personal communication, Appendix A). This demonstrates that the poor efficiency is not the result of technical difficulties during the enrichment studies but the result of the inclusion of repetitive elements in our design. As a consequence, I excluded fifty probes sequences from the original design to use in future targeted sequencing studies.

One key aspect of targeted sequencing is variant discovery. Based on variant comparison to the 1000 Genomes dataset, I found that the inclusion of variants located in repetitive regions had a negative effect on the sensitivity of detection (170). In contrast, when I included variants from non-repetitive regions, I reached approximately 98% sensitivity and specificity for both technologies. Repetitive sequences have always presented computational challenges(170) as they create ambiguities in alignment and in the assembly process which hinder the interpretation of results(153).

Finally, I also validated eight variants in a 684bp region from Chr11:320,543-321,227 by Sanger sequencing in all nine samples. I showed that PacBio sequencing provides the most accurate results in terms of genotype accuracy (100% accuracy). In contrast, only 76% of Illumina calls for the pulldown study were accurate. This finding highlights the advantages of using long read sequences for ‘difficult’ to sequence regions. The application of PacBio sequencing to target high-repeat regions is well documented. For example, Loomis and colleagues successfully sequenced the human fragile X mental retardation 1 (*FMR1*) gene which contains a (CGG)n repeat (162, 168). They generated PacBio reads for the expanded CGG repeats contained within the gene and observed that PacBio sequencing was not adversely affected by the length of the repeats, only by the lifetime of the polymerase(168).

Despite the clear advantages of PacBio sequencing for hard-to-sequence regions, the use of this technology in combination with SureSelect targeted sequencing is limited for practical reason. For example, there are over one hundred steps required for the target enrichment and PacBio library preparations; and most of these steps cannot be automated due to low input concentration from the pulldown method. This lack of scalability is a major drawback for the utility of the target enrichment for PacBio sequencing at the moment.

In contrast, Illumina is amenable to robotic manipulation and does not require manual library preparations. Furthermore, Illumina methods are commonly available for indexing DNA libraries prior sequencing. Indeed, the ability to pool samples together reduces the cost of sequencing and increases the number of samples that can be sequenced with one probe library design at a time. Although PacBio has also introduced indexing for their libraries, attempts to apply these to the targeted pulldown were unsuccessful (Dr. Irene Bassano, personal communication).

These results clearly show the value of both technologies at capturing genomic variation. I found that the advantages of using PacBio for certain regions of the genome do not justify the higher human cost of performing target sequencing of the *IFITM* region using this technology. Furthermore, these results highlight the value of sequencing a small number of target regions at high sequencing depth where higher depth provides better resolution and can help uncover variants that had been previously inaccessible by low-coverage whole genome sequencing.

4. Assessing the contribution of IFITM variation to HIV-1 disease progression

4.1. Introduction

4.1.1. Global burden of HIV

It was not until 1982, following five years of reports on unusual series of infection cases such as *Pneumocystis* pneumonia (PCP) and Kaposi's sarcoma in young gay men, that the Centres for Disease Control and Prevention (CDC) finally acknowledged the USA was experiencing an epidemic. They called it 'acquired immune deficiency syndrome' (AIDS). Since then, and despite global efforts of more than 160 countries to combat the transmission of HIV, facilitate treatment and support current HIV research, HIV infections and AIDS continue to be one of the greatest public health challenges of the 21st century, especially for low-income countries. There are on average 36.7 million HIV infected individuals worldwide, of which 2.1 million were reported as new infections in 2016 (Global Aids Update, 2016). The risk of infection continues to be high among population groups of sex workers and their clients, men that have sex with men (MSM)(172), transgender groups and injection drug users(173).

4.1.2. HIV infection

HIV-1 and HIV-2 are distinct retroviruses of different origins. The primate reservoir of HIV-1 is the chimpanzee *Pan troglodytes troglodytes* (Ptt) populations from the southern region of Cameroon(174) whilst HIV-2 originates from the sooty mangabey monkeys (*Cercocebm atys)*(175) from West African regions. The incidence of HIV-2 worldwide is low with approximately 144 patients reported to be infected with HIV-2 in the UK alone in 2013 (http://www.aidsmap.com). The incidence of HIV-1, on the other hand, remains high, with most HIV-positive individuals infected with this virus type (Global Aids Update, 2016).

In-depth analysis of historic blood samples from the Democratic Republic of Congo and Republic of Congo, places the origin of HIV-1 group M (responsible for the worldwide pandemic) in Kinshasa (Figure 29) and supports the hypothesis that changes in population movement and sexual behaviours contributed to the establishment and dissemination of the pandemic HIV-1 group M strain (176)

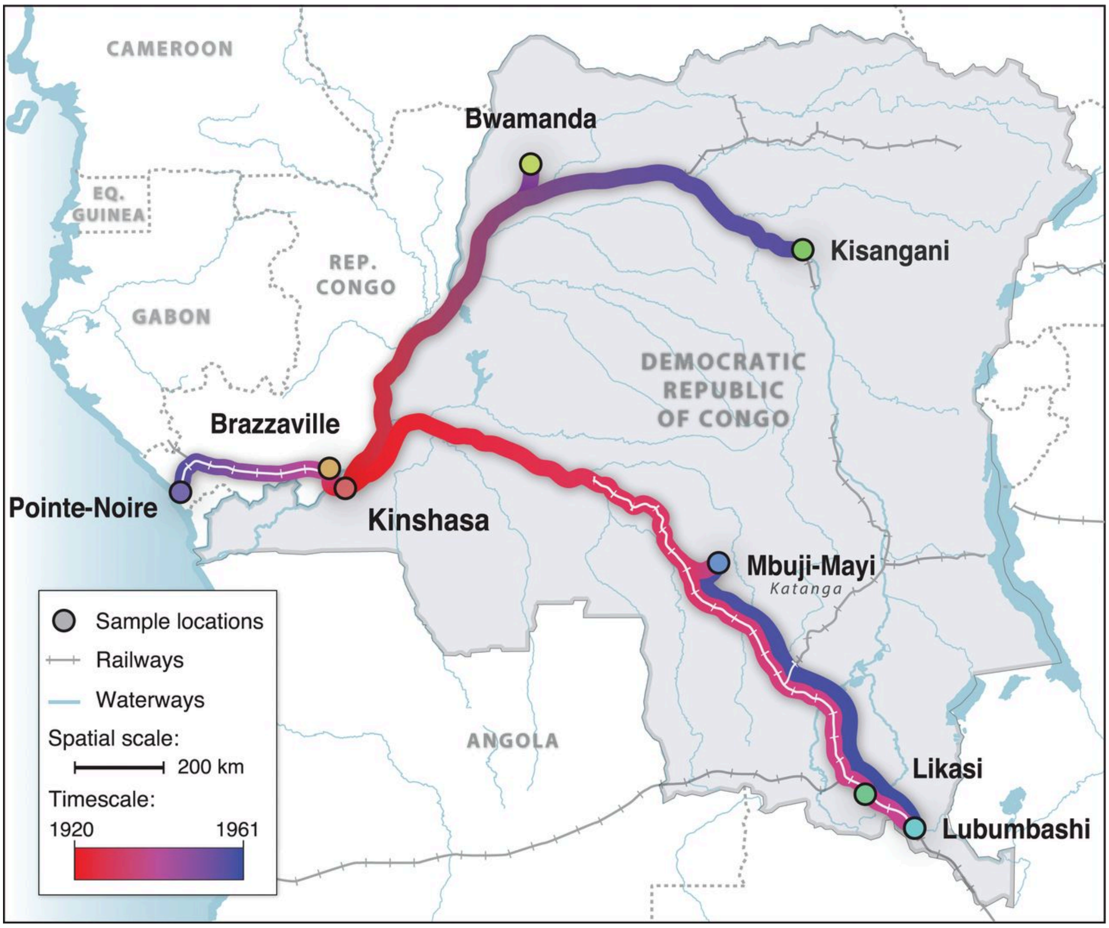


Figure 29. **Representation of the spatial movement of HIV-1 group M strain in Kinshasa**. The circles represent the locations where samples were available. They are coloured according to the time HIV-1 M viruses were introduced. The rate of special movement for the virus were projected onto railway and waterways transportation systems and coloured according to time scale of virus movement. Figure from Faria *et al.*, 2014.

HIV-1 viruses are also further characterised by the types of cells they infect and can be divided into two main groups: macrophage-tropic nonsyncytium inducing (NSI) or T-tropic syncytium-inducing (SI) isolates(96, 177, 178). NSI isolates preferentially target macrophages and constitute the most common type of transmitted viruses. SI T-tropic viruses, on the other hand, target T-cells and generally appear during later stages of infection. To infect cells, both macrophage-tropic and T-tropic must bind CD4 receptors in susceptible cells although their fusion is mediated by different accessory proteins or chemokine co-factors (CCR5 or CXCR4). For example, by expressing CCR5 receptors in human and murine cell lines and testing infectivity of these cells by HIV-1 pseudotyped viruses, Deng *et al*. 1996 established that macrophage-tropic viruses require β-chemokine receptor CCR5 for viral entry into the cell(179, 180). Pseudotyped viruses are replication defective and provide a safer alternative to live viruses.

T-tropic viruses require α-chemokine receptor CXCR4 to infect cells(178). CXCR4 co-factor was first discovered back in 1996, when Feng, *et al*. engineered murine NIH 3T3 fibroblast cells (that readily transfect) to express CXCR4 co-factor, CD4 receptor or both. They found that HIV entry and infection occurred only in cells co-expressing CD4 and CXCR4, judged by the presence of syncytia and levels of beta-galactosidase β-Gal levels in these cells(178). Once the Env protein binds the CD4$^+$ receptor and co-receptors, it triggers the structural changes in the gp41 virus envelop protein and enables the fusion of the virion to the host membrane(181) (Figure 30). Once the virion penetrates the membrane, it starts the process of uncoating, followed by reverse transcription of its RNA into cDNA. This cDNA is then transported into the nucleus, where the viral integrase enzyme catalyses the integration of viral cDNA in the host cell DNA leading to provirus formation(181).

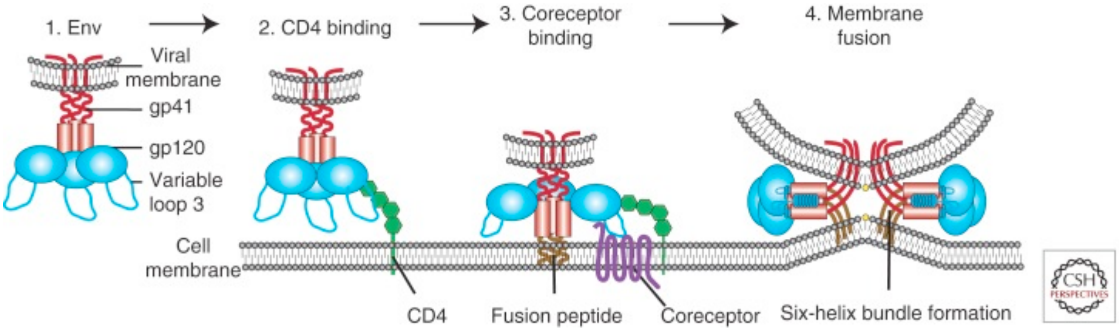


Figure 30. **Overview of HIV entry**. To deliver the viral proteins into cells, HIV Env, comprised of gp120 and gp41 subunits (1), first attaches to the host cell, binding CD4 (2). This causes conformational changes in Env, allowing coreceptor binding, which is mediated in part by the V3 loop of Env (3). This initiates the membrane fusion process as the fusion peptide of gp41 inserts into the target membrane, followed by six-helix bundle formation and complete membrane fusion (4). Adapted from Wilen *et al*., 2012.

4.1.3. Pathogenesis of infections

Following HIV entry and infection, a typical pattern of HIV infection *in vivo* develops and can be classified into three broad phases: the acute or primary infection, the chronic (clinical latency) and the symptomatic phase (AIDS). The acute infection occurs in the first two to four weeks and it is associated with rapid CD4 T-cell depletion and high viral load, sometimes in excess of 1 million copies of virus RNA per ml(182). The period 2-10 weeks post infection marks the start of the chronic phase when the viral load drops to stable levels (viral set point) possibly due to initial intracellular immunological responses to infection. This viral set point following acute infection serves as a good predictor of disease progression(183). For example, early studies evaluating the viral load in 180 individuals found that 50% of individuals with high viral load (>10,900 HIV-1 RNA molecules/ml) died within 6 years of the study entry despite having CD4+ T-cell counts >500 cells/μl (normal range of T-cell count). In contrast, only 5% of subjects with similar CD4+ T-cell count but <10,900 HIV-1 RNA molecules/ml died in a similar time period(183). This phase is characterised by the expansion of cytotoxic T-cells (CD8+ T-cell) that specifically target the viral particles leading to a fall in viral load (182, 184-186). Neutralising antibody activity targeted to the *env* HIV gene also develops 2-5 weeks after infection but the ENV viral protein evades complete neutralisation(187). In typical progressors, the chronic stage for the disease can last for around ten years until a rapid depletion of CD4+ T-cells emerge, leading to AIDS. For several years it was believed that this depletion of CD4+ T-cells occurred due to a process of exhaustion but the overall consensus in the field points to a mechanism of immune activation(188). This mechanism involves a subset of CD4+ and CD8+ T cells that rapidly proliferate and die and are able to recruit other T-cells to this dying pool. This eventually leads to a depletion of naïve T-cells and CD4+ and CD8+ T-cell numbers(188). Concomitant with this cell depletion, there is also a resurgence of high viral load and uncontrolled replication due to a gradual failure of CD8+ T cells to control the virus. Some groups have hypothesised that the gradual failure of CD8+ T-cells to control replication can be explained by a 'viral escape' mechanism(189). This 'viral escape' theory supports the idea that cells stop recognising HIV's genetic sequences due to high levels of viral turnover as a result of string of selection pressures from the host(190). In one study, it was observed that CD8+ T-cells from individuals in the symptomatic stages of the disease despite being able to recognise and kill laboratory HIV strains, were unable to target their own infected cells(191, 192). There is however, an 'evolutionary penalty' incurred by the virus as a result of these escape mutations. In some instances these escape mutant viruses will revert in the absence of the host pressures that initially selected them, thus remaining useful for cytotoxic T-lymphocyte (CTL) vaccine designs(190).

4.1.3.1. Extreme HIV phenotypes

Patients infected with HIV can be classified into three broad groups depending on the speed of their progression to AIDS (Figure 31). Patients are classified as 'long term non progressors (LTNP) if they maintain a stable CD4+ count, typically over 500 cells/μl and remain AIDS-free for at least ten years(193). The 'elite controllers' also satisfy some of the definitions of LTNP but constitute an independent group. These patients naturally supress the virus to undetectable levels (<50 copies/ml) without antiretroviral therapy (ART)(193) and rarely progress to developing AIDS. Chronic or typical progressors suffer a gradual decline in their CD4+ T-cell levels and develop AIDS only after a period of 8-10 years. On the opposite end of the spectrum, there are the rapid progressors, who manifest a rapid decline in the levels of CD4+ cell counts, typically <200 cells/μl and progress to AIDS within 2-5 years due to uncontrolled replication (> 10,000 HIV RNA molecules/ml) in the absence of ART. Typically, the optimal viral suppression measurements are defined as the viral load below the level of detection (HIV RNA <20 to 75 copies/mL).

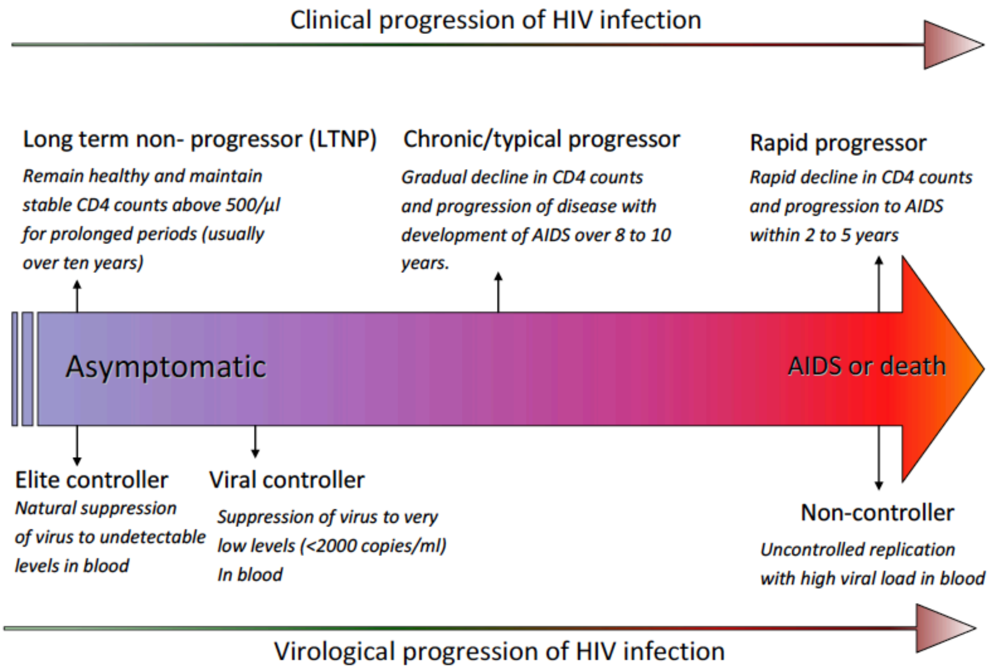

Clinical progression of HIV infection
Long term non- progressor (LTNP)
Remain healthy and maintain stable CD4 counts above 500/μl for prolonged periods (usually over ten years)
Chronic/typical progressor
Gradual decline in CD4 counts and progression of disease with development of AIDS over 8 to 10 years.
Rapid progressor
Rapid decline in CD4 counts and progression to AIDS within 2 to 5 years
Asymptomatic
AIDS or death
Elite controller
Natural suppression of virus to undetectable levels in blood
Viral controller
Suppression of virus to very low levels (<2000 copies/ml) In blood
Non-controller
Uncontrolled replication with high viral load in blood
Virological progression of HIV infection

Figure 31. **Classification of HIV disease based on clinical and virological progression.** Figure from Gurdasani *et al.* 2014.

4.1.3.2. Host genetic determinants of HIV extreme phenotypes: Long term non-progressors and elite controllers

The natural control of viral replication in the absence of treatment is rare and only observed in approximately 1% or less of the HIV-infected individuals(194). The mechanisms by which these patients are able to control HIV replication are still being elucidated but are likely to be influenced by a combination of factors, including the virus strain, background genetic factors and individual immune responses(195).

4.1.3.2.1. Viral genetics

Variations and mutation in HIV genes have been suggested to play a role. For example, Rhodes *et al*., 1995 reported a case study where the individual carried a defected virus with large deletion in the *nef-nef* and *nef-U3* region(196). Similarly, Wang *et al*., found that virus replication was hampered in one individual possibly due to stop codons in HIV matrix protein GAG p17 and capsid protein GAG p24 and in polymerase reverse transcriptase(197). However, studies of virus genomes isolated from 10 elite controllers did not find evidence of large deletions(198). Furthermore, analysis of full-length plasma virus and provirus sequences from around 95 elite controllers, found no evidence of gross viral defects(199).

4.1.3.2.2. Host genetics

Several host factors have been reported that appear to explain, at least in part, the observed virus control in LTNP and elite controllers. Some studies have shown that protective MHC class I alleles (HLB*B57 and *B27) are enriched in the elite controller population(200, 201). More specifically, HLA-B*57:01 allele is observed in greater frequency in Europeans and North American elite control cohorts whilst HLA-B*57:03 is enriched in elite control populations of African descent(202).

Previous studies have proposed that CD4+ T cells' resistance to infection explained the control of viral populations observed in elite controllers and LTNP. However, CD4+ T-cells from elite controllers support infection of both T-tropic and M-tropic viruses *in vitro* to comparable levels to CD4+ T-cells from HIV-free individuals*(203)*, suggesting that resistant CD4+ T-cells in elite controllers may not play an important role in the natural control of these individuals.

Increased levels of broadly neutralising antibodies have also been suggested as a possible reason for disease control in controllers but reports of lower neutralising antibody (NAb) activity amongst the elite controller groups discard this hypothesis and suggest that Nab may not play a major role in the natural control of HIV infection(187). Other factors such as CD8+ mediated control due to secretion of IL-2 and IFN-g by CD8+ T-cells have also been reported in the elite control groups(200). Similarly, a 32-base pair deletion in the CCR5 receptor (CCR5Δ32) has been identified to confer protection to HIV-1 viruses that require CCR5 receptor for entry into the cell (HIV1- R5 tropic strains)(95, 204). Taken together, all the evidence suggests that other factors are also at play.

4.1.3.3. Rapid disease progression

Some viruses require CCR5 receptors for cell entry (R5 tropic strains) whilst other HIV-1 strains required a different receptor CXCR4 to infect the cells (X4 tropic strains)(205). It is believed that the presence of X4-tropic viruses increases the risk of HIV progression and serves as a predictor of poor immunological response and death. A UK study of 289 HIV-1 positive individuals during 12 months' prior antiretroviral treatment (ART) found that patients infected with X4 T-tropic virus or X4 dual/mixed viruses (n=60) had significantly greater decrease in CD4+ T-cell counts and were more likely to experience clinical adverse events, compared to the R5-tropic infected individuals(206). Follow up studies of both groups also found that once patients had started therapy, the CD4+ cell count and the viral suppression was comparable between both groups(206). In a similar study, a US team determined that the presence of dual/mixed R5/X4 HIV viruses also increased disease progression. They found that R5/X4 HIV infected patients (n=30) were twice as likely to have reduced CD4+ cell counts (<350 cell/mm^3), be initiated into antiretroviral treatment or die compared to R5-tropic infected individuals (n=270)(207). Although follow-up information was missing, it suggests, in line to what was previously reported, that X4-tropic viruses may speed-up the rate of HIV progression to AIDS.

4.1.4. IFITM genes in the context of HIV-1

Upon initial stages of HIV-1 and SIV infection, systemic type 1 interferon production is one of the first lines of defence elicited by the host immune system(208, 209). In rhesus macaques, induction of Type I IFN-α reduced the number of transmitter founder viruses and led to an increase in the number of challenges required to achieve SIV infection in these animals(210). Although virus evolution to counteract restriction by interferon stimulated genes (ISGs) such as APOBEC3G and Tetherin is well documented, reports of other ISGs that appear to restrict HIV-1 replication suggests that their role on the pathogenesis of HIV/AIDS remains to be elucidated(211). For example, myxovirus resistance 2 (MX2), an IFN-inducible GTPase protein is a strong HIV-1 inhibitor in interferon-treated cells(212). MX2 localises to the nuclear membrane and nuclear pores, possibly interfering with the accumulation of viral cDNA in the nucleus(212, 213).

The role of IFITMs for virus transmission and replication has remained elusive. Previous research suggests that the incorporation of IFITMs into HIV-1 viral membranes is associated with restriction of virus fusion and spread(214). When IFITM proteins are expressed in non-infected lymphocytes, the proteins exert some protective effects upon cell-free virus infection, but the restriction disappears during cell-to-cell HIV infection. In contrast, when IFITM is expressed in virus-producing lymphocytes, IFITM proteins are observed to co-localise with viral HIV proteins in nascent virions and restrict virus spread(214). In addition, Foster and colleagues demonstrated Transmitted Founder viruses (TF), viruses that establish *de novo* mutations, are resistant to IFITM inhibition. In contrast, HIV-1 clones generated from individuals after six months of infection display an increase in sensitivity to inhibition by IFITM2 and IFITM3 proteins. They also demonstrate that this HIV-1 restriction is dependent on the strain's co-receptor usage and the localisation of the IFITM proteins within the cells(211). Specifically, they found that X4-tropic viruses are more sensitive to IFITM2 and IFITM3 than R5-tropic viruses. When specific mutation were introduced at the N-terminal region of either IFITM2 or IFITM3, R5-tropic viruses rather than X4-tropic viruses displayed greater sensitivity(211) (Figure 32).

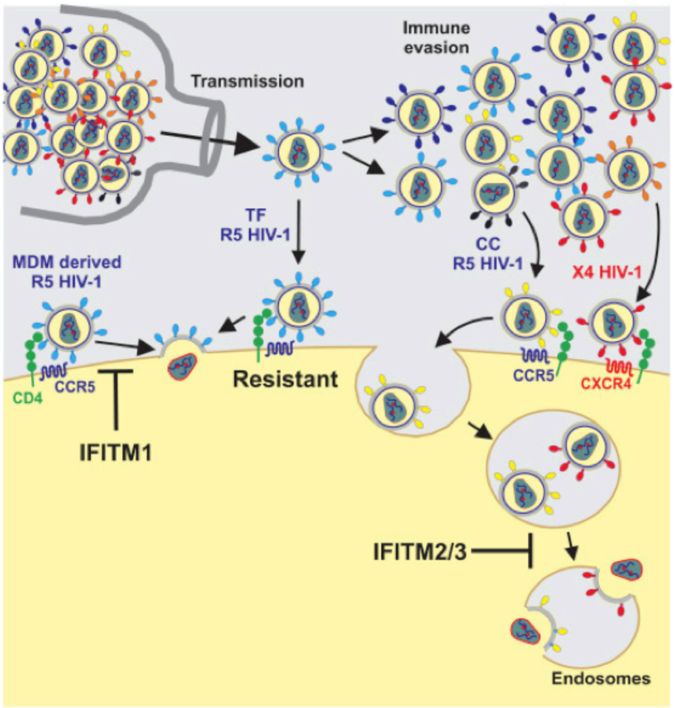


Figure 32. **IFITM inhibition of HIV**. Transmitter founder viruses (TF) are resistant to IFITM restriction. Escape mutations that allow the virus to escape detection from the host, and co-receptor tropism for CCR5 or CXCR4 make the virus sensitive to restriction by IFITM proteins in the endosomal compartments. Figure from Sauter, *et al.*, 2016

Recently, a polymorphism in *IFITM3*, rs12252, was associated with faster HIV-1 progression in China(89). Zhang and colleagues analysed a cohort of 74 patients classified as rapid progressors and 104 non-progressors from PRIMO cohort(215). By sanger sequencing 300bp at the 5' end of the *IFITM3* genes encompassing the rs12252 SNP, they reported a higher frequency that 68 out of 74 individuals carried the CC/CT alleles in rapid progressors compared to 78 out of a 104 non-progressors (P=0.004, OR 3.8(95% CI – 1.4-9.7). The higher CC/CT allele frequency in the PRIMO cohort (75% CT/CC genotype carriers) compared to the frequency of the same alleles in the European CASCADE cohort (2–8% CT/CC genotypes carriers) probably facilitated the study of this particular variant in this population(215).

The role of IFITMs for virus transmission and replication has remained elusive. Previous research suggests that the incorporation of IFITMs into HIV-1 viral membranes is associated with restriction of virus fusion and spread(214). When IFITM proteins are expressed in non-infected lymphocytes, the proteins exert some protective effects upon cell-free virus infection, but the restriction disappears during cell-to-cell HIV infection. In contrast, when IFITM is expressed in virus-producing lymphocytes, IFITM proteins are observed to co-localise with viral HIV proteins in nascent virions and restrict virus spread(214). In addition, Foster and colleagues demonstrated Transmitted Founder viruses (TF), viruses that establish *de novo* mutations, are resistant to IFITM inhibition. In contrast, HIV-1 clones generated from individuals after six months of infection display an increase in sensitivity to inhibition by IFITM2 and IFITM3 proteins. They also demonstrate that this HIV-1 restriction is dependent on the strain's co-receptor usage and the localisation of the IFITM proteins within the cells(211). Specifically, they found that X4-tropic viruses are more sensitive to IFITM2 and IFITM3 than R5-tropic viruses. When specific mutation were introduced at the N-terminal region of either IFITM2 or IFITM3, R5-tropic viruses rather than X4-tropic viruses displayed greater sensitivity(211) (Figure 32). Taken together, both *in vitro* and *in vivo* assays provide some evidence on the role of IFITMs as modulators of HIV-1 transmission.

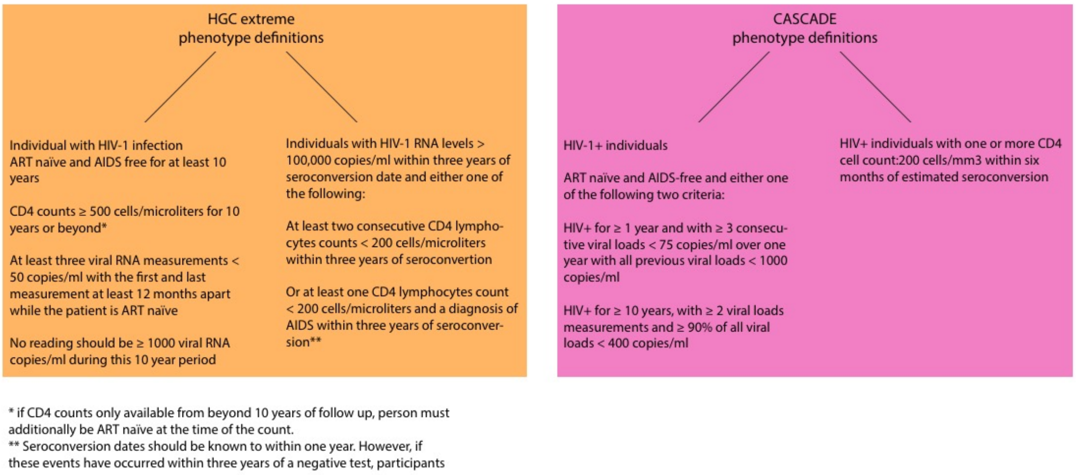
4.2. Aims

The aim of this project was to identify whether specific mutations in *IFITM* genes contribute to disease progression of HIV-1. In order to do this, I carried out targeted sequencing of the *IFITM* locus (Chr11:280,000-380,000) in patients classified as long-term non-progressors and rapid progressors from HIV Genome Consortium (HGC), UK Register of HIV Seroconverters and Conserted Action on SeroConversion to AIDS and Death in Europe (CASCADE) cohorts.

4.3. Methods

4.3.1. Study populations

I sequenced a total of 255 patients from three main cohort studies: 21 patients from HIV Genome Consortium (HGC), 52 patients from UK Register of HIV Seroconverters and 182 patients from Conserted Action on SeroConversion to AIDS and Death in Europe (CASCADE) Definitions for eligible participants changed slightly across cohorts. For example, HGC cohort recruitment came first with definitions under the title: HGC extreme phenotype definitions. Recruitment for CASCADE used a more relaxed definition also listed under CASCADE extreme phenotype definition. The UK Register of HIV Seroconverters started using HGC definitions for participant selection but widened it later on to include participants using the revised definitions set out by CASCADE (Figure 33).



HGC extreme phenotype definitions

Individual with HIV-1 infection ART naïve and AIDS free for at least 10 years

CD4 counts ≥ 500 cells/microliters for 10 years or beyond*

At least three viral RNA measurements < 50 copies/ml with the first and last measurement at least 12 months apart while the patient is ART naïve

No reading should be ≥ 1000 viral RNA copies/ml during this 10 year period

Individuals with HIV-1 RNA levels > 100,000 copies/ml within three years of seroconversion date and either one of the following:

At least two consecutive CD4 lymphocytes counts < 200 cells/microliters within three years of seroconvertion

Or at least one CD4 lymphocytes count < 200 cells/microliters and a diagnosis of AIDS within three years of seroconversion**

* if CD4 counts only available from beyond 10 years of follow up, person must additionally be ART naïve at the time of the count.
** Seroconversion dates should be known to within one year. However, if these events have occurred within three years of a negative test, participants

CASCADE phenotype definitions

HIV-1+ individuals

ART naïve and AIDS-free and either one of the following two criteria:

HIV+ for ≥ 1 year and with ≥ 3 consecutive viral loads < 75 copies/ml over one year with all previous viral loads < 1000 copies/ml

HIV+ for ≥ 10 years, with ≥ 2 viral loads measurements and ≥ 90% of all viral loads < 400 copies/ml

HIV+ individuals with one or more CD4 cell count:200 cells/mm3 within six months of estimated seroconversion

Figure 33. **Definitions adopted by the cohorts used in this study**. On the left (orange) is the definition from the HIV Genome Consortium (HGC). On the right in pink is the latest definition adopted by CASCADE and by HGC

4.3.2. Ethics approval

Ethics approval was granted by the ethics committees of each participating cohorts according to their local regulations with HDMCM numbers 11/070, 11/012 and 13/038. This included written informed consent from all participants taking part in the study. Consent was obtained for blood sampling, DNA sequencing and analysis of samples, storage of blood for future research, collection of demographics and anthropometric data and for access to clinical records.

4.3.3. Probe design

This study employed Agilent SureSelect targeted sequencing method to pulldown a region in chr11:280,000-380,000 referred to as 'The *IFITM* locus'. All the sample processing and sequencing was performed at The Wellcome Trust Sanger Institute. Briefly, genomic DNA was sheared to an average of 500bp using Covaris E210 (Covaris Massachusetts, USA). Sheared samples were used in the Illumina library preparation and enriched for the *IFITM* locus using SureSelect Agilent probes (Agilent technologies, Santa Clara, USA ELID number 0798441).

Samples were sequenced using the HiSeq 2500 (Illumina, SanDiego, USA) as paired-end 300bp reads. I sequenced 60 samples per lane in duplicates.

4.3.4. Sequencing analysis

I performed the alignment of the raw sequencing data to the to the human reference genome build GRCh37 using the Burrows-Wheeler Aligner (BWA-mem)(216) and marked duplicates using Picard (http://broadinstitute.github.io/picard, version 2.7.2). Because the same samples were sequenced in two different lanes, I merged bam files belonging to the same samples using SAMtools(217) and followed the GATK(218) best practice guidelines(166, 167) for bam improvement prior variant calling. The steps included duplicate marking (http://broadinstitute.github.io/picard, version 2.7.2) and base quality score recalibration (GATK 3.6). Variant calling was performed at the single sample level using the Haplotype Caller (GATK 3.6) and then jointly across individuals using GATK CombineVCF and GenotypeVCF (GATK 3.6).

For variant QC, I applied GATK hardfiltering recommendations(167) as described in the previous Chapter (Figure 34). The hardfiltering parameters included were: QualByDepth (QD), Fisher Strand (FS), StrandOddsRatio (SOR), RMSMapping Quality (MQ) MappingQualityRankSumTest (MQRankSum), ReadPosRankSum. I tested various filters by applying different filtering thresholds. Ultimately, I decided to use the filters that provided me with a ratio of transition to transversion (Ts/Tv) > 2.8 in all samples.

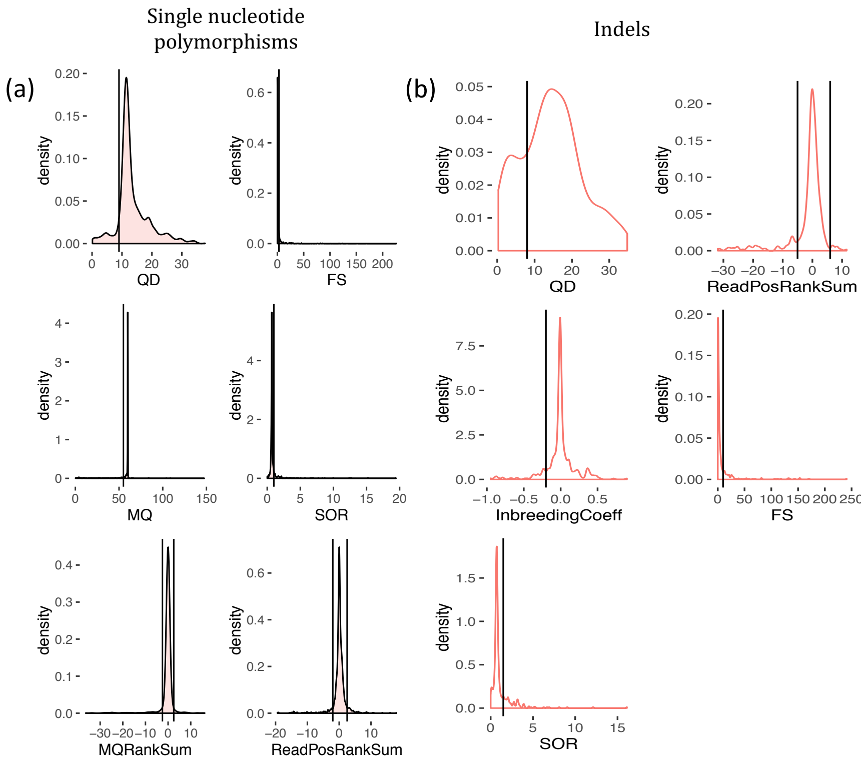


Figure 34. **Density plots of QC parameters used to filter out bad quality SNPs and Indels.** On the x-axis are the annotation values used to assess the quality of the dataset, and on the y-axis, are the density values. The black line represents the threshold values used to hardfilter variants **(a)** Distribution of several QC parameters on single nucleotide polymorphisms (SNPs). The distribution of Quality by Depth (QD) values occur from 0-30, with two peaks representing heterozygous reads (QD =12) and homozygous calls with approximately double the number of reads (QD = 18). Variants with QD values < 9 represent low quality calls and were excluded from the dataset. SNPs have Fisher Strand (FS) values close to zero showing no strand bias. The black lines show the cut off value FS >1 used to exclude variants that deviate from zero. SNPs Mapping Quality (MQ) values are close to 60 and of the highest quality. Any variants with MQ <55 were excluded from the analysis. The majority of SNPs had Strand Odds Ratio (SOR) values close to zero. To exclude variants with some degree of strand bias, SOR values > 3 were used as cut off values. Variants with MQRankSum < -3 (x3 more reads supporting reference alleles) and MQRankSum > +3 (3x more reads supporting alternate alleles) were also excluded in the analysis. Finally, the dataset showed that the distribution of the Read Position Rank Sum (ReadPosRankSum) is close to zero indicating that there is little or no difference between the SNP positions within the reads. **(b)** Distribution of several QC parameters on indels. Similar to SNPs QC steps, Quality by Depth (QD) values occur from 0-30, with two peaks representing heterozygous reads (QD =15) and homozygous calls with approximately double the number of reads (QD = 30). Variants with QD values < 9 represent low quality indels and were excluded from the dataset. Indels that exhibited any positional bias (ReadPosRankSum) because they were either seen always at the start or end of reads were excluded from the dataset. Likewise, indels that displayed some level of strand bias were also excluded (FS and SOR). The inbreeding coefficient provides a measure of inbreeding within the data and measures an excess of heterozygous sites in the dataset. Higher than expected heterozygous indels were excluded from further analysis.

4.3.5. Sequencing analysis

Following variant calling and QC, I annotated variants using dbSNP v137. Functional annotations were then added using Ensembl Variant Effect predictor (VEP, version 84) keeping the most severe consequences per gene(219). I followed Scoring Intolerance from Tolerance (SIFT)(220) and Polyphen-2(221) predictions to determine whether the variants were likely to affect amino acid sites and the Sequence Ontology terms and description in Ensembl to score variants for their functional impact. In this study, I found 1,286 variants with various coding consequences (Figure 35). I also found 331 variants that were not reported in the 1000 Genomes dataset. In total, I discovered 267 novel variants that so far have not been reported in any dataset including dbSNP. Of these novel variants, 42% were missense mutations and the rest were synonymous mutations (Figure 35).

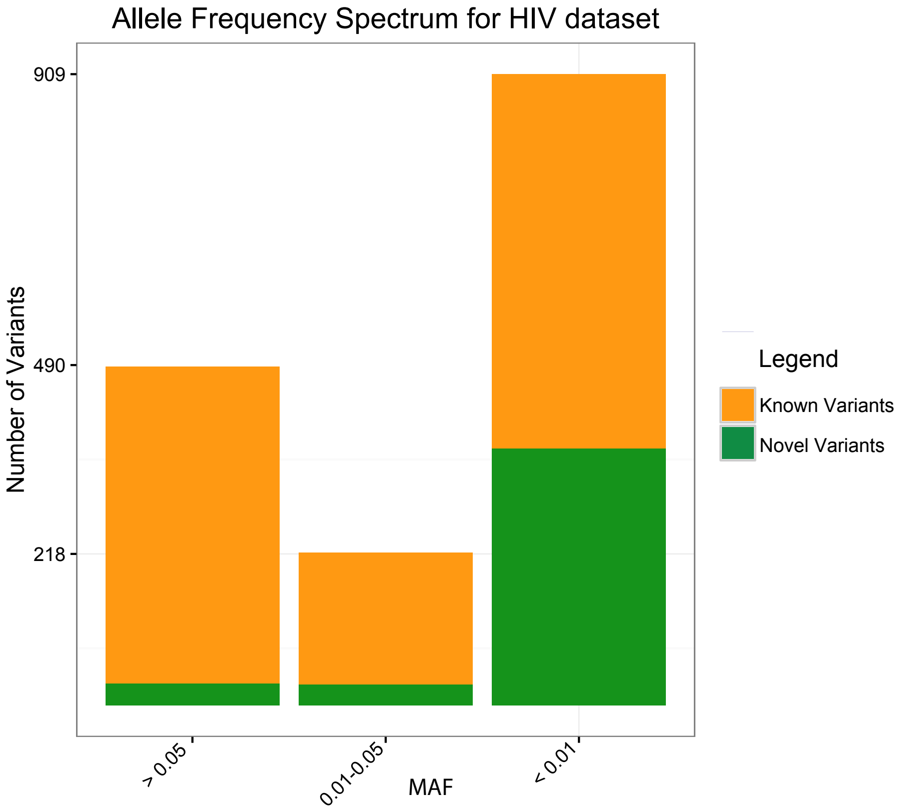


Figure 35. **Minor Allele Frequency (MAF) spectrum for variants in the HIV dataset.** Distribution of variants in the dataset that are common (≥ 5% frequency), medium (1-5% frequency) and rare (< 1% frequency). Within each group, variants were coloured yellow if they had been previously reported in 1000 Genomes Project and green if they had not been reported before on any known dataset.

4.3.6. Identification of individuals with elevated missing rate or outlying heterozygosity

Before embarking in downstream analysis, I carried out a number of QC steps to assess the quality of the sequencing data. These steps were conducted at the individual and variant levels.

Individual level QC

Generally, scrutinising the distribution of missing genotypes for individual samples is the best strategy to identify an adequate threshold to filter out individuals with excessive missing genotypes. Typically samples with more than 3-7% missing genotypes are removed(132). For this dataset, a group of 28 samples (22 elite controllers and 6 rapid progressors) had approximately 10% of missing genotypes and were excluded for downstream analysis (Figure 36).

Similarly, the distribution of mean heterozygosity (with the exception of the sex chromosomes) at the individual level can identify samples with an excessive or reduced number of heterozygous calls, that may indicate DNA sample contamination or inbreeding, respectively(222). For this dataset, only two samples were excluded as a result of low heterozygosity rate (Figure 36).

Variant-level QC

In case-control genetic studies, it is important to exclude variants with genotype missing rates >20% between cases and controls. This QC step is essential to ensure that differences observed between Elite controllers and Rapid progressors are not due to technical artefacts as a result of poor genotype calls. I excluded 17 out of a total 1,618 variants after imposing a cut off threshold of 20% for missing genotype calls (Figure 37).

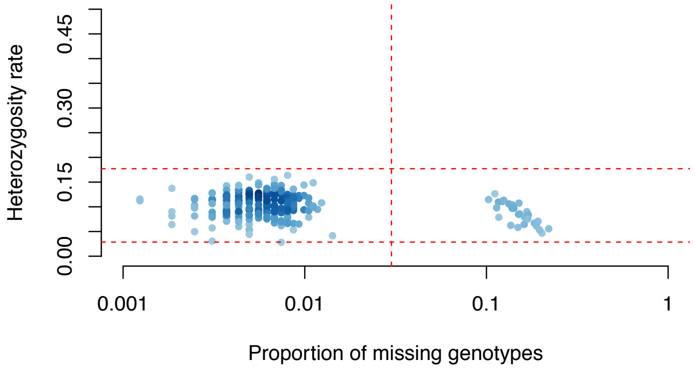


Figure 36. **Representation of heterozygosity and missing rate in the HIV cohort**. The x axis represents the distribution of missing genotypes for each individual. Any individuals with > 2% missing genotype calls were excluded (vertical dashed line). This resulted in a total of 28 samples excluded from downstream analysis. The y-axis represents the distribution of the heterozygosity rate for each individual. Individuals with heterozygosity rate ± 3.5 standard deviations from the mean were also excluded from downstream analysis. A total of two samples were excluded due to reduced heterozygosity.

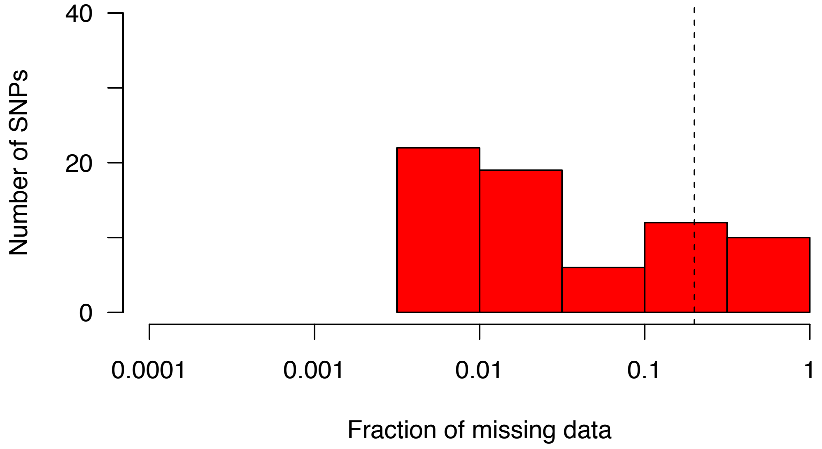


Figure 37. **Histogram of the number of variants with excessive missing data rate.** The missing data rate was plotted across all individuals that passed the per-individual QC. The dash line represents the threshold (3%) for missing data which was imposed on the dataset. SNPs with this level of missigness were excluded due to excess failure rate.

4.3.7. Other QC metrics on whole exome data for the same samples

My dataset only included genomic information for a 100kb region, Chr11:280,000-380,000, and did not allow for the evaluation of ethnicity, sex discordance or individual relatedness. However, whole exome data for 238 of the total 255 individuals, was kindly obtained from Prof Paul McClaren and Prof. Kholoud Porter. This exome dataset allowed me to evaluate ethnicity, sex discordance and relatedness for 94% of my samples. The remaining 6% (15 individuals) that were not included in the whole exome sequencing data provided, were excluded from all subsequent analysis.

One of the main sources of cofounding in candidate gene studies is population stratification; that is, genotyping differences that can be attributed to divergent population origins rather than differences within the specified disease trait. The most common analysis to detect individuals with differing ancestries is the principal component analysis (PCA). Genome-wide data is used due to the large number of SNPs (markers) needed to make accurate PCA predictions (> 50,000). PCA predictions require a set of observations (i.e. individuals) and co-related variables (i.e. the markers). Filtered genome wide datasets from 1000 Genomes for each population are commonly used to detect larger continental level ancestries. PCA calculations produce a set of uncorrelated variables (or principal components) from the information matrix that contains the observations and the variables for each individual(132).

I evaluated the ethnicity of my samples via Principal Component Analysis (PCA)(223, 224). To do this analysis, I included only autosomal bi-allelic variants, with minor allele frequency > 5% that did not deviate significantly from the Hardy-Weinberg equilibrium (HWE < 10^{-5}) and with a call rate > 90% across all samples. I then took the overlapping variants between the whole exome dataset and 1000

Genome, phase 3 dataset and pruned markers in high linkage disequilibrium (LD)(225) using PLINK version 1.9. The PCA calculation was carried out using EIGENSTRAT package(224). Due to the genetic differences across populations in the 1000 Genome dataset, only two principal components were sufficient to cluster individuals in the exome dataset alongside individuals from 1000 Genomes dataset. The results from the PCA analysis demonstrated that a total of 66 samples were not of European descent and should be excluded from downstream analysis (Figure 38).

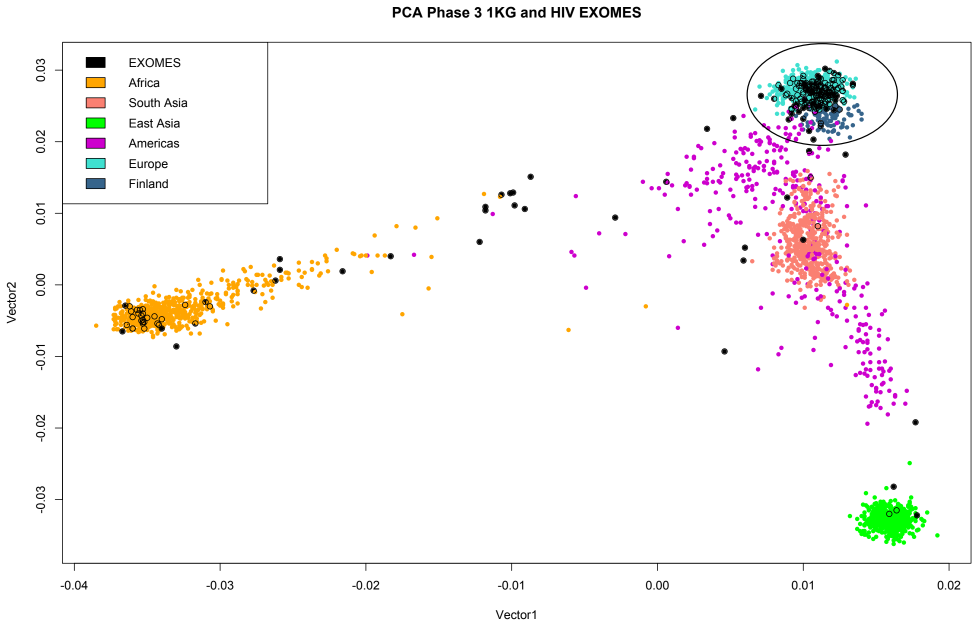


Figure 38. **Principal component analysis (PCA) of whole exome HIV sequencing data.** Principal component clustering was built using 2,504 individuals from African, South Asian, East Asian, American, Finish and European populations from 1000 Genomes phase 3 dataset. These populations were used to predict the ancestry of the 240 samples (of supposed European ancestry) from HIV whole exome dataset. A total of 68 samples clustered away from the European samples and were excluded from the study. The circle surrounding the black dots represent the 172 HIV whole exome samples that clustered with European samples and thus, included in downstream analysis.

Furthermore, I identified 4 samples with identity by state (IBD) score > 0.185, a value which is suggestive of relatedness or duplication between individuals(132). Any standard population-based study require that all the samples are unrelated. In datasets that contain related individuals, there will be an overrepresentation of family genotypes. IBD is calculated based on the average common proportion of alleles shared by all. Typically, the estimated threshold to remove samples if IBD > 0.185 which represents a midway value of what it is expected for third and second-degree relatives. Finally, IBD > 0.98 identifies duplicates or monozygotic twins(132). I did not find discordant sex information amongst the remaining samples. As a result of all QC steps, I was able to test association between 92 rapid progressors and 60 elite controllers (152 individuals).

4.3.8. Statistical power to detect association

The statistical power to identify genetic variants of genome-wide significance and with different effect sizes given the sample size was estimated using QUANTO software (http://biostats.usc.edu/software). These calculations were done by Dr. Neneh Sallah and Fernando Riveros Aguilera at the Wellcome Trust Sanger Institute (Figure 39 and Figure 40).

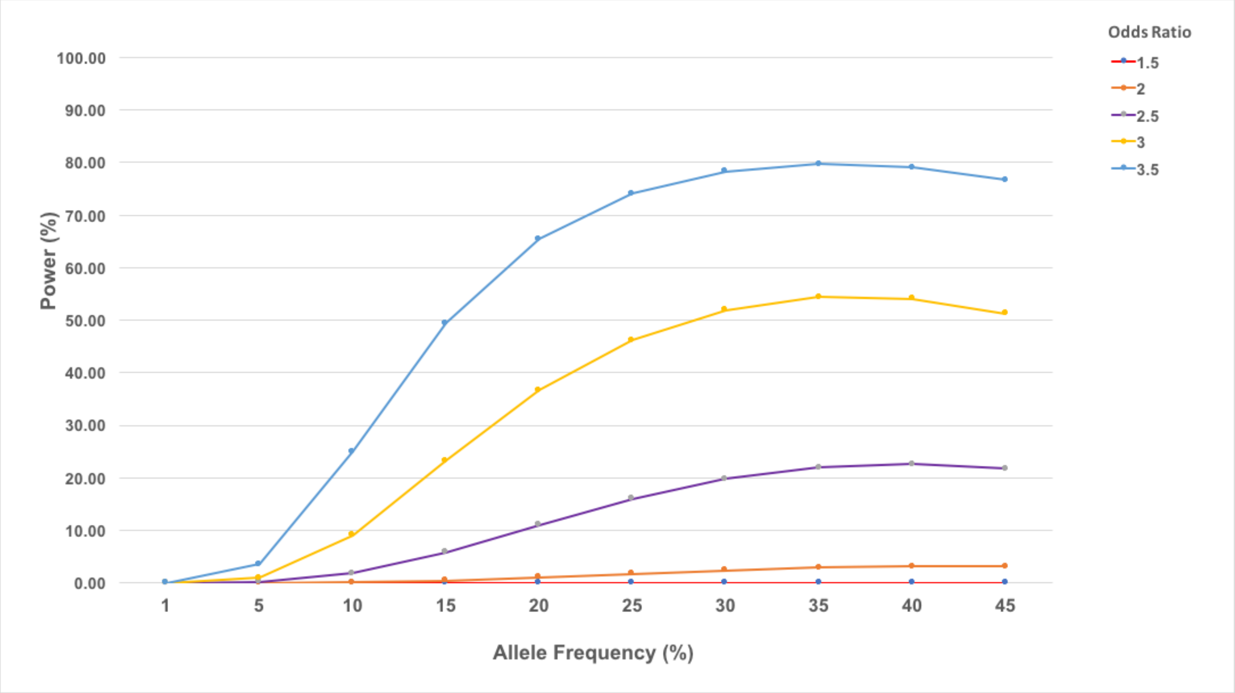


Figure 39. **Power calculations for original number of samples**. Statistical power (%) to identify genetic variants at p<5x10$^{-8,}$ given different allele frequencies (%) and effect sizes (OR) (N=255).

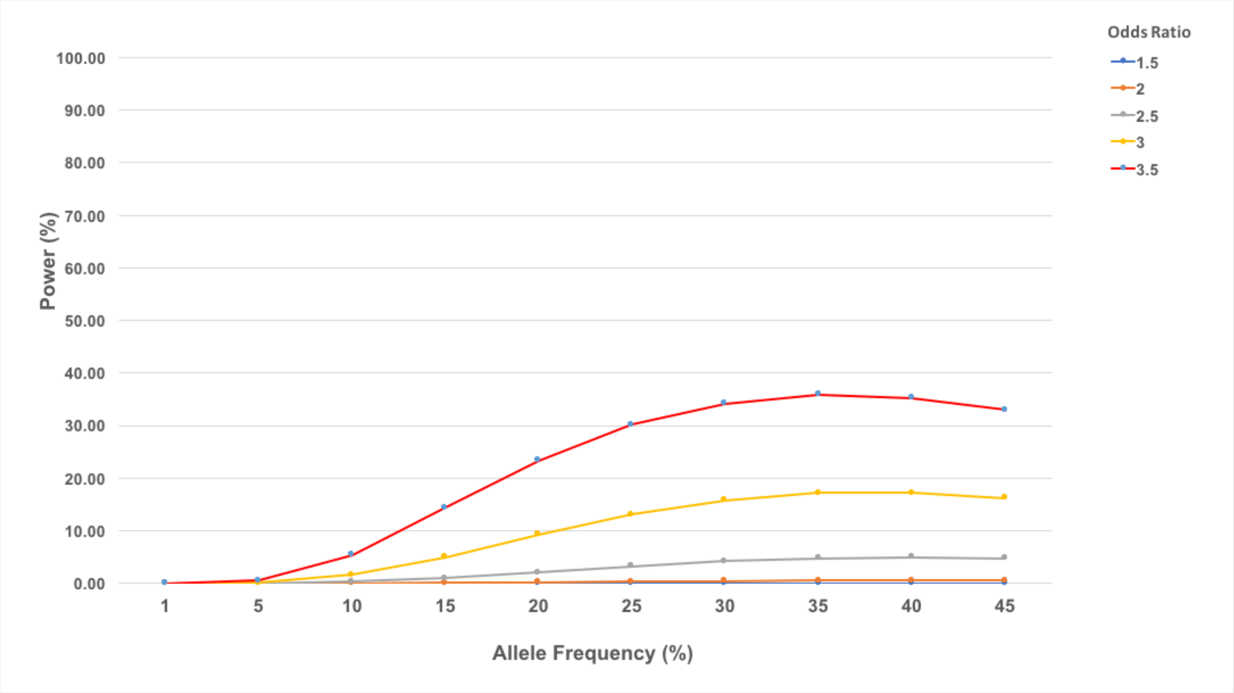


Figure 40. **Power calculations after QC**. Statistical power (%) to identify genetic variants at $p<5x10^{-8,}$ given different allele frequencies (%) and effect sizes (OR) (N=255).

4.4. Results

4.4.1. Searching for association to HIV progression.

4.4.1.1. Single variant association tests

I tested genetic variants in and around *IFITM1, IFITM2* and *IFITM3* for association with the rapid disease progression in HIV using a cohort of 60 elite controllers and 92 rapid progressors. To ensure that any differences in the frequency of variants between elite controllers and rapid progressors were not the result of different ethnicities, I ensured that all cases and controls were of European descent. As highlighted by the PCA analysis, I had to exclude approximately 25% of the original samples due to the inclusion of non-European individuals. I tested all variants that passed the quality control filter (n =1,617) as described in my methods. In this study, no variants reached genome-wide significance. The nominally significant variants are shown in Table 18.

Table 18. Case-control association tests (Fisher exact test) for variants in the *IFITM* locus. P-values are not corrected for multiple testing but none reached genome-wide significance (5x10^{-8})

Genes	rs ID	AF cases	AF controls	P- value	OR (95% CI)	Consequence	AF (1KP)	AF (ExAC)
B4GALNT4	rs1134699	0.47	0.30	0.0029	2.1 (1.2-3.4)	synonymous_variant	0.50	0.45
B4GALNT4	rs10902142	0.38	0.22	0.0037	2.2 (1.2-3.6)	intron_variant	0.45	
B4GALNT4	rs10794316	0.37	0.22	0.0052	2.1 (1.2-3.6)	intron_variant	0.37	
B4GALNT4	rs10751657	0.37	0.22	0.0052	2.1 (1.2-3.6)	intron_variant	0.45	
IFITM3	rs55671406*	0.46	0.30	0.0058	2.0 (1.2-3.3)	upstream_gene_variant	-	
B4GALNT4	rs12360752	0.46	0.30	0.0058	2.0 (1.2-3.2)	intron_variant	0.50	
B4GALNT4	rs7481525	0.47	0.31	0.0061	2.0 (1.2-3.2)	intron_variant	0.50	
IFITM3	rs56232455	0.35	0.20	0.0065	2.1 (1.2-3.7)	upstream_gene_variant	0.17	
B4GALNT4	rs12361394	0.38	0.23	0.0078	2.1 (1.2-3.8)	intron_variant	0.41	
ATHL1	rs56069858	0.32	0.18	0.0082	2.1 (1.2-3.7)	upstream_gene_variant	0.51	
B4GALNT4	rs7483942	0.46	0.31	0.0085	1.9 (1.2-3.1)	intron_variant	0.50	
IFITM3	rs56228238*	0.32	0.19	0.0170	2.0 (1.2-3.5)	upstream_gene_variant	-	
B4GALNT4	rs55794317	0.03	0.09	0.0179	0.28 (0.1-0.8)	intron_variant	0.02	0.06
B4GALNT4	rs7120441	0.34	0.21	0.0192	1.9 (1.1-3.3)	intron_variant	0.34	
B4GALNT4	rs7396812	0.37	0.24	0.0233	1.8 (1.1-3.0)	intron_variant	0.40	0.33
B4GALNT4	rs7395781	0.37	0.24	0.0233	1.8 (1.1-3.0)	intron_variant	0.40	0.13
B4GALNT4	rs35842721	0.33	0.21	0.0265	1.9 (1.1-3.2)	intron_variant	0.38	
B4GALNT4	rs34063493	0.33	0.21	0.0265	1.9 (1.1-3.2)	missense_variant	0.38	
B4GALNT4	rs35475866	0.33	0.21	0.0265	1.9 (1.1-3.2)	synonymous_variant	0.33	8.88x10^{-6}
IFITM2	rs9704108	0.05	0.13	0.0276	0.36 (0.2-0.9)	5_prime_UTR_variant	0.97	

4.4.1.2. Replicating the association for *IFITM3* SNP rs12252

In a genetic association study of HIV-1 infection prognosis consisting of 74 rapid progressors and 104 elite controllers, Zhang and colleagues reported an association for SNP rs12252 (P = 0.002, OR = 3.778 (95CI- 1.5–9.7) under a dominant model. They found that 92% of rapid progressors (68/74) carried the CC/CT genotypes, compared to 75% of non-progressors (78/108)(88).

In my study, I was unable to replicate this finding. In fact, I found that the direction of effect for the C allele is in the opposite direction to what is reported in their study (OR = 0.64 (95CI - 0.12-3.25, additive model) (Table 19). Although, I carried out the analysis under different models: Cochran-Armitage trend test, a 1df allelic test, a dominant model (for the minor allele) test and a recessive model (for the minor allele), I was unable to obtain significant evidence of replication under the dominant or recessive models. One potential reason for the lack of replication is the low frequency of this risk allele (give the frequency and the allele in your cohort) in the cohort analysed in this study, compared to the PRIMO cohort(88). Of course, another potential reason for the lack of replication could be that the initial association is a false-positive finding.

Table 19. Table of association tests for SNP rs12252. The top row "This study" shows the association summary statistics in my analysis. The values for Zhang *et al,* are shown underneath. *P*-values are uncorrected for multiple testing but do not reach genome-wide significance (5x10^{-8})

Study	**ID**	**Cases**	**Control**	**P**	**OR (95% CI)**	**AF**	**EAS AF**	**EUR AF**
This study	rs12252	0.02	0.28	0.59	0.64 (0.12-3.25)	0.24	0.53	0.04
Zhang *et al, 2015*	rs12252	0.91	0.75	0.004	3.78 (1.51–9.72)	0.24	0.53	0.04

4.4.1.3. Aggregate variant tests

To assess the role of low frequency variants, a way to achieve higher statistical power is to combine those variants that are likely to have an impact on gene function and compare their distribution in two phenotypic groups(135) using aggregate variant tests. Because there is no set strategy for variant selection in collapsing tests, scientists often select variants based on their minor allele frequencies (MAF) information and their predicted functional consequence.

To test different hypotheses regarding the role of specific sets of rare variants on HIV prognosis, I performed a series of different analyses. In a first collapsing test, I included all missense variants, regardless of their MAF. I also grouped all missense variants for *IFITM*1, 2 and 3 genes because the number of variants in individual genes was insufficient to provide even reasonable statistical power. In a second test, I only included missense variants that had been reported as rare by the 1000

Genomes dataset (MAF < 1%, European 1KP populations) and again grouped variants for *IFITM*1, 2, 3. In the final test, I collapsed all rare missense variants (MAF <1%) for the whole locus (Chr11:280,000-380,000) that were predicted to be deleterious by Polyphen-2(221) or SIFT(220) as define by VEP(219) annotation. I ran two statistical tests: Variable threshold (VT)(226) and SKAT-O. The VT test, does not rely on fixed MAF thresholds; instead, the test is aimed at specific scenarios when the likely allele frequency and effect sizes of the variants in question are unknown. Another reason for using VT tests is that mean-based tests are reportedly more powerful than other alternatives for less stringent levels of significance(227). The Sequencing Kernel Association test (SKAT)(228) is a type of dispersion test that is robust against the presence of non-causal, protective and non-deleterious variants. Power of the SKAT test can be compromised if the conditions of a burden test are met, for this reason Lee *et al.*, (2012) proposed SKAT-O(229), a test that combines burden and SKAT statistics and the one I used in this study (Table 20). As a result of these tests, I found no genome-wide significant associations.

Table 20. Case-control aggregate variant tests. P-values are uncorrected for multiple testing but none reach genome-wide significance

Gene	**Variants in cases**	**Variants in controls**	**Scenario**	**SKAT-O *P*-value**	**VT *P*-value**
ATHL1	5	3	All missense variants	0.93	0.83
IFITMs (1,2,3)	4	3	All missense variants	0.8	1
B4GALNT4	7	11	All missense variants	0.12	1
ATHL1	5	3	Rare missense variants (1KP)	0.93	0.83
IFITMs (1,2,3)	1	1	Rare missense variants (1KP)	1	1
B4GALNT4	7	11	Rare missense variants (1KP)	0.12	1
Whole locus	9	3	All missense variants with most deleterious effects	0.73	1

4.5. Discussion

The patients that I describe in this study represent two phenotypically distinct groups of HIV+ positive individuals, that constitute less than 0.8% of the HIV infected population(128). Their extreme phenotype characteristics had been carefully defined according to stringent clinical definitions, set-up by the relevant cohorts. In this study, HIV-elite individuals are natural suppressors of HIV (ART naïve) who are able to control viral replication to undetectable levels, for at least a year. In contrast, rapid progressors have very low CD4+ cell counts (<200 cell/mm^3) and can progress within six months of seroconversion. As I mentioned in the introduction, despite great efforts to characterise the genetic factors driving such extreme phenotypes, the discovery of new important loci has not been successful(99).

In this Chapter, I explored the contributions of the *IFITM* genes on HIV disease progression. The in-depth functional data on the role of *IFITM2* and *3* on HIV-1 infection(211, 214), coupled with reports of association of *IFITM3* SNP rs12252 with rapid HIV progression(89), motivated us to consider *IFITM* genes as potential gene candidates for HIV progression. I used a similar probe design from Chapter 3 to target a 100Kb region of the *IFITM* locus (Chr11:280,000-380,000) which is flanked by *B4GALNT4* and *ATHL1,* encompassing four *IFITM*-family genes (*IFITM*1,2,3,5). To my knowledge, this is the first time an *IFITM* candidate gene study has been carried out which has looked at most of the variation in the region(88, 230).

Although the study was limited by the sample size (due to the rarity of the phenotype), I controlled for some of the power limitations by sampling from the most extreme of the disease spectrum and enriching for severity alleles(231). The initial design of this study with 122 elite controllers and 133 rapid progressors, using a genome-wide threshold of P < 5x10^{-8}, had 80% power to detect common variants with minor allele frequencies of at least 30% with very large effect sizes

(OR = 3.5) (Figure 39). This effect size and allele frequency are similar to those reported by Zhang and colleagues for their influenza study. For lower frequency variants of 5%, the study had less than 5% power to detect variants with large effect size (OR = 3.5). Following the exclusion of a total 86 individuals from the study, a substantial drop in power occurred. Following quality control assessments, I had a total of 60 elite controllers and 92 rapid progressors, and 18% power to detect common variants with minor allele frequencies of at least 30% with very large effect size (OR = 3.5) (
Figure 40). The study had no power to detect lower frequency variants of 5%. No genome-wide significant loci were discovered although this is not surprising given the low sample sizes and power to detect associations.

4.5.1. The role of rare variants in HIV-1 disease progression

The involvement of rare variants in HIV-1 disease progression is a topic of debate in the field(232). To identify the involvement of rare variants in the architecture of the disease, I performed a number of aggregate variant tests in a case-control design. These approaches are advantageous because they consider alleles that may have full or incomplete penetrance on the disease. Although, I did not find any associated variants, I cannot rule out that given a larger cohort of individuals, rare variants in the IFITM locus could contribute to HIV progression.

There is good precedent for exploring the rare variant hypothesis in HIV, especially given the number of reports that have demonstrated, *in silico*, that a burden of rare mutation is important for several complex traits. For example, whole exome sequencing of 81 unrelated individuals suffering from the atrioventricular septal defect (AVSD) established that genes with previously known biological association to AVSD such as cohesion loading factor (*NIPBL*) and Zinc Finger Protein (*ZFPM2*) were enriched for rare and damaging non-synonymous variants(139). Using a similar rare burden analysis approach, Grozeva, and colleagues, found that highly pathogenic loss of function and missense variants were present in ~11% of intellectual disability cases from a cohort of 986 individuals (113 variants in 107 individuals)(145). Using a much larger cohort of individuals of 4,264 cases and 9,343 controls, Singh and colleagues detected an enrichment of rare loss of function variants in SET Domain Containing 1A (*SETD1A*) gene that encodes a protein from a histone methyltransferase (HMT) complex(141). Although the number of samples in each individual cohort varied and functional follow-up studies will be necessary, it appears that the contribution of some loss of function and missense rare variants to the risk of disease such as heart conditions, intellectual disability or schizophrenia, is considerable. It is possible that the same approach could be used in future experimental designs to study HIV.

The design of a comprehensive analysis of host factor contribution to HIV-1 progression is still of paramount importance and future studies could also investigate other forms of variation such as copy number variation(128). Differences in copy number variants have been shown to be important for HIV-1 disease progression within specific populations. For example, if an individual carries copy number of C-C Motif Chemokine 3-Like 1 (*CCL3L1*) lower than the average copy number observed for the general population from which that individual originates, he/she will be more susceptible to HIV-1 infection(128). It is believed that increased doses of *CCL3L1* could affect HIV-1 in a number of ways: by inhibiting the binding of gp120 to the CCR5 co-receptor, inducing CCR5 co-receptor internalisation, leading to a decrease in the receptors on cell membranes or influencing leukocyte trafficking involved in virus restriction(128).

4.5.2. Future candidate gene studies for *IFITM*

Despite the lack of association signals, this study had the advantage of controlling for a number of factors such as ethnicity, sex and relatedness that would not normally be possible in standard candidate gene designs. In future, possibly through data sharing and meta-analysis, I hope the results presented in this chapter can be combined with those from other sequencing studies to enable more powerful tests to be performed to more fully assess the role of *IFITM* genetic variation in HIV progression.

5. Assessing the contribution of IFITM variation to dengue haemorrhagic fever

5.1. Introduction

5.1.1. Global burden of dengue fever

Dengue fever is a mosquito-borne infection, endemic in approximately 100 countries in the Americas, Southeast Asia, Africa, the Western Pacific, Africa, and the eastern Mediterranean area. There are between 350 million infections every year caused by four known serotypes of dengue virus (DENV1-4) (WHO, fact sheet, July 2016). In Vietnam alone, as of 31st August 2016, there were 63,504 reported cases of dengue and 20 deaths, in 44 out of 63 provinces in 2016. In August 2016 alone, there were 16,547 cases reported including 4 fatalities. Compared to the same period in 2015, the cumulative number of cases increased by 97%. The number of cases in 2016, also represents an increase of 99.7% compare to the median in the period between 2011-2015 (WHO, Dengue situation update 504). Furthermore, this number is likely to be a gross underestimation of dengue incidence due to lack of effective disease surveillance, misdiagnosis and low reporting levels(233).

The clinical manifestation for the dengue varies from asymptomatic infections to mild (dengue fever) and severe (dengue haemorrhagic fever and dengue shock syndrome). The WHO defines dengue fever (DF) as an acute condition that can cause fever, headaches, exanthema, severe muscle and joint pain, and bleeding of gums. Although dengue fever is an incapacitating disease, the prognosis is generally favourable for most patients, who recover without other clinical complications after several weeks(234).

Dengue haemorrhagic fever (DHF) and dengue shock syndrome (DSS) are the most severe forms of the disease. DHF is usually correlated with secondary dengue infections(235) but may sometimes also occur at primary infections. DHF has similar symptoms to DF but patients also suffer from severe abdominal pain that may be a sign of abdominal haemorrhage(235). There are also atypical complications such as damage to specific organs such as kidneys and liver that may cause severe complications in patients(236). More importantly, if left untreated, DHF could progress to the potentially fatal dengue shock syndrome (DSS). The WHO defines DSS as a form of hypovolaemic shock resulting from continued vascular permeability and plasma leakage. This usually takes place around defervescence, on days 4–5 of illness. At least in Vietnam, epidemiological studies of dengue infections have found that both DHF and DSS are also most commonly observed in children(237). For example, an elegant retrospective study looking at dengue patients admitted to three hospitals in Ho Chi Minh City, Vietnam reported that from a total of 14,079 DSS patients diagnosed between 1996 and 2009, 96.6% were children. The mortality was also highest amongst children (< 15 years old) than in adult dengue patients (case fatality rate 0.20% versus 0.11%; *P*=0.002). However, amongst those adults with DSS (N=484) the mortality was higher than in paediatric cases (CFR 5.5% versus 1.4%; P <0.001)(237). The same study also reported that the majority of deaths occurred in patients diagnosed with DSS. The overall case fatality rate amongst DSS patients was 1.6% (153/9,784) compared with 0.03% (28/92,683) among clinically diagnosed dengue patients that did not develop DSS(237). Many individuals with DSS respond to resuscitation with isotonic crystalloid solutions, but patients not responding to treatment often require fluid resuscitation with crystalloid or colloid solution and blood transfusion. Mortality rates for DSS vary from (<1% to >10%) depending on access to healthcare(237).

5.1.2. Dengue infection and transmission

Dengue is transmitted primarily by the *Aedes aegypti* mosquito but also by *Aedes albopictus* which is the main vector in Asia. The virus is transmitted to humans through the bites of infected female mosquitoes. Meta-analysis of dengue virus incubation period suggests that the extrinsic incubation period (EIP), which represents the time when a mosquito takes a blood meal and becomes infected, is between 5 and 33 days at 25°C, and 2 and 15 days at 30°C. The intrinsic incubation period (IIP) which is the time between a person being infected and the onset of symptoms due to the infection, has an estimate of 5.9 days(238). Dengue virus can also be transmitted in utero, with one specific study reporting that vertical transmission occurred in 18 of 143 (12.3%) pregnant women in their case report(239). One comparative study also found evidence of vertical transmission in 2.5% of cases (n=63), with a vertical transmission rate of 1.6%(240) in pregnant women. It has been proposed that dengue-positive individuals who experience a subsequent infection with another serotype are at higher risk of developing severe dengue (DHF/DSS). A retrospective study of 1,757 children (aged between 4-16 years) found that zero of 47 children with primary dengue infections were hospitalised, whereas 7 of 56 children with secondary infections required hospital care (P = 0.012). They concluded that pre-existent dengue immunity to one strain, as detected by conventional serologic techniques, places the patient at a significant risk (odds ratio greater than or equal to 6.5) for development of dengue haemorrhagic fever when infected with a different DENV strain. These observations support the hypothesis of antibody-dependent enhancement (ADE) in dengue infections. One of the first lines of evidence for ADE in dengue came to light when it was reported that passive transfer of maternal dengue antibodies to the foetus significantly increase the likelihood of acquiring DHF/DSS in infants(241). One hypothesis is that the pre-existing antibodies use Fc receptors in the target cells to form complexes that facilitate the infection of cells such as monocytes, macrophages and dendritic cells(242). Several *in vitro* studies have since reproduced the enhanced infection of Fc-receptor bearing cells to mimic the cellular infection observed in DHF/DSS patients(243). Furthermore, the transfer of DENV-specific monoclonal antibodies into juvenile rhesus monkeys resulted in a notable clinical manifestation and viraemia(244).

5.1.3. The genetics of dengue infection

So far, no single gene has been associated with susceptibility to DENV infections but several host factors have been found to play a role in the severity of the disease. For example, a study of the *HLA-A* and B genes in a Thai population (DF=149 and DHF=114 and control=140 individuals) constituted the first evidence of an association (albeit with no genome wide significance) between severe dengue phenotype predisposition and individual genetic composition, virus serotype, and primary/secondary virus infection(245). *HLA-A*0203* was associated with the less severe DF (*P*=0.012, OR=3.09), regardless of the secondary infecting virus serotype. Conversely, *HLA-B*52* (*P*=9.6x10$^{-6,}$ OR=26) was associated with DF in patients with secondary DENV-1 and DENV-2 infections(245).

In Vietnam, for example, several studies have determined that close to 85% of the population is exposed to dengue virus by the time they turn fifteen years old, but only 1% suffer the most severe symptoms. These epidemiological studies have served to implicate host factors with susceptibility to severe dengue infections(246). Two regions: major histocompatibility complex (*MHC*) class I polypeptide sequence B (*MICB*, *P*=4.4x10^{-11}, OR=1.34) locus and phospholipase C, epsilon 1 (*PLCE, P*=3.1x10^{-10}, OR=0.8) gene confer susceptibility to the most severe forms of dengue infection in children(113). Although identification of the causative loci was not possible, MICB appears to be the most likely candidate for disease severity in this cohort(113) (Table 21).

Table 21. Table showing the most relevant genetic studies of dengue fever or dengue shock syndrome

Genes	SNP	P value	Odds ratio	Phenotype	gene location	Sample collection	sample size	publication	Type of study
MICB	rs3132468	P= $4.4\mathrm{x}10^{-11}$	1.34	Dengue Shock syndrome	Chromosome 6: 31,494,881-31,511,124	Vietnamese	2008 DSS and 2,018 controls. Follow up replication studies in 1,737 cases and 2,934 controls	Khor, *et al.* 2011	GWAS
PLCE1	rs3765524	P= $3.1\mathrm{x}10^{-10}$	0.8	Dengue Shock syndrome	Chromosome 10: 93,993,989-94,332,823	Vietnamese	2008 DSS and 2,018 controls. Follow up replication studies in 1,737 cases and 2,934 controls	Khor, *et al* . 2011	GWAS
MICB	rs3132468	P =0.0213	1.58	Dengue Shock Syndrome	Chromosome 6: 31,494,881-31,511,124	Thailand	76 DSS and 841 non-DSS (DHF and normal controls)	Dhang, *et al.* 2014	Genotyping. Replication study
PLCE1	rs3765524	P = 0.0252	1.49	Dengue Shock Syndrome	Chromosome 10: 93,993,989-94,332,823	Thailand	76 DSS and 841 non-DSS (DHF and normal controls)	Dhang, *et al.* 2014	Genotyping. Replication study
CD209	rs4804803	P = $1.4\mathrm{x}10^{-7}$	5.84	Carrier frequency of 4.7% in individuals with DF compared with 22.4% in individuals with DHF	Chromosome 19: 7,739,994-7,747,564	Thailand	606 individuals with dengue disease and 696 healthy population controls from the same hospitals	Sakuntabhai, *et al.* 2005	CD209 screen
JAK1	rs11208534	P= 2.9×10^{-3}	5.2	Increased risk for DHF.	Chromosome 1:64833229-64966504	Brazil	50 Brazilians with probable or possible DHF, 236 with DF, and 236 Brazilians with asymptomatic infections	Silva, *et al.* 2010	Genotyped for 593 SNPs in 56 genes across the type 1 interferon (IFN) response pathway as well as other important candidate genes
JAK1	rs2780831	P= 2.7×10^{-3}	2.6	Increased risk for DHF.	Chromosome 1:64833229-64966504	Brazil	50 Brazilians with probable or possible DHF, 236 with DF, and 236 Brazilians with asymptomatic infections	Silva et al. (2010)	Genotyped for 593 SNPs in 56 genes across the type 1 interferon (IFN) response pathway as well as other important candidate genes
JAK1	rs310196	P=3.0×10^{-4}	0.3	Protection for DHF.	Chromosome 1:64833229-64966504	Brazil	50 Brazilians with probable or possible DHF, 236 with DF, and 236 Brazilians with asymptomatic infections	Silva et al. (2010)	Genotyped for 593 SNPs in 56 genes across the type 1 interferon (IFN) response pathway as well as other important candidate genes

5.1.4. IFITM restriction of dengue virus *in vitro*

Interferon transmembrane proteins IFITM1,2 and 3 have been identified as antiviral mediators that confer resistance to a number of viruses, including dengue 2 viruses(53, 73). Brass *et al*. established that siRNA depletion of IFITM3 protein led to an increase of replication of DENV serotype 2 virus (New Guinea C strain)(53). These results were later replicated in another study that also observed a similar level of restriction (around 80%) to DENV serotype 2 (New Guinea C strain) in A549 cells expressing IFITM3 proteins(73). Investigations on the mechanisms of IFITM3 antiviral effects have reported that the role of *IFITM3* is not limited to the intercellular space but can also be observed across cells(247). Interestingly, another study revealed that IFITM proteins restrict antibody-dependent enhancement (ADE) infection as efficiently as direct infection(248). Zhu *et. al*, investigated the propagation of *IFITM3* antiviral activity upon dengue infection via exosomes. Exosomes are small vesicles (30-100nm) that are of endocytic origin which have been associated with cell-to-cell transmission of HIV(247). This study demonstrated the presence of IFITM-containing exosome in the extracellular environment. Furthermore, they identified that exosomes delivered IFITM3 to non-infected cells; thereby propagating its antiviral effect(247).

Despite the number of open questions that remain regarding the mechanisms of antiviral restriction of dengue and other virus infections, it appears that IFITM proteins may be good therapeutic targets. If their extracellular antiviral function with regards to dengue are replicated, this would provide great potential for antiviral drug development in the future. Given the incredible burden of dengue worldwide, it is important to understand the role of IFITM proteins and host factors in the context of dengue virus (DENV) infections. I hypothesise that *IFITM3* is one of a number of dengue restriction factors involved in the restriction of these viruses and this constitutes the main motivation for this chapter.

I used genotype data for the *IFITM* locus (11:280,000-380,000) from the largest genetic case-control study on dengue susceptibility in a cohort of 2008 children diagnosed with DSS and 2018 control individuals(113). I had detailed knowledge from Chapter 2 of this thesis, of the coverage of *IFITM* genes in Illumina 660W Beadchip, the array used in the dengue GWAS I found that the coverage for Asian populations (as reported in Chapter 2) was the following: *IFITM2* =12% and *IFITM3* = 2% (IFITM1 was not included due to low number of common SNPs within the gene region). In addition, we had evidence from Chapter 3 that there may be regions near the *IFITM* locus that are not captured or do not have correct genotypes in the 1000 Genome Phase 3 panel. To bridge the gap between the scarce number of directly genotype SNPs included in Illumina Illumina 660W array in this locus, I proceeded to use the data to impute from a number of reference panels, including a deep-sequenced Vietnamese panel that I constructed from lymphoblastoid cell lines. This enabled me to impute and test for association using a greater number of SNPs in the *IFITM* genes.

5.2. Aims

The aim of this project was to identify whether specific mutations in *IFITM* genes contribute to dengue haemorrhagic fever. In order to do this, I use genotype data across a cohort of 2,008 Vietnamese children diagnosed with dengue haemorrhagic fever (DHF) and 2,018 cord blood controls, to test if common variants are associated with the disease.

5.3. Material and Methods

5.3.1. GWAS Genotype dataset

5.3.1.1. Vietnamese genotype data from Dengue GWAS by Khor et. al. 2011

I accessed a section of genotype data (chr11:0-650,000) from the largest GWAS study of Dengue Shock Syndrome carried out so far(113). This dataset contained 579 directly genotyped SNPs from 2008 cases of Vietnamese children with dengue shock syndrome and 2018 controls cord blood controls. In the original GWAS, randomised samples from cases and controls were genotyped using Illumina 660W Quad BeadChips. The QC steps for these samples were performed by Khor and colleagues(113). As a result of these QC steps, samples with sex discordant information, high relatedness IBD scores, per-sample call rates of less than 95% were excluded. Markers with high missing rates or MAF < 1% were also excluded from downstream analysis(113).

5.3.1.2. Constructing a Vietnamese reference panel to impute into the Vietnam genotype data

5.3.1.2.1. Access to Vietnamese samples DNA from 1000 Genomes

I selected the Human variation DNA panel (catalogue number MGP00014) from 100 unrelated Kinh individuals in Ho Chi Minh City, Vietnam from the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository at the Coriell Cell Repositories. I processed 100 DNA samples containing 2μg of DNA, normalised to a concentration of 100ng/μl. These samples were originally used in the 1000 Genomes Sequencing Project.

5.3.1.2.2. Illumina Hiseq sequencing and variant QC

This study employed the Agilent SureSelect targeted sequencing method to pulldown a region in chr11:280,000-380,000, described as in Chapter 4. All the sample processing and sequencing was performed at The Wellcome Trust Sanger Institute. Genomic DNA was sheared to an average was 500bp using Covaris E210 (Covaris Massachusetts, USA). Sheared samples were used in the Illumina library preparation and enriched for the *IFITM* locus using SureSelect Agilent probes (Agilent technologies, Santa Clara, USA ELID number 0798441). Samples were sequenced using the HiSeq 2500 (Illumina, SanDiego, USA) as paired-end 300bp reads. I sequenced 50 samples per lane in duplicates.

5.3.1.2.3. Sequencing analysis

Sequencing analysis was carried out as in previous Chapters. I performed the alignment of the raw sequencing data to the to the human reference genome build GRCh37 using the Burrows-Wheeler Aligner (BWA-mem)(216) and marked duplicates using Picard (http://broadinstitute.github.io/picard, version 2.7.2). Because the same samples were sequenced in two different lanes, I merged bam files belonging to the same individuals using SAMtools(217) and followed the GATK(218) best practice guidelines(166, 167) for bam improvement prior to variant calling. The steps included duplicate marking (http://broadinstitute.github.io/picard, version 2.7.2) and base quality score recalibration (GATK 3.6). Variant calling was performed at the single sample level using the Haplotype Caller (GATK 3.6) and then joint-called using GATK CombineVCF and GenotypeVCF (GATK 3.6). One sample failed sequencing QC and a further two samples failed the variant calling step. One sample was excluded from the 1000 Genomes final dataset and subsequently, I also excluded it from my panel. This left a total of 96 samples for downstream analysis.

For variant QC, I applied GATK hardfiltering recommendations(167) as in Chapter 3 and 4 (Figure 41). The hardfiltering parameters included were: QualbyDepth (QD), Fisher Strand (FS), StrandOddsRatio (SOR), RMSMapping Quality (MQ), MappingQualityRankSumTest (MQRankSum) and ReadPosRankSum. After filtering 909 variants were retained.

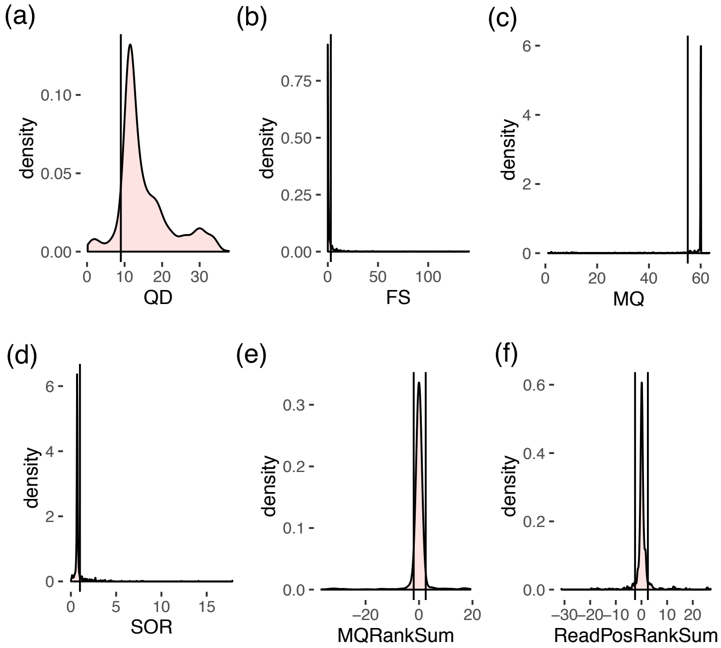


Figure 41. **Density plots of QC parameters used to filter out bad quality SNPs.** On the x-axis are the annotation values used to assess the quality of the dataset, and on the y-axis, are the density values. The black line represents the threshold values used to filter variants **(a)**The distribution of the dataset shows Quality by Depth (QD) values from 0-30, with two peaks representing heterozygous reads (QD =13) and homozygous calls with double the number of reads (QD =28). Variants with QD values < 9 represent low quality calls and were excluded from the dataset. **(b)** Variants have Fisher Strand (FS) values close to zero showing no strand bias. The black lines show the cut off value FS >1 used to exclude variants that deviate from zero. **(c)** Mapping Quality (MQ) values are close to 60 and of the highest quality. Any variants with MQ <55 were excluded from the analysis. **(d)** The majority of the dataset had Strand Odds Ratio (SOR) values close to zero. To exclude variants with some degree of strand bias, SOR values > 3 were used as cut off values. **(e)** Variants with MQRankSum < -3 (more reads supporting reference alleles) and MQRankSum > +3 **(**more reads supporting alternate alleles**)** were also excluded in the analysis. **(f)** Finally, the dataset showed that the distribution of the Read Position Rank Sum (ReadPosRankSum) is close to zero indicating that there is little difference between the positions of the reference and alternate alleles within the reads. Any variants deviating from zero were also excluded from the analysis.

5.3.1.2.4. Variant Annotation

Following variant calling and QC, I annotated variants using dbSNP v137. I used Ensembl Variant Effect predictor (Ensembl variation release 76) for annotation and used the most severe predicted consequence for each gene variant. I used Scoring Intolerance from Tolerance (SIFT)(220) and Polyphen-2(221) predictions to determine whether the variants were likely to affect protein function, and the Sequence Ontology terms and description in Ensembl to score variants for their functional impact on the function of the protein

5.3.1.2.5. Identification of Individuals with elevated missing genotype rate or outlying heterozygosity

I followed the data quality in case-control association studies by Anderson, *et al*. 2011 to identify individuals with elevated missing genotype rates (Figure 42), followed by identification of markers with excessive missing genotype rate using PLINK 1.9. In the first instance, I excluded a total of five samples due to high heterozygosity rate (Figure 42) leaving me with the final total number of 91 individuals in the Vietnamese panel. For the second step of the QC, I excluded 28 variants out of 909 total number of variants after imposing a cut off threshold of 2% on the dataset (Figure 43).

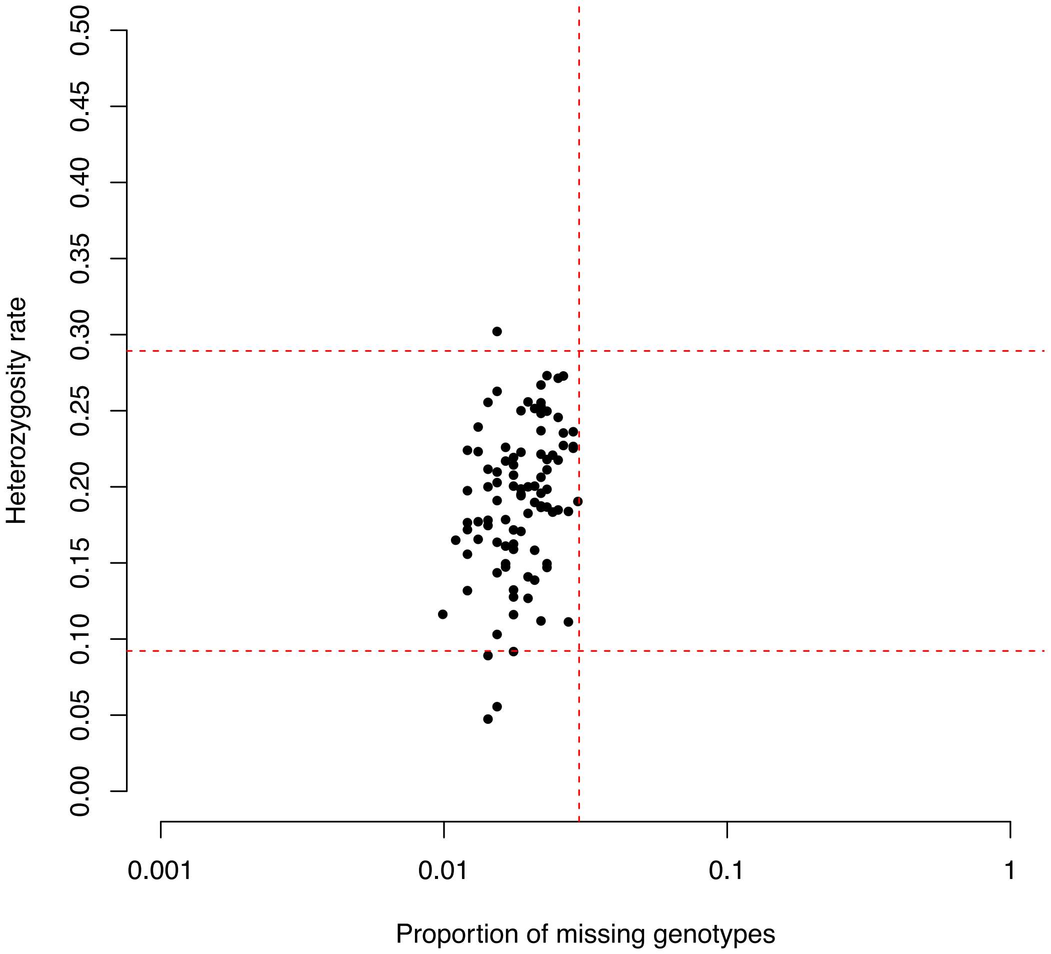


Figure 42. **Identification of individuals with high missing genotype and outlying heterozygosity rates.** Each black dot represents an individual sample and the dashed red lines represent the QC threshold imposed on a dataset. Individuals with more than 2% of missing genotypes were excluded. In addition, individuals with too high (>29%) or too low (<9.0%) values of heterozygosity were also excluded from the analysis.

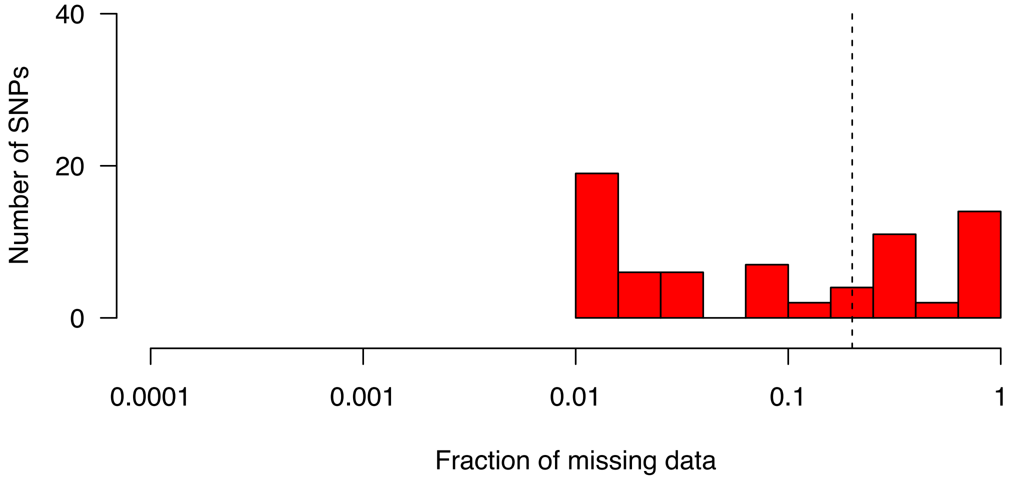


Figure 43. **Histogram of the number of variants with missing genotype data**. The dashed vertical line represents the QC threshold (13%) used to filter out SNPs with high missing rate. In total, 28 SNPs were removed out of 909 SNPs.

5.3.1.2.6. Phasing the Vietnamese sequencing data using SHAPEIT2

In order to construct the reference panel, I proceeded to phase the QC'ed sequencing data using SHAPEIT2(249),(250). There are three phasing strategies currently listed by SHAPEIT2. Because, it was unknown which specific strategy would work best on my data, I phased with all the three methods and evaluated according to switch errors and Info metrics. In the following section, I give more details on the different phasing strategies that I employed:

phasing with a reference panel scaffold, phasing without a reference and Read Aware Phasing (PIRS). I then proceeded to evaluate the phasing output by comparing the switch error rate and the flip errors between outputs.

1. Phasing the Vietnamese samples using 1000 Genomes Phase 3 Reference Panel as scaffold

   I first downloaded 1000 Genomes Reference panel from https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html and proceeded to compare variants called in my dataset with variants in the

1000 Genome dataset. I excluded all sites that were highlighted as different or missing from the 1000 Genomes Phase 3 Reference panel. The excluded sites were comprised of 126 sites missing in the 1000 Genomes Phase 3 reference panel and a further 36 misaligned sites within the panels. Once the data was pruned, I then phased with SHAPEIT2 using standard phasing options as described in the SHAPEIT2 best practices. I had a total of 736 sites after phasing with 1 Genome dataset.

2. Phasing without a scaffold

   It has been documented that phasing without a scaffold gives less accurate results. However, I also employed this strategy to use as comparison tool for imputation accuracy. I had a total of 881 sites after phasing.

3. Phasing using "Phase Informative Reads (PIRS)"(251)

   Next-generation sequencing data produces reads that may contain phase information when spanning two or more heterozygous sites. Pirs constitutes an extension of SHAPEIT2 method to infer haplotypes from genotype data. In this method, Delaneau and colleagues make used of base quality scores within reads to feed into a probabilistic model of haplotype estimation(251). Because the method was originally designed for high coverage dataset, I reasoned that it would be suitable for the Vietnamese dataset. Furthermore, phasing with 1000 Genomes Phase 3 reference panel (see explanation above) resulted in over 100 sites excluded for further analysis. Using the read-aware strategy avoids the exclusion of some of these sites because as long as the base quality for a particular site is high (>30) in the bam files, the sites would be included in the final haplotype estimation output. This left me with a total of 891 sites.

The initial metrics I used to assess the phasing accuracy of the three phasing strategies were switch error and imputation quality(252). I used an 'in-house' script' written by Tommy Carstensen, staff scientists at the Wellcome Trust Sanger Institute. The switch error strategy generally requires a gold standard phased dataset. Because I did not have a gold standard phased dataset for the Vietnamese samples, I carried out several combinations of comparisons between phased haplotypes from all three phased panels. I found that phasing with the 1000 Genomes as a scaffold resulted in 1,127 error rates; without a scaffold, 1,066 and with PIRs, 575. The results were similar but suggested that PIRS contained the least number of switch error rates. This is consistent with previous findings that found using PIRS in real data from 1000 Genomes dataset would reduce the switch error rate substantially(251)

5.3.1.2.7. Imputation using IMPUTE2

To impute untyped SNPs from reference panels I used the default imputation commands. As suggested in the IMPUTE2 guidelines, I excluded variants with MAF < 0.01 as these SNPs are expected to provide the least power for association studies and increase the error rate, they are often also harder to impute and may affect imputation quality for the rest of the sites.

To impute using the Sanger Imputation server, I followed the set of instructions provided in the Sanger imputation website. The first step was to ensure that the data was in the correct format. I used BCFtools (1.3.1-htslib-1.3.2) to convert the genotype gen file to vcf files. I then annotated the chromosomes to match Ensembl-style chromosome names also with BCFtools and I ensured that sites in the genotype data matched the sites in GRCh37 reference fasta. Finally, I converted the data from Illumina TOP convention to forward reference strand with BCFtools fixref plugin and validated my vcf using vcf-validator using VCFtools (v 1.3.1-htslib-1.3.2).

5.4. Results

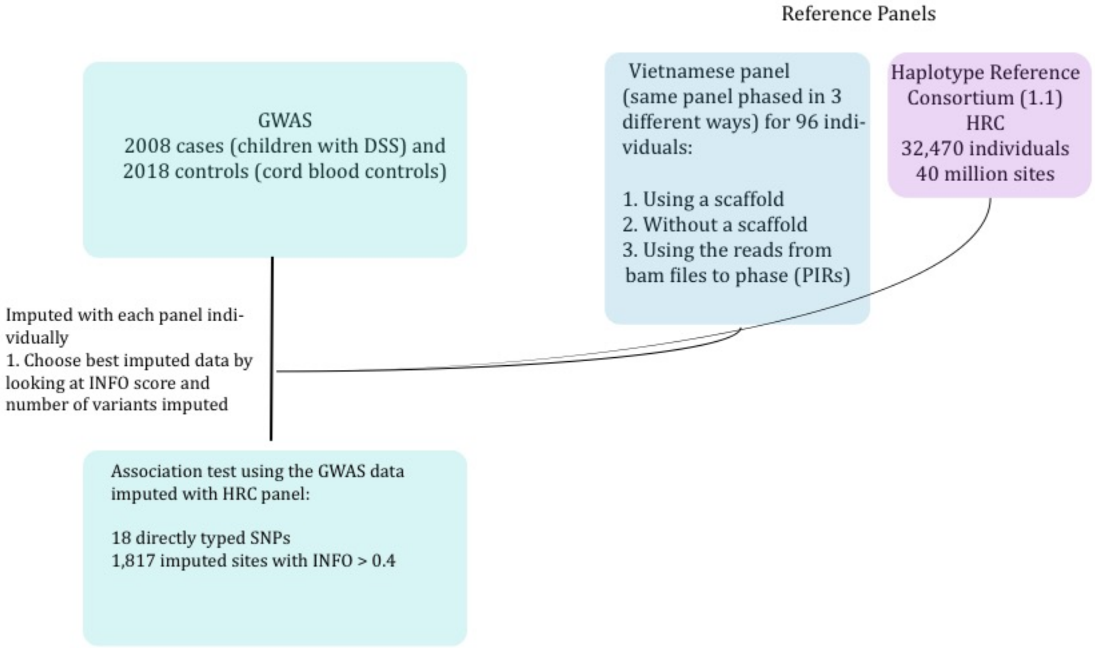
5.4.1. Imputation of GWAS genotype data using Vietnamese sequencing panel

Before association analysis, I imputed the GWAS data(113) with each of the three phased Vietnamese panels that I constructed from the previous step. I then assessed the imputation accuracy for each panel. The output of the imputation resulted in:

a) 718 SNPs were imputed from the panel phased with scaffold.
b) 873 SNPs were imputed from the panel phased without scaffold.
c) 833 SNPs were imputed SNPs from the panel phased with reads from bam files – Pirs

In a second step, I also decided to impute the GWAS Genotype dataset using the HRC reference panels available from the Sanger Imputation Website(86). The output of this imputation resulted in 1,835 imputed sites and 18 sites from directly genotyped SNPs (Figure 44).

To determine the imputation accuracy, I used the INFO score calculated by the imputation software. The INFO score used the input dataset to quantify the relevant statistical information about the variant's MAF in the input dataset. An INFO score = 0.9 represents highly accurate genotype data (Figure 45).



Reference Panels

GWAS
2008 cases (children with DSS) and
2018 controls (cord blood controls)

Vietnamese panel (same panel phased in 3 different ways) for 96 individuals:

1. Using a scaffold
2. Without a scaffold
3. Using the reads from bam files to phase (PIRs)

Haplotype Reference Consortium (1.1)
HRC
32,470 individuals
40 million sites

Imputed with each panel individually
1. Choose best imputed data by looking at INFO score and number of variants imputed

Association test using the GWAS data imputed with HRC panel:

18 directly typed SNPs
1,817 imputed sites with INFO > 0.4

Figure 44. **Summary of the GWAS analysis using imputation**. I constructed a reference panels of 96 Vietnamese individuals. The same reference panel was phased in three different ways and used to imputed the GWAS dataset. In parallel, I used the HRC panel to also impute. After the comparison, The HRC panel came out on top with the highest INFO scores and number of imputed variants.

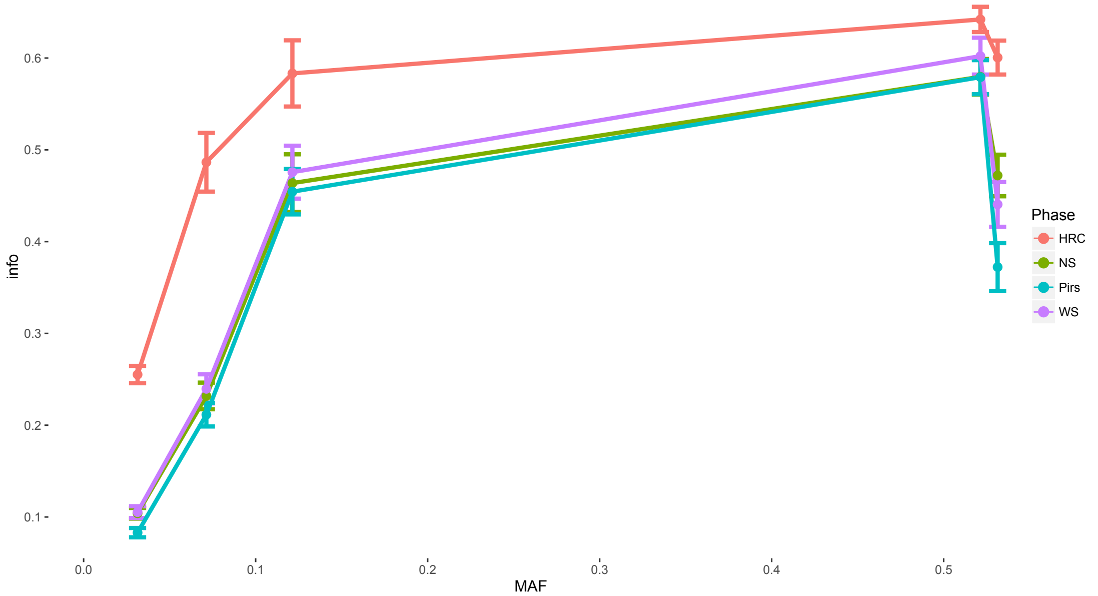


Figure 45. **The INFO quality for imputed variants in dengue cohort**. Variants obtained from each imputation analysis show that the INFO score for variants imputed with the HRC panels (red line) have substantially higher INFO scores than variants imputed with the Vietnamese panel alone.

5.4.3. Power to detect associations

Following QC of SNPs in the severe dengue dataset, 18 directly genotyped and 1,817 imputed SNPs of MAF ≥1% were available for GWAS across the *IFITM* region (chr11:280,000-380,000). With 2,008 cases of Vietnamese children with dengue shock syndrome and 2,018 controls cord blood controls and using a genome-wide significance threshold of P values <5x10^{-8}, this study had >90% power to detect common variants with allele frequencies of at least 5% and moderate and large effect sizes (OR≥2.0) (Figure 46). For low-frequency variants of 1%, 80% power was only achieved for large effect sizes (OR≥3) (Figure 46).

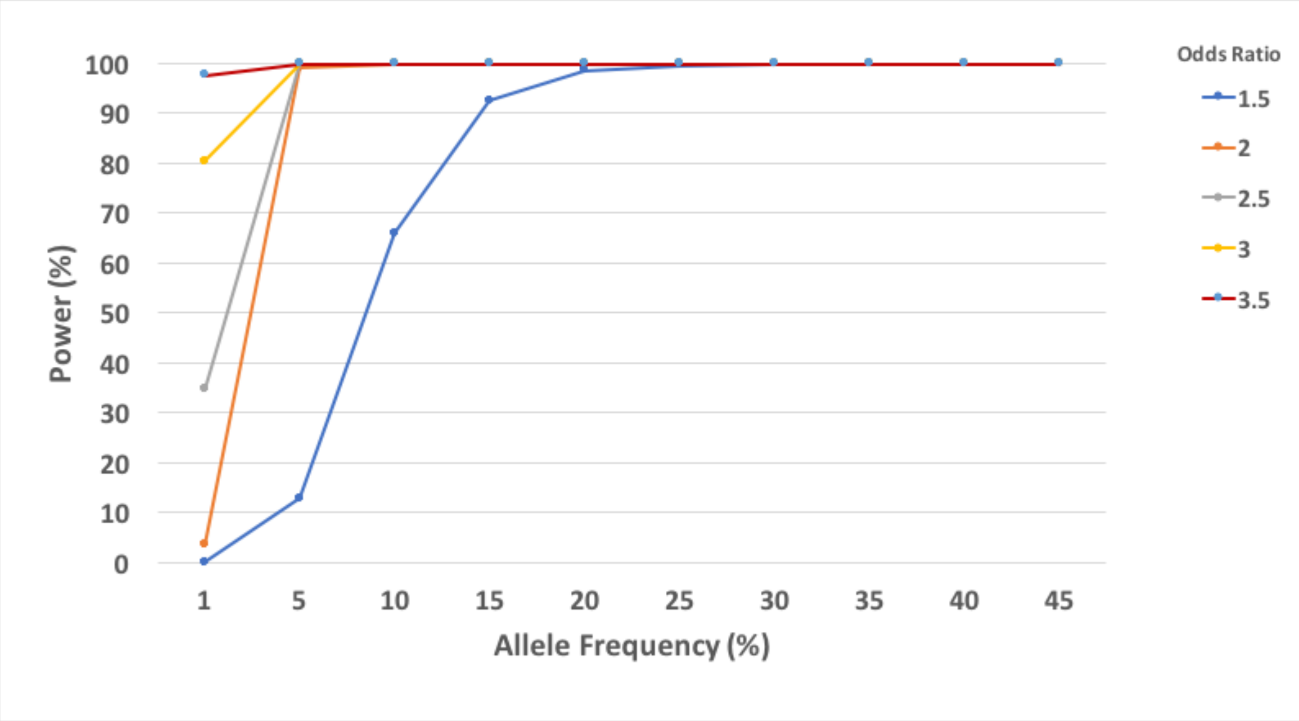


Figure 46. **Statistical power.** The statistical power to identify genetic variants of genome-wide significance and with different effect sizes given the sample size was estimated using QUANTO software (http://biostats.usc.edu/software). These calculations were done by Dr. Neneh Sallah and Fernando Riveros Aguilera at the Wellcome Trust Sanger Institute.

5.4.2. Single variant association test

For genetic association, I used data imputed with the HRC panels. 2,008 individuals with dengue and 2018 controls were available for analysis and 1,836 sites were available for association testing. I tested for association using SNPTEST v2.5, performing an additive frequentist association on variants with INFO≥0.4. I tested 1836 sites for association of which 20 variants showed nominal evidence of significance (P < 0.05) and none showed significant evidence of genome-wide association (5x10^{-8}) (Table 22).

Table 22. Case-control association tests (SNPTEST) for variants in the *IFITM* locus. P-values are not corrected for multiple testing but none reached genome-wide significance (5x10^{-8}).

rsid	chromosome	position	alleleA	alleleB	all maf	cases maf	controls maf	OR	OR lower	OR_upper	P value
rs760060	11	288883	A	G	0.13	0.11	0.14	0.81	0.71	0.93	0.0025
rs10902119	11	288115	T	C	0.12	0.11	0.13	0.82	0.72	0.94	0.0043
rs61876261	11	356080	T	C	0.14	0.16	0.13	1.20	1.06	1.36	0.0043
rs186373656	11	302039	C	T	0.20	0.21	0.19	1.18	1.05	1.31	0.0048
rs11246052	11	286134	G	A	0.10	0.09	0.11	0.82	0.71	0.94	0.0057
rs11246117	11	353630	G	A	0.26	0.24	0.27	0.87	0.79	0.96	0.0065
rs11600194	11	287959	G	A	0.10	0.10	0.11	0.82	0.71	0.95	0.0073
rs78642272	11	351271	G	A	0.17	0.16	0.18	0.86	0.76	0.97	0.0090
rs183397121	11	334170	C	G	0.12	0.13	0.12	1.20	1.05	1.36	0.0094
rs3932433	11	289553	G	C	0.10	0.09	0.11	0.83	0.72	0.96	0.0107
rs3809112	11	307036	T	C	0.31	0.30	0.33	1.13	1.03	1.24	0.0109
rs10902121	11	306791	T	C	0.31	0.30	0.33	1.13	1.03	1.24	0.0126
rs117779161	11	343820	G	A	0.14	0.13	0.14	0.85	0.75	0.97	0.0129
rs11246099	11	347878	T	A	0.17	0.16	0.18	1.16	1.03	1.30	0.0139
rs11246059	11	305961	A	C	0.49	0.50	0.47	1.12	1.02	1.22	0.0145
rs7948108	11	323507	G	A	0.31	0.30	0.33	0.89	0.81	0.98	0.0147
rs11246062	11	313755	C	G	0.45	0.43	0.46	1.12	1.02	1.22	0.0157
11:308178	11	308178	T	C	0.37	0.36	0.38	1.12	1.02	1.22	0.0171
rs10398	11	308180	A	G	0.39	0.38	0.40	1.11	1.02	1.22	0.0188
11:344626	11	344626	T	C	0.02	0.03	0.02	1.42	1.06	1.91	0.0196

5.5. Discussion

Recent reports suggest that IFITM proteins restrict dengue infection with several studies observing 70-80% restriction in cells that overexpress IFITM3(53, 73). As a consequence, I investigated the host genetic contribution of *IFITM* genes to dengue shock syndrome in a Vietnamese cohort comprised of 4,026 individuals.

The design of this study with 2,008 cases and 2,018 controls, using a genome-wide threshold of $P < 5x10^{-8}$, had 100% power to detect common variants with minor allele frequencies of at least 30% with effect sizes (OR > 2.0) (Figure 46). However, despite having the power, I do not find any genome-wide significant associations. This means that at least in this cohort, common variants of *IFITM* genes do not contribute to the genetics of severe dengue haemorrhagic fever. We had very little power (<10%) to detect associations of alleles with modest Odds ratio and MAF (2%). Thus, I cannot rule out that multiple common variants of small effect, population-specific variants or rare variants exist that influence dengue severity.

Another possibility that could impact the findings is the study design. As I mentioned in the introduction, it is well documented that in order to avoid severe symptoms of dengue, a crucial step is the early recognition of signs of dengue haemorrhagic fever. Thus, dengue infected individuals that develop DHS respond well to fluid replacement therapies and recover if medical attention is sought on time. Although there are few detailed epidemiological data in publication with regards to dengue management of infections in Vietnam in recent years, there have been some reports that highlight the poor control of dengue infection by healthcare workers in the country. For example, healthcare provisions were assessed between April 2001 to March 2002 in a cohort of two thousand ninety-six patients. This study found that the diagnostic and therapeutic response of healthcare workers was 'unspecific' and reflected lack of understanding of the disease(253). In rural areas of Vietnam, where healthcare provisions are not adequate, there is generally a reliance on simple tourniquet tests (approved by the WHO to test capillary fragility) for diagnosis of dengue haemorrhagic fever. It has been reported that this test differentiates badly between dengue haemorrhagic fever (45% positive) and dengue fever (38% positive)(254). Probably the best study yet to highlight the challenges of good study design for dengue disease was undertaken Anders KL, *et. al.*, in 2011. This study highlighted the burden of dengue infection in Vietnam by analysing admissions data from over 100,000 patients to three large hospitals in Ho Chi Minh City, Vietnam, for the period of 1996-2009. One key finding was that girls had a significant higher risk to suffer a fatal outcome from DSS than boys. Unsurprisingly, a greater proportion of girls also developed DSS (a complication of DHF). It is interesting to note that girls did not account for the majority of admissions for dengue fever (the mild form of dengue) for the same period. The report goes on to highlight a substantial bias to male admissions to hospital amongst dengue cases which cannot be explained by local demographics (male: female ratio in children is around 108:100) at the time. The authors therefore attributed the differences to as yet unexplained healthcare-seeking behaviour, which saw almost twice as many boys as girls admitted to hospitals during that period for dengue fever symptoms. Although the authors attribute the higher mortality rate observed in females suffering with DSS to biological differences between sexes; they also point out that behavioural factors such as the differences of care provided to girls and boys could also influence the disease outcome(237). The proportion of females in this study is 48.5% between the age of 6-11. Therefore, problems with early diagnosis of dengue haemorrhagic fever and the health-seeking behaviour displayed by the Vietnamese population could affect the severity of the disease. It is an open question whether the same bias of male admissions compared to females impacted the results of this study.

Finally, it has been demonstrated that GWAS continues to have the potential to uncover association of novel loci with modest effect sizes (1.07-1.20) as evident by a recent GWAS of the well-studied inflammatory bowel disease (IBD) (109). In this study, the authors used genotyping to scan the variation of 12,160 IBD cases and 13,145 population controls of European ancestry and test for association. In order to achieve a similar level of success in terms of the number of identified loci with genome-wide levels of significance in dengue, a similar number of cases and controls would be necessary. Due to the rising number of dengue cases worldwide surpassing 300 million, a similar number of cases and controls can be achieved.

6. General Discussion

6.1. Summary of my research

In this dissertation, I described four projects, all with the common purpose of understanding the genetic and phenotypic characteristics of the *IFITM* locus. In Chapter 1, I gave an overview of the biology and function of IFITM proteins, highlighting the role of IFITMs as potent antiviral factors. I also provided details of several candidate gene association studies that have reported correlations between genetic variation in *IFITM* genes and virus susceptibility and disease progression.

Despite all the evidence from *in vitro* and *in vivo* studies demonstrating the important role of *IFITMs* as modulators of restriction, no genome-wide association studies have reported any significant associations to genetic variants in or around these genes. I addressed why this may be the case in Chapter 2, by calculating the coverage of these genes by several commercially available genotyping arrays. I found that less than 25% of the common variation (minor allele frequency > 1%) in *IFITM2* and *IFITM3* is covered on these arrays. Both genes are therefore in the bottom 7% of genes across the genome in terms of coverage, and suggests that poor coverage could explain the lack of genome-wide significant associations in the region. I concluded that other methods such as next generation sequencing would be required to ascertain the full degree of variation in the region.

In Chapter 3, I explored the utility of a conventional targeted sequencing method to detect variation in the *IFITM* locus. As part of this work, I also worked to adapt this method for PacBio library preparations. Conventional pulldown protocols for targeted sequencing have not been designed for single molecule, real-time sequencing platforms such as PacBio *RS*. Although I used a similar approach as that used in a recent publication(151), I developed the technique independently, before this report was published. I found that both methods captured most of the variation in the target region. Although PacBio sequencing marginally outperformed Illumina sequencing in 'difficult' regions near the *IFITM3*, the high cost of reagents and manual effort involved in adapting the technology to work with PacBio *RS*, guided the decision to employ Illumina in future *IFITM* targeted sequencing studies.

In Chapter 4, I applied the conventional targeted sequencing method described in Chapter 3 to test genetic variants in and around *IFITM1, IFITM2* and *IFITM3* for association with the rapid disease progression in HIV. I also explored the burden of low-frequency and rare genetic variants (MAF < 5%) to this phenotype by testing for a differential enrichment between HIV elite controllers and rapid progressors across each of the three genes. Because of the limited sample size in this study, I also decided to analyse other larger cohorts of infectious disease cases and controls to test for significant associations to the *IFITM* genes. In Chapter 5, I described the association analysis in a cohort of 2,008 Vietnamese children diagnosed with dengue haemorrhagic fever (DHF) and 2,018 cord blood controls. To increase the number of variants to test for association, I also constructed an *IFITM* imputation panel by deep-sequencing the locus in 100 Vietnamese individuals from the 1000 Genomes Consortium. I evaluated the use of these haplotypes with other currently accessible imputation panels such as the Human Reference Consortium (HRC) and the 1000 Genomes Phase 3 panels. I found that the HRC panel outperformed all other panels tested.

In the following pages, I will discuss general lessons learnt during my PhD research and what solutions can be adopted to address some of the limitations presented by studying complex infectious diseases.

6.2. Candidate gene studies and population stratification

As I mentioned in the introduction, most candidate gene studies for *IFITM* genes have concentrated on one SNP rs12252 situated at the 5' of the *IFITM3* gene. The evidence of association between this SNP and diseases such as HIV and influenza have often been conflicting(54, 91, 93, 255). Crucially, to my knowledge, none of these studies have accounted for population stratification. With a variant such as the rs12252 SNP, with MAF = 0.5 in Asian populations and MAF = 0.02 in European populations, accounting for population stratification will be crucial. Indeed, the study reported in Chapter 3 highlights the potential for population stratification to be a major source of false-positives associations when using a candidate gene approach. The project started with 126 rapid progressors and 99 elite controllers, all of alleged European descent. Because approximately 94% of these samples had been whole-exome sequenced in previous studies, I was able to test for population stratification using these data. As detailed previously, a large number of samples had to be excluded (n=66) for further analysis, the result of which was a significant reduction in power to detect true associations but a decreased false-positive rate due to population stratification.

Confounding factors such as population admixture and population stratification, are thought to be major factors contributing to the lack of replication of candidate gene studies(84). Indeed, there are two classic examples where genetic admixture and population stratification have led to reports of false associations. One of the most widely known examples comes from a study that reported as association between the HLA haplotype and diabetes mellitus in individuals on a Pima Indian reservation. Once this analysis was restricted to full-heritage Pima-Papago Indians, the association disappeared(84). Similarly, a metanalysis of association studies between alcoholism and dopamine D2 receptor established that most likely all previously reported associations were the result of population stratification due to the reliance on self-reported ancestries for most studies(256). Although I am aware that targeted sequencing studies are the most feasible and cost-effective strategies to many scientists who want to capture all genetic variation within a locus, I believe that such studies should be complemented with other similarly affordable technologies such as genome-wide genotyping, which would ensure that appropriate quality control steps can be adopted.

6.3. Genome-wide association studies in current times

Genome-wide association studies continue to provide the scientific community with invaluable information regarding the genetic architecture of complex diseases and their underlying biology. One example that highlights the important role of genome-wide studies is the UK Biobank's initiative to whole-exome sequence and genotype approximately 500,000 UK individuals using a custom design array carrying 820,967 markers. Importantly, over 10% of these markers had been chosen because of previous known genetic associations or possible roles in phenotypic variation for a number of diseases, such as neurological diseases, cancer and inflammatory or autoimmune disorders. Crucially, the UK Biobank is making a considerable effort to link participants' health-records from General Practice (GP) centres to their genotype information, thus providing scientists with relevant epidemiological information that would prove invaluable to the understanding of disease.

Despite the relevance of genome-wide scans in large initiatives such as UK Biobank, as I demonstrated for *IFITM* genes in Chapter 2, there are regions of the genome that are still not represented in genotype arrays. Furthermore, the substantial population sampling bias in genome-wide association studies, exemplified by the dire statistics showing that only 4% of GWAS up to 2011 were carried in non-European populations(257), means that there is generally a lack of coverage of variants present in these non-European populations in most genotyping chips. This historical population bias in genotyping arrays is changing and there are currently several genotyping arrays available to the scientific community that capture variation in non-European populations (258).

The coverage is further reduced following typical quality control steps in genome-wide association studies. For example, in a study of Primary Sclerosing Cholangitis (PSC) in 2014, Liu and colleagues genotyped 3,789 PSC cases of European ancestry and 25,079 population controls using the Immunochip. Close examination of their quality control steps revealed that approximately 61,000 variants (out of ~196,000) included in the array, did not pass the quality control assessments generally employed in these studies. Similarly, Kumar and colleagues, obtained only 118,989 (out of ~196,000) SNPs from the 217 candidaemia cases and 11,920 healthy controls after applying standard quality parameters(259). There will be a number of factors that will contribute to the failure of a particular SNP to pass quality controls. It is certainly the case that although quality control steps have been standardised for genome wide association studies, each centre will employ specific threshold that are more appropriate for their data. Importantly, many of the variants included in arrays are selected from publicly available datasets such as the 1000 Genomes. In Chapter 3, I provided a snapshot of the problem when I showed that for a 600bp region at the 5' end of the *IFITM3*, all non-reference calls for 8 sampled SNPs were incorrectly called in that dataset. Although similar situations in the rest of the genome may be rare, when they happen, it could negatively impact the discovery of new associations.

6.4. Samples size in infectious diseases.

Sample size is a critical parameter in any scientific experiment where a hypothesis is being tested. In this dissertation, low sample sizes had a negative influence on the results reported in Chapter 3. Albeit one study, all of the reported candidate studies of *IFITM* so far have been carried out with very low samples sizes

(88, 260). To ascertain the role of genetic variation *IFITM* locus in virus susceptibility and prognosis, greater emphasis must be placed on the recruitment of larger cohorts of individuals. Although there are several examples of excellent genetic studies in infectious diseases, especially in HIV(99, 102, 202), and tuberculosis(105), there is a generally a lack of well-powered (with adequately large samples sizes) studies being reported in the field. By definition, cases of rare phenotypes such HIV rapid progressors and elite controllers, are very difficult to obtain(99). As a consequence, many scientists have called for greater worldwide collaborations that could help provide access to larger study cohorts(99). HIV/AIDS is responsible for approximately 65 deaths for every 100,000 individuals in low-income countries(261). As a result, future collaborations in those countries will be key to the understanding of the architecture of disease. Indeed, there are important initiatives set up by the Wellcome Trust to try to facilitate partnerships between the United Kingdom and several countries in the African continent. For example, the KEMRI-Wellcome Trust Research Programme (KWTRP) in Kenya, has been, for the past 25 years, focusing their research in areas such as malnutrition, HIV/AIDS and respiratory diseases. Another important example is the Wellcome Trust MRC Unit in the Gambia that also focuses in eradicating diseases such as tuberculosis and HIV in the country. More recently, other collaborations have been formed to facilitate the data analysis of African populations. The African Genome Project has provided dense genotypes from 1,481 individuals and whole-genome sequencing data for a further 320 individuals across sub-Sahara Africa(262), thus becoming an invaluable resource for scientists who are interested in studying genetics in these populations(262). Ultimately, the hope is that initiatives such as these would not only increase our understanding of disease, but it would help build the infrastructure of those countries and contribute towards the formation of future scientists in the region.

Generally, significantly increasing the sample size has led to a greater number of novel association being discovered, although this is not always the case. For example, a major study of schizophrenia in 36,989 cases and 113,075 controls identified 128 independent associations, 83 of which were novel(263). In contrast, a study of major depressive disorder (MDD) in 22,158 cases and 133,749 controls identified only one significant association: rs7647854 ($p = 5.2 \times 10^{-11}$)(264). The study of MDD has been challenging due to a moderate heritability (20%) and the heterogeneity of the genetic factors that contribute to the condition(265). Although we hope that infectious diseases will not be as challenging for GWAS as MDD is, what these studies suggest is that other approaches will be needed alongside genome-wide studies to understand the genetic architecture of disease.

6.5. Future study designs for complex infectious diseases

6.5.1. Host and pathogen interactions

Most genetic studies of infectious diseases to date are limited to either the study of the pathogen or the host. However, extrapolating from a recent study by Bartha and colleagues, it is apparent that the advantages of studying the genome of the pathogen alongside the genome of the infected individual will result in an increase in the power to detect significant associations. For example, Bartha and colleagues, with only 1,071 individuals infected with HIV were able to report a number of significant associations between human single nucleotide variants and viral amino acid sites. Their most significant association was between human rs72845950 and virus Nef position 135 ($p = 2.7 \times 10^{-66}$). Interestingly, the authors also discovered an association between the SNP rs2395029 which acts as a proxy for HLA-B*57:01 and amino acid in Gag at position 242 (an escape position from HLA-B*57:01). Previous association studies of HIV control using viral load also detected the same SNP but with a weaker association (1.21×10^{-6}). Ultimately, this study has great implications for the development of treatments and vaccines and most importantly, it is an approach that can be employed in the future to determine the co-evolution of other viruses that are also under great pressure from the host, such as influenza virus.

6.5.2. The relevance of co-infections

Despite some evidence on the effect of co-infections on disease severity and progression, there are relatively few examples of studies that have explored these complex pathogens' interactions in the host. A study in Malawi several years ago established that the viral load in HIV positive individuals increased with the severity of malaria, potentially increasing the likelihood of HIV transmission(266). Furthermore, mathematical modelling trying to estimate the effect of malaria on HIV transmission established that interactions between both diseases resulted in approximately 8,500 HIV infections and 980,000 malaria infections in the region since 1980(267). Although challenging, future genetic studies on co-infection would provide greater understanding of disease and how concurrent infections may affect treatment response and disease susceptibility to other infections.

6.5.3. The future of next-generation targeted enrichment strategies

Finally, in the past few years, the rapid decline in the cost of large scale sequencing has resulted in tremendous advancement in the field of human genetics and clinical research. Targeted sequencing continues to provide an affordable and effective strategy to target specific regions of the genome, however, they often require several days for completion due to the number of laboratory steps involved. Currently, next generation targeted sequencing methods rely on a separate step for hybridisation where biotinylated DNA or RNA capture probes are hybridised with the target DNA. Recently, an amplification-free Single Molecule Targeted Sequencing (SMTS) method has been proposed to directly target and sequence sample molecules without the need for PCR amplification(268). If successful, this technique would result in reduced library preparation time and greater sensitivity for capturing mutations(268). This would be particularly useful for clinicians and diagnostic companies that aim to provide accurate and fast sequencing results.

6.6. Concluding remarks

The Human Genome Project demonstrated the power of collaborative processes and data sharing and this dogma still stands today. I believe that better collaborations especially with countries with the greatest burden of disease could provide the scientific community with invaluable insights into the architecture of infectious diseases in future. Furthermore, initiatives such as GSK's Open Targets in collaboration with Biogen, the European Bioinformatics Institute and the Wellcome Trust Sanger Institute provide an invaluable resource of freely available genetic and biological data for scientists across the globe. It is certainly the case that we are currently experiencing a revolution in science and I believe it is a privilege to be a part of it.

Appendix A

In Chapter3, I mentioned that the method I developed to sequence with PacBio *RS* has been used to sequence chicken cell lines. The picture below illustrates the coverage for the PacBio method, compared to the Illumina method

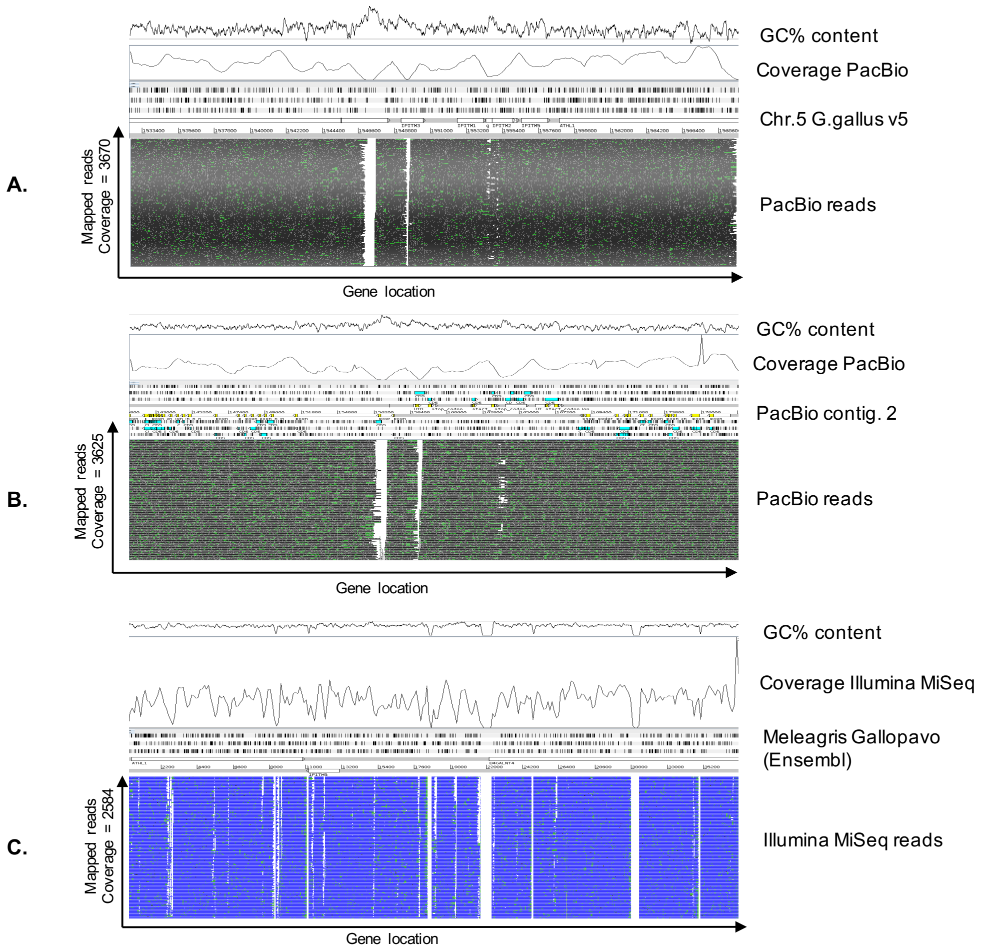


Figure 47. **Artemis coverage and stack view of the IFITM locus in DF1 cells following pull down of the IFITM locus using SureSelect probes and sequencing with PacBio**. The figure shows an intact locus and successful mapping of the IFITM locus. B. Artemis coverage and stack view of the IFITM locus in turkey breast tissue following pull down of the IFITM locus using SureSelect probes and sequencing with Illumina MiSeq. The graph shows successful mapping of MiSeq reads despite using chicken probes to pull down the locus in turkey tissue. The white bars represent actual gaps in the turkey reference as published on both Ensemble and NCBI and to which the probes will not eventually map as gaps are shown in the reference as “NNN”. From Dr. Irene Bassano, personal communication.

References

1. Kumar H, Kawai T, Akira S. Pathogen recognition in the innate immune response. The Biochemical journal. 2009;420(1):1-16.

2. Kauffmann SaK, D. Introduction: The immune response to infectious agents. In: Kauffman SHEaKD, editor. Methods in Microbiology. 251998. p. 4.

3. Mogensen TH. Pathogen Recognition and Inflammatory Signaling in Innate Immune Defenses. Clinical microbiology reviews. 222009. p. 240-73.

4. Takeuchi O, Akira S. Pattern recognition receptors and inflammation. Cell. 2010;140(6):805-20.

5. Gay NJ, Symmons MF, Gangloff M, Bryant CE. Assembly and localization of Toll-like receptor signalling complexes. Nature Reviews Immunology. 2014;14:546-58.

6. Yin Q, Fu T-M, Li J, Wu H. Structural Biology of Innate Immunity. http://dxdoiorg/101146/annurev-immunol-032414-112258. 2015.

7. Alexopoulou L, Holt AC, Medzhitov R, Flavell RA. Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3. Nature. 2001;413(6857):732-8.

8. Negishi H, Osawa T, Ogami K, Ouyang X, Sakaguchi S, Koshiba R, et al. A critical link between Toll-like receptor 3 and type II interferon signaling pathways in antiviral innate immunity. 2008.

9. Hardarson HS, Baker JS, Yang Z, Purevjav E, Huang CH, Alexopoulou L, et al. Toll-like receptor 3 is an essential component of the innate stress response in virus-induced cardiac injury. American journal of physiology Heart and circulatory physiology. 2007;292(1):H251-8.

10. Oshiumi H, Okamoto M, Fujii K, Kawanishi T, Matsumoto M, Koike S, et al. The TLR3/TICAM-1 pathway is mandatory for innate immune responses to poliovirus infection. Journal of immunology (Baltimore, Md : 1950). 2011;187(10):5320-7.

11. Oshiumi H, Matsumoto M, Funami K, Akazawa T, Seya T. TICAM-1, an adaptor molecule that participates in Toll-like receptor 3-mediated interferon-beta induction. Nature immunology. 2003;4(2):161-7.

12. Matsumoto M, Funami K, Tanabe M, Oshiumi H, Shingai M, Seto Y, et al. Subcellular localization of Toll-like receptor 3 in human dendritic cells. Journal of immunology (Baltimore, Md : 1950). 2003;171(6):3154-62.

13. Gowen BB, Hoopes JD, Wong M-H, Jung K-H, Isakson KC, Alexopoulou L, et al. TLR3 Deletion Limits Mortality and Disease Severity due to Phlebovirus Infection. 2006.

14. Le Goffic R, Balloy V, Lagranderie M, Alexopoulou L, Escriou N, Flavell R, et al. Detrimental contribution of the Toll-like receptor (TLR)3 to influenza A virus-induced acute pneumonia. PLoS Pathog. 2006;2(6):e53.

15. Krug A, French AR, Barchet W, Fischer JA, Dzionek A, Pingel JT, et al. TLR9-dependent recognition of MCMV by IPC and DC generates coordinated cytokine responses that activate antiviral NK cell function. Immunity. 2004;21(1):107-19.

16. Monteiro JT, Lepenies B. Myeloid C-Type Lectin Receptors in Viral Recognition and Antiviral Immunity. Viruses. 2017;9(3).

17. Geijtenbeek TB, van Kooyk Y. DC-SIGN: a novel HIV receptor on DCs that mediates HIV-1 transmission. Current topics in microbiology and immunology. 2003;276:31-54.

18. Gringhuis SI, van der Vlist M, van den Berg LM, den Dunnen J, Litjens M, Geijtenbeek TB. HIV-1 exploits innate signaling by TLR8 and DC-SIGN for productive infection of dendritic cells. Nature immunology. 2010;11(5):419-26.

19. Jensen S, Thomsen AR. Sensing of RNA Viruses: a Review of Innate Immune Receptors Involved in Recognizing RNA Virus Invasion. 2012.

20. Sabbah A, Chang TH, Harnack R, Frohlich V, Tominaga K, Dube PH, et al. Activation of innate immune antiviral response by NOD2. Nature immunology. 2009;10(10):1073-80.

21. Barlan AU, Griffin TM, McGuire KA, Wiethoff CM. Adenovirus membrane penetration activates the NLRP3 inflammasome. Journal of virology. 2011;85(1):146-55.

22. Wang X, Jiang W, Yan Y, Gong T, Han J, Tian Z, et al. RNA viruses promote activation of the NLRP3 inflammasome through a RIP1-RIP3-DRP1 signaling pathway. Nature immunology. 2014;15:1126-33.

23. Allen IC, Scull MA, Moore CB, Holl EK, McElvania-TeKippe E, Taxman DJ, et al. The NLRP3 Inflammasome Mediates in vivo Innate Immunity to Influenza A Virus through Recognition of Viral RNA. Immunity. 2009;30(4):556-65.

24. Brubaker SW, Bonham KS, Zanoni I, Kagan JC. Innate Immune Pattern Recognition: A Cell Biological Perspective. http://dxdoiorg/101146/annurev-immunol-032414-112240. 2015.

25. Seth RB, Sun L, Ea CK, Chen ZJ. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. Cell. 2005;122(5):669-82.

26. Hornung V, Ellegast J, Kim S, Brzózka K, Jung A, Kato H, et al. 5'-Triphosphate RNA Is the Ligand for RIG-I. 2006.

27. Lee YF, Nomoto A, Detjen BM, Wimmer E. A protein covalently linked to poliovirus genome RNA. 1977.

28. Flint S.J. ELW, Racaniello V. R., and Skalka, A. M. Principles of Virology, 3rd Ed.: American Society for Microbiology Press; 2009.

29. Fensterl V, Chattopadhyay S, Sen GC. No Love Lost Between Viruses and Interferons. http://dxdoiorg/101146/annurev-virology-100114-055249. 2015.

30. Isaacs A, Lindenmann J. Virus Interference. I. The Interferon. 1957.

31. Ryman KD, Klimstra WB, Nguyen KB, Biron CA, Johnston RE. Alpha/beta interferon protects adult mice from fatal Sindbis virus infection and is an important determinant of cell and tissue tropism. Journal of virology. 2000;74(7):3366-78.

32. Platanias LC. Mechanisms of type-I- and type-II-interferon-mediated signalling. Nature Reviews Immunology. 2005;5(5):375-86.

33. Dalton D, Pitts-Meek S, Keshav S, Figari I, Bradley A, Stewart T. Multiple defects of immune cell function in mice with disrupted interferon-gamma genes. 1993.

34. Jouanguy E, Altare F, Lamhamedi S, Revy P, Emile J-F, Newport M, et al. Interferon-γ –Receptor Deficiency in an Infant with Fatal Bacille Calmette–Guérin Infection. http://dxdoiorg/101056/NEJM199612263352604. 2009.

35. Newport MJ, Huxley CM, Huston S, Hawrylowicz CM, Oostra BA, Williamson R, et al. A Mutation in the Interferon-γ –Receptor Gene and Susceptibility to Mycobacterial Infection. http://dxdoiorg/101056/NEJM199612263352602. 2009.
36. Bellanti F, Vendemiale G, Altomare E, Serviddio G. The impact of interferon lambda 3 gene polymorphism on natural course and treatment of hepatitis C. Clinical & developmental immunology. 2012;2012:849373.
37. Ge D, Fellay J, Thompson AJ, Simon JS, Shianna KV, Urban TJ, et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature. 2009;461(7262):399-401.
38. Rauch A, Kutalik Z, Descombes P, Cai T, Di Iulio J, Mueller T, et al. Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. Gastroenterology. 2010;138(4):1338-45, 45.e1-7.
39. Prokunina-Olsson L, Muchmore B, Tang W, Pfeiffer RM, Park H, Dickensheets H, et al. A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus. Nature Genetics. 2013;45:164-71.
40. O'Brien TR, Pfeiffer RM, Paquin A, Lang Kuhs KA, Chen S, Bonkovsky HL, et al. Comparison of functional variants in IFNL4 and IFNL3 for association with HCV clearance. Journal of hepatology. 2015;63(5):1103-10.
41. Hamming OJ, Terczyńska-Dyla E, Vieyres G, Dijkman R, Jørgensen SE, Akhtar H, et al. Interferon lambda 4 signals via the IFNλ receptor to regulate antiviral activity against HCV and coronaviruses. 2013.
42. Lauber C, Vieyres G, Terczyńska-Dyla E, Anggakusuma, Dijkman R, Gad HH, et al. Transcriptome analysis reveals a classical interferon signature induced by IFN|[lambda]|4 in human primary cells. Genes and Immunity. 2015;16(6):414-21.
43. Noureddin M, Rotman Y, Zhang F, Park H, Rehermann B, Thomas E, et al. Hepatic expression levels of interferons and interferon-stimulated genes in patients with chronic hepatitis C: A phenotype|[ndash]|genotype correlation study. Genes and Immunity. 2015;16(5):321-9.
44. Sung PS, Hong S-H, Chung J-H, Kim S, Park S-H, Kim HM, et al. IFN-λ4 potently blocks IFN-α signalling by ISG15 and USP18 in hepatitis C virus infection. Scientific Reports. 2017;7(1):3821.
45. Schneider WM, Chevillotte MD, Rice CM. Interferon-Stimulated Genes: A Complex Web of Host Defenses. http://dxdoiorg/101146/annurev-immunol-032713-120231. 2014.
46. Zhang X, Bogunovic D, Payelle-Brogard B, Francois-Newton V, Speer SD, Yuan C, et al. Human intracellular ISG15 prevents interferon-[agr]/[bgr] over-amplification and auto-inflammation. Nature. 2014;517:89-93.
47. Flint S.J. ELW, Racaniello V. R., and Skalka, A. M. Principles of Virology 3rd edition. Washington DC, USA: ASM Press; 2009.
48. Mashimo T, Lucas M, Simon-Chazottes D, Frenkiel MP, Montagutelli X, Ceccaldi PE, et al. A nonsense mutation in the gene encoding 2'-5'-oligoadenylate synthetase/L1 isoform is associated with West Nile virus susceptibility in laboratory mice. Proceedings of the National Academy of Sciences of the United States of America. 2002;99(17):11311-6.
49. Yoshii K, Moritoh K, Nagata N, Yokozawa K, Sakai M, Sasaki N, et al. Susceptibility to flavivirus-specific antiviral response of Oas1b affects the

neurovirulence of the Far-Eastern subtype of tick-borne encephalitis virus. Archives of virology. 2013;158(5):1039-46.

50. Silverman RH. Viral Encounters with 2′,5′-Oligoadenylate Synthetase and RNase L during the Interferon Antiviral Response. 2007.

51. Lin R-J, Yu H-P, Chang B-L, Tang W-C, Liao C-L, Lin Y-L. Distinct Antiviral Roles for Human 2′,5′-Oligoadenylate Synthetase Family Members against Dengue Virus Infection. 2009.

52. Yakub I, Lillibridge KM, Moran A, Gonzalez OY, Belmont J, Gibbs RA, et al. Single nucleotide polymorphisms in genes for 2'-5'-oligoadenylate synthetase and RNase L inpatients hospitalized with West Nile virus infection. J Infect Dis. 2005;192(10):1741-8.

53. Brass AL, Huang IC, Benita Y, John SP, Krishnan MN, Feeley EM, et al. IFITM Proteins Mediate the Innate Immune Response to Influenza A H1N1 Virus, West Nile Virus and Dengue Virus. Cell. 2009;139(7):1243-54.

54. Everitt AR, Clare S, Pertel T, John SP, Wash RS, Smith SE, et al. IFITM3 restricts the morbidity and mortality associated with influenza. Nature. 2012;484(7395):519-23.

55. Ifitm3 Limits the Severity of Acute Influenza in Mice. 2017.

56. Lange UC, Adams DJ, Lee C, Barton S, Schneider R, Bradley A, et al. Normal Germ Line Establishment in Mice Carrying a Deletion of the Ifitm/Fragilis Gene Family Cluster. 2008.

57. Hickford D, Frankenberg S, Shaw G, Renfree MB. Evolution of vertebrate interferon inducible transmembrane proteins. BMC genomics. 2012;13(1):155.

58. Ling S, Zhang C, Wang W, Cai X, Yu L, Wu F, et al. Combined approaches of EPR and NMR illustrate only one transmembrane helix in the human IFITM3. Sci Rep. 62016.

59. Xu Y, Yang G, Hu G. Binding of IFITM1 enhances the inhibiting effect of caveolin-1 on ERK activation. Acta biochimica et biophysica Sinica. 2009;41(6):488-94.

60. Yang G, Xu Y, Chen X, Hu G. IFITM1 plays an essential role in the antiproliferative action of interferon-|[gamma]|. Oncogene. 2006;26(4):594-603.

61. A Membrane Topology Model for Human Interferon Inducible Transmembrane Protein 1. 2017.

62. Lu J, Pan Q, Rong L, Liu S-L, Liang C. The IFITM Proteins Inhibit HIV-1 Infection. 2011.

63. Bailey CC, Kondur HR, Huang I-C, Farzan M. Interferon-Induced Transmembrane Protein 3 is a Type II Transmembrane Protein. 2013.

64. Zhu H, Liu C. Interleukin-1 Inhibits Hepatitis C Virus Subgenomic RNA Replication by Activation of Extracellular Regulated Kinase Pathway. Journal of virology. 772003. p. 5493-8.

65. Wilkins C, Woodward J, Lau DT, Barnes A, Joyce M, McFarlane N, et al. IFITM1 is a tight junction protein that inhibits hepatitis C virus entry. Hepatology (Baltimore, Md). 2013;57(2):461-9.

66. Narayana SK, Helbig KJ, McCartney EM, Eyre NS, Bull RA, Eltahla A, et al. The Interferon-induced Transmembrane Proteins, IFITM1, IFITM2, and IFITM3 Inhibit Hepatitis C Virus Entry. The Journal of biological chemistry. 2015;290(43):25946-59.

67. Takahashi S, Doss C, Levy S, Levy R. TAPA-1, the target of an antiproliferative antibody, is associated on the cell surface with the Leu-13 antigen. 1990.

68. Lindenbach BD, Rice CM. The ins and outs of hepatitis C virus entry and assembly. Nature Reviews Microbiology. 2013;11:688-700.

69. Weston S, Czieso S, White IJ, Smith SE, Wash RS, Diaz-Soria C, et al. Alphavirus Restriction by IFITM Proteins. Traffic (Copenhagen, Denmark). 2016;17(9):997-1013.

70. Feeley EM, Sims JS, John SP, Chin CR, Pertel T, Chen LM, et al. IFITM3 inhibits influenza A virus infection by preventing cytosolic entry. PLoS Pathog. 2011;7(10):e1002337.

71. Mudhasani R, Tran JP, Retterer C, Radoshitzky SR, Kota KP, Altamura LA, et al. IFITM-2 and IFITM-3 but not IFITM-1 restrict Rift Valley fever virus. Journal of virology. 2013;87(15):8451-64.

72. IFITM Proteins Restrict HIV-1 Infection by Antagonizing the Envelope Glycoprotein: Cell Reports. 2017.

73. John SP, Chin CR, Perreira JM, Feeley EM, Aker AM, Savidis G, et al. The CD225 domain of IFITM3 is required for both IFITM protein association and inhibition of influenza A virus and dengue virus replication. Journal of virology. 2013;87(14):7837-52.

74. Huang IC, Bailey CC, Weyer JL, Radoshitzky SR, Becker MM, Chiang JJ, et al. Distinct patterns of IFITM-mediated restriction of filoviruses, SARS coronavirus, and influenza A virus. PLoS Pathog. 2011;7(1):e1001258.

75. Wee YS, Roundy KM, Weis JJ, Weis JH. Interferon-inducible transmembrane proteins of the innate immune response act as membrane organizers by influencing clathrin and v-ATPase localization and function. http://dxdoiorg/101177/1753425912443392. 2012.

76. Diamond MS, Farzan M. The broad-spectrum antiviral functions of IFIT and IFITM proteins. Nature Reviews Immunology. 2012;13(1):46-57.

77. Amphotericin B Increases Influenza A Virus Infection by Preventing IFITM3-Mediated Restriction.

78. Qian J, Le Duff Y, Wang Y, Pan Q, Ding S, Zheng YM, et al. Primate lentiviruses are differentially inhibited by interferon-induced transmembrane proteins. Virology. 2015;474:10-8.

79. Amini-Bavil-Olyaee S, Choi YJ, Lee JH, Shi M, Huang I-C, Farzan M, et al. The Antiviral Effector IFITM3 Disrupts Intracellular Cholesterol Homeostasis to Block Viral Entry. Cell host & microbe. 2013;13(4):452-64.

80. Mehle A, Strack B, Ancuta P, Zhang C, McPike M, Gabuzda D. Vif Overcomes the Innate Antiviral Activity of APOBEC3G by Promoting Its Degradation in the Ubiquitin-Proteasome Pathway. 2004.

81. Antagonism of Tetherin Restriction of HIV-1 Release by Vpu Involves Binding and Sequestration of the Restriction Factor in a Perinuclear Compartment. 2017.

82. Burgner D, Jamieson SE, Blackwell JM. Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? Lancet Infect Dis. 62006. p. 653-63.

83. CN H, RG J, Herndon CN, Jennings RG. A twin-family study of susceptibility to poliomyelitis. 1951.

84. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet (London, England). 2003;361(9357):598-604.

85. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061-73.

86. Consortium tHR. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics. 2016;48:1279-83.

87. Wang Z, Zhang A, Wan Y, Liu X, Qiu C, Xi X, et al. Early hypercytokinemia is associated with interferon-induced transmembrane protein-3 dysfunction and predictive of fatal H7N9 infection. 2014.

88. Zhang Y-H, Zhao Y, Li N, Peng Y-C, Giannoulatou E, Jin R-H, et al. Interferon-induced transmembrane protein-3 genetic variant rs12252-C is associated with severe influenza in Chinese individuals. Nature Communications. 2013;4:1418.

89. Zhang Y, Makvandi-Nejad S, Qin L, Zhao Y, Zhang T, Wang L, et al. Interferon-induced transmembrane protein-3 rs12252-C is associated with rapid progression of acute HIV-1 infection in Chinese MSM cohort. AIDS (London, England). 2015;29(8):889-94.

90. Xu-yang Z, Pei-yu B, Chuan-tao Y, Wei Y, Hong-wei M, Kang T, et al. Interferon-Induced Transmembrane Protein 3 Inhibits Hantaan Virus Infection, and Its Single Nucleotide Polymorphism rs12252 Influences the Severity of Hemorrhagic Fever with Renal Syndrome. Front Immunol. 2016;7.

91. Mills TC, Rautanen A, Elliott KS, Parks T, Naranbhai V, Ieven MM, et al. IFITM3 and susceptibility to respiratory viral infections in the community. J Infect Dis. 2014;209(7):1028-31.

92. Gaio V, Nunes B, Pechirra P, Conde P, Guiomar R, Dias CM, et al. Hospitalization Risk Due to Respiratory Illness Associated with Genetic Variation at IFITM3 in Patients with Influenza A(H1N1)pdm09 Infection: A Case-Control Study. PLoS One. 2016;11(6):e0158181.

93. Lopez-Rodriguez M, Herrera-Ramos E, Sole-Violan J, Ruiz-Hernandez JJ, Borderias L, Horcajada JP, et al. IFITM3 and severe influenza virus infection. No evidence of genetic association. European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology. 2016;35(11):1811-7.

94. Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, et al. Genetic Restriction of HIV-1 Infection and Progression to AIDS by a Deletion Allele of the CKR5 Structural Gene. 1996.

95. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, et al. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell. 1996;86(3):367-77.

96. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber CM, et al. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature. 1996;382(6593):722-5.

97. Husman A-MdR, Koot M, Cornelissen M, Keet IPM, Brouwer M, Broersen SM, et al. Association between CCR5 Genotype and the Clinical Course of HIV-1 Infection. Annals of internal medicine. 2017;127(10):882-90.

98. Migueles SA, Sabbaghian MS, Shupert WL, Bettinotti MP, Marincola FM, Martino L, et al. HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. 2000.

99. McLaren PJ, Coulonges C, Bartha I, Lenz TL, Deutsch AJ, Bashirova A, et al. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. 2015.

100. Syvänen A-C. Toward genome-wide SNP genotyping. Nature Genetics. 2005;37.

101. Consortium TIH. A haplotype map of the human genome. Nature. 2005;437(7063):1299-320.

102. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, et al. A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. 2007.

103. Common Genetic Variation and the Control of HIV-1 in Humans. 2017.

104. Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PI, Walker BD, et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science (New York, NY). 2010;330(6010):1551-7.

105. Curtis J, Luo Y, Zenner HL, Cuchet-Lourenço D, Wu C, Lo K, et al. Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. Nature Genetics. 2015;47:523-7.

106. Jallow M, Teo YY, Small KS, Rockett KA, Deloukas P, Clark TG, et al. Genome-wide and fine-resolution association analysis of malaria in West Africa. Nature Genetics. 2009;41(6):657-65.

107. Allison AC. Protection Afforded by Sickle-cell Trait Against Subtertian Malarial Infection. Br Med J. 1954;1(4857):290-4.

108. JL T, GW N, JA L, L C, C M, RC J, et al. Genome-wide Association Study Implicates PARD3B-based AIDS Restriction. 2011.

109. Lange KMd, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nature Genetics. 2017;49:256-61.

110. Luo Y, Lange KMd, Jostins L, Moutsianas L, Randall J, Kennedy NA, et al. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. Nature Genetics. 2017;49:186-92.

111. Hill AVS. Evolution, revolution and heresy in the genetics of infectious disease susceptibility. 2012.

112. Zhang F-R, Huang W, Chen S-M, Sun L-D, Liu H, Li Y, et al. Genomewide Association Study of Leprosy. http://dxdoiorg/101056/NEJMoa0903753. 2010.

113. Khor CC, Chau TN, Pang J, Davila S, Long HT, Ong RT, et al. Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. Nat Genet. 2011;43(11):1139-41.

114. Ha NT, Freytag S, Bickeboeller H. Coverage and efficiency in current SNP chips. European journal of human genetics : EJHG. 2014;22(9):1124-30.

115. Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet. 2006;38(6):659-62.

116. Li M, Li C, Guan W. Evaluation of coverage variation of SNP chips for genome-wide association studies. European journal of human genetics : EJHG. 2008;16(5):635-43.

117. Risch N, Merikangas K. The Future of Genetic Studies of Complex Human Diseases. 1996.

118. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The International HapMap Project. Nature. 2003;426(6968):789-96.
119. T. Heim, L.-C. Tranchevent, E. Carlon, †,‡ and, Barkema§ GT. Physical-Chemistry-Based Analysis of Affymetrix Microarray Data. 2006.
120. Consortium TGP. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56-65.
121. Consortium TGP. A global reference for human genetic variation. Nature. 2015;526:68-74.
122. K W, JL M, J H, L C, Y M, S M, et al. The UK10K project identifies rare variants in health and disease. 2015.
123. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics. 2010;11(7):499-511.
124. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS genetics. 2012;8(8):e1002793.
125. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics. 2011;43:1193-201.
126. Parkes M, Cortes A, Heel DAv, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nature Reviews Genetics. 2013;14:661-73.
127. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, et al. A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. Science (New York, NY). 2007;317(5840):944-7.
128. McLaren PJ, Carrington M. The impact of host genetic variation on infection with HIV-1. Nature immunology. 2015;16:577-83.
129. Ji S-G, Juran BD, Mucha S, Folseraas T, Jostins L, Melum E, et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. Nature Genetics. 2016.
130. Suppiah V, Moldovan M, Ahlenstiel G, Berg T, Weltman M, Abate ML, et al. IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. Nat Genet. 2009;41(10):1100-4.
131. Browning BL, b.browning@auckland.ac.nz, Department of Statistics UoA, Auckland 1142, New Zealand, Browning SR, Department of Statistics UoA, Auckland 1142, New Zealand. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics. 2009;84(2):210-23.
132. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. Nat Protoc. 2010;5(9):1564-73.
133. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. 2017.
134. Yu Z, Schaid DJ. Methods to impute missing genotypes for population data. Human genetics. 2007;122(5):495-504.
135. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. Human molecular genetics. 2012;21(R1):R1-9.

136. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biology. 2009;10(3).

137. Beck TF, Mullikin JC, Program obotNCS, Biesecker LG. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. 2016.

138. Asan, Xu Y, Jiang H, Tyler-Smith C, Xue Y, Jiang T, et al. Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biol. 2011;12(9):R95.

139. D'Alessandro LC, Al Turki S, Manickaraj AK, Manase D, Mulder BJ, Bergin L, et al. Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect. Genetics in medicine : official journal of the American College of Medical Genetics. 2016;18(2):189-98.

140. Chilamakuri CS, Lorenz S, Madoui MA, Vodak D, Sun J, Hovig E, et al. Performance comparison of four exome capture systems for deep sequencing. BMC genomics. 2014;15:449.

141. Singh T, Kurki MI, Curtis D, Purcell SM, Crooks L, McRae J, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nature Neuroscience. 2016;19:571-7.

142. Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, Hannula S, et al. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. 2011;12(9):R94.

143. Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, et al. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome research. 2010;20(10):1420-31.

144. Mertes F, ElSharawy A, Sauer S, van Helvoort JM, van der Zaag P, Franke A, et al. Targeted enrichment of genomic DNA regions for next-generation sequencing. Brief Funct Genomics. 102011. p. 374-86.

145. Grozeva D, Carss K, Spasic-Boskovic O, Tejada MI, Gecz J, Shaw M, et al. Targeted Next-Generation Sequencing Analysis of 1,000 Individuals with Intellectual Disability. Human mutation. 2015;36(12):1197-204.

146. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. Nature Methods. 2007;4(11):903-5.

147. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, et al. Target-enrichment strategies for next-generation sequencing. Nature Methods. 2010;7(2):111-8.

148. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology. 2009;27(2):182-9.

149. Meienberg J, Center for Cardiovascular Genetics and Gene Diagnostics FfPwRD, Schlieren-Zurich CH-8952, Switzerland, Zerjavic K, Center for Cardiovascular Genetics and Gene Diagnostics FfPwRD, Schlieren-Zurich CH-8952, Switzerland, Keller I, Department of Clinical Research UoB, Berne CH-3010, Switzerland, et al. New insights into the performance of human whole-exome capture platforms. Nucleic acids research. 2017;43(11).

150. Parla JS, Iossifov I, Grabill I, Spector MS, Kramer M, McCombie WR. A comparative analysis of exome capture. Genome Biol. 2011;12(9):R97.

151. Wang M, Beck CR, English AC, Meng Q, Buhay C, Han Y, et al. PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. BMC genomics. 2015;16:214.

152. Witek K, Jupe F, Witek AI, Baker D, Clark MD, Jones JDG. Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. Nature Biotechnology. 2016;34:656-60.

153. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2011;13(1):36-46.

154. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2014;517:608-11.

155. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;17:333-51.

156. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.

157. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics. 2012;13(1):341.

158. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic acids research. 2008;36(16):e105.

159. Complete annotated genome sequence of Mycobacterium tuberculosis (Zopf) Lehmann and Neumann (ATCC35812) (Kurono) - Tuberculosis. 2017.

160. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biology. 2013;14(5).

161. Rutledge GG, Böhme U, Sanders M, Reid AJ, Cotton JA, Maiga-Ascofare O, et al. Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution. Nature. 2017.

162. PacBio Sequencing and Its Applications. 2015;13(5):278–89.

163. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC genomics. 2012;13(1):375.

164. IFITM Proteins Restrict Antibody-Dependent Enhancement of Dengue Virus Infection. 2017.

165. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics (Oxford, England). 252009. p. 1754-60.

166. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011;43:491-8.

167. Van der Auwera GA, Massachusetts GSaAGBIC, Carneiro MO, Massachusetts GSaAGBIC, Hartl C, Massachusetts GSaAGBIC, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. 2017:11.0.1-.0.33.

168. Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, et al. Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. Genome research. 2013;23(1):121-8.

169. Clark MJ, Chen R, Lam HY, Karczewski KJ, Euskirchen G, Butte AJ, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol. 2011;29(10):908-14.
170. Li H, Medical Population Genetics Program BIoHaM, Cambridge, MA 02142, USA. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics (Oxford, England). 2017;30(20):2843-51.
171. Dapprich J, Ferriola D, Mackiewicz K, Clark PM, Rappaport E, D'Arcy M, et al. The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. BMC genomics. 172016.
172. Osmond DH, Page K, Wiley J, Garrett K, Sheppard HW, Moss AR, et al. HIV infection in homosexual and bisexual men 18 to 29 years of age: the San Francisco Young Men's Health Study. http://dxdoiorg/102105/AJPH84121933. 2011.
173. Nelson KE, knelson3@jhu.edu, Health BSoP, Vlahov D, dvlahov1@jhu.edu, Galai N, et al. Preparations for AIDS vaccine trials. Incident human immunodeficiency virus (HIV) infections in a cohort of injection drug users in Baltimore, Maryland. AIDS Research and Human Retroviruses. 2017;10(SUPPL. 2).
174. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature. 1999;397(6718):436-41.
175. Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH, Johnson PR. An African primate lentivirus (SIVsmclosely related to HIV-2. Nature. 1989;339(6223):389-92.
176. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science (New York, NY). 2014;346(6205):56-61.
177. Alkhatib G, Combadiere C, Broder CC, Feng Y, Kennedy PE, Murphy PM, et al. CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. Science (New York, NY). 1996;272(5270):1955-8.
178. Feng Y, Broder CC, Kennedy PE, Berger EA. HIV-1 Entry Cofactor: Functional cDNA Cloning of a Seven-Transmembrane, G Protein-Coupled Receptor. 1996.
179. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, et al. Identification of a major co-receptor for primary isolates of HIV-1. Nature. 1996;381(6584):661-6.
180. Dragic T, Litwin V, Allaway GP, Martin SR, Huang Y, Nagashima KA, et al. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. Nature. 1996;381(6584):667-73.
181. Wilen CB, Tilton JC, Doms RW. HIV: Cell Binding and Entry. 2012.
182. Kahn JO, Walker BD. Acute Human Immunodeficiency Virus Type 1 Infection. http://dxdoiorg/101056/NEJM199807023390107. 2009.
183. Mellors JW, Rinaldo CR, Jr., Gupta P, White RM, Todd JA, Kingsley LA. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science (New York, NY). 1996;272(5265):1167-70.
184. Koup RA, Safrit JT, Cao Y, Andrews CA, McLeod G, Borkowsky W, et al. Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. Journal of virology. 1994;68(7):4650-5.

185. Borrow P, Lewicki H, Wei X, Horwitz MS, Peffer N, Meyers H, et al. Antiviral pressure exerted by HIV-l-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nature Medicine. 1997;3(2):205-11.
186. Kalia V, Sarkar S, Ahmed R. CD8 T-Cell Memory Differentiation during Acute and Chronic Viral Infections. 2013.
187. The origin of genetic diversity in HIV-1. 2012;169(2):415–29.
188. Hazenberg MD, Otto SA, van Benthem BH, Roos MT, Coutinho RA, Lange JM, et al. Persistent immune activation in HIV-1 infection is associated with progression to AIDS. AIDS (London, England). 2003;17(13):1881-8.
189. Timm J, Lauer GM, Kavanagh DG, Sheridan I, Kim AY, Lucas M, et al. CD8 Epitope Escape and Reversion in Acute HCV Infection. The Journal of experimental medicine. 2002004. p. 1593-604.
190. Leslie AJ, Pfafferott KJ, Chetty P, Draenert R, Addo MM, Feeney M, et al. HIV evolution: CTL escape mutation and reversion after transmission. Nature Medicine. 2004;10(3):282-9.
191. Hess C, Altfeld M, Thomas SY, Addo MM, Rosenberg ES, Allen TM, et al. HIV-1 specific CD8+ T cells with an effector phenotype and control of viral replication. Lancet (London, England). 2004;363(9412):863-6.
192. Baarle Dv, d_van_baarle@clb.nl, Dept of Clinical Viro-Immunology SRaCLLotAMC, University of Amsterdam, Amsterdam, The Netherlands, Kostense S, Dept of Clinical Viro-Immunology SRaCLLotAMC, University of Amsterdam, Amsterdam, The Netherlands, Oers MHJv, et al. Failing immune control as a result of impaired CD8+ T-cell maturation: CD27 might provide a clue. Trends in Immunology. 2002;23(12):586-91.
193. Gurdasani D, Iles L, Dillon DG, Young EH, Olson AD, Naranbhai V, et al. A systematic review of definitions of extreme phenotypes of HIV control and progression. AIDS (London, England). 2014;28(2):149-62.
194. Grabar S, Selinger-Leneman H, Abgrall S, Pialoux G, Weiss L, Costagliola D. Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the French Hospital Database on HIV. AIDS (London, England). 2009;23(9):1163-9.
195. Martina BE, Koraka P, Osterhaus AD. Dengue virus pathogenesis: an integrated view. Clinical microbiology reviews. 2009;22(4):564-81.
196. Rhodes DI, Ashton L, Solomon A, Carr A, Cooper D, Kaldor J, et al. Characterization of Three nef-Defective Human Immunodeficiency Virus Type 1 Strains Associated with Long-Term Nonprogression. Journal of virology. 742000. p. 10581-8.
197. Wang B, Mikhail M, Dyer WB, Zaunders JJ, Kelleher AD, Saksena NK. First demonstration of a lack of viral sequence evolution in a nonprogressor, defining replication-incompetent HIV-1 infection. Virology. 2003;312(1):135-50.
198. Blankson JN, Bailey JR, Thayil S, Yang HC, Lassen K, Lai J, et al. Isolation and Characterization of Replication-Competent Human Immunodeficiency Virus Type 1 from a Subset of Elite Suppressors▿. Journal of virology. 812007. p. 2508-18.
199. Miura T, Brockman MA, Brumme CJ, Brumme ZL, Carlson JM, Pereyra F, et al. Genetic Characterization of Human Immunodeficiency Virus Type 1 in Elite Controllers: Lack of Gross Genetic Defects or Common Amino Acid Changes. 2008.

200. Pereyra F, Partners AIDS Research Center MGHaDoA, Harvard Medical School, Boston, Massachusetts, Brigham and Women's Hospital DoID, Boston, Massachusetts, Addo MM, Partners AIDS Research Center MGHaDoA, Harvard Medical School, Boston, Massachusetts, Kaufmann DE, et al. Genetic and Immunologic Heterogeneity among Persons Who Control HIV Infection in the Absence of Therapy. The Journal of Infectious Diseases. 2017;197(4):563-71.

201. Lambotte O, Boufassa F, Madec Y, Nguyen A, Goujard C, Meyer L, et al. HIV controllers: a homogeneous group of HIV-1-infected patients with spontaneous control of viral replication. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2005;41(7):1053-6.

202. Pelak K, Center for Human Genome Variation DUMS, Durham, North Carolina, Goldstein DB, Center for Human Genome Variation DUMS, Durham, North Carolina, Walley NM, Center for Human Genome Variation DUMS, Durham, North Carolina, et al. Host Determinants of HIV-1 Control in African Americans. The Journal of Infectious Diseases. 2017;201(8):1141-9.

203. Julg B, Pereyra F, Buzon MJ, Piechocka-Trocha A, Clark MJ, Baker BM, et al. Infrequent recovery of HIV from but robust exogenous infection of activated CD4(+) T cells in HIV elite controllers. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2010;51(2):233-8.

204. Hütter G, Nowak D, Mossner M, Ganepola S, Müßig A, Allers K, et al. Long-Term Control of HIV by CCR5 Delta32/Delta32 Stem-Cell Transplantation. http://dxdoiorg/101056/NEJMoa0802905. 2009.

205. Grivel JC, Margolis LB. CCR5- and CXCR4-tropic HIV-1 are equally cytopathic for their T-cell targets in human lymphoid tissue. Nat Med. 1999;5(3):344-6.

206. Waters L, St. Stephens AIDS Trust DoGHM, Chelsea and Westminster Hospital, London, United Kingdom, Mandalia S, St. Stephens AIDS Trust DoGHM, Chelsea and Westminster Hospital, London, United Kingdom, Randell P, St. Stephens AIDS Trust DoGHM, Chelsea and Westminster Hospital, London, United Kingdom, et al. The Impact of HIV Tropism on Decreases in CD4 Cell Count, Clinical Progression, and Subsequent Response to a First Antiretroviral Therapy Regimen. Clinical Infectious Diseases. 2017;46(10):1617-23.

207. Goetz MB, Leduc R, Kostman JR, Labriola AM, Lie Y, Weidler J, et al. Relationship between HIV Co-Receptor Tropism and Disease Progression in Persons with Untreated Chronic HIV Infection. J Acquir Immune Defic Syndr. 2009;50(3):259-66.

208. Abel K, Rocke DM, Chohan B, Fritts L, Miller CJ. Temporal and anatomic relationship between virus replication and cytokine gene expression after vaginal simian immunodeficiency virus infection. Journal of virology. 2005;79(19):12164-72.

209. Stacey AR, Norris PJ, Qin L, Haygreen EA, Taylor E, Heitman J, et al. Induction of a striking systemic cytokine cascade prior to peak viremia in acute human immunodeficiency virus type 1 infection, in contrast to more modest and delayed responses in acute hepatitis B and C virus infections. Journal of virology. 2009;83(8):3719-33.

210. Sandler NG, Bosinger SE, Estes JD, Zhu RT, Tharp GK, Boritz E, et al. Type I interferon responses in rhesus macaques prevent SIV infection and slow disease progression. Nature. 2014;511(7511):601-5.

211. Resistance of Transmitted Founder HIV-1 to IFITM-Mediated Restriction: Cell Host & Microbe. 2017.
212. Kane M, Yadav SS, Bitzegeio J, Kutluay SB, Zang T, Wilson SJ, et al. MX2 is an interferon-induced inhibitor of HIV-1 infection. Nature. 2013;502:563-6.
213. Doyle T, Goujon C, Malim MH. HIV-1 and interferons: who's interfering with whom? Nature reviews Microbiology. 2015;13(7):403-13.
214. Compton AA, Bruel T, Porrot F, Mallet A, Sachse M, Euvrard M, et al. IFITM proteins incorporated into HIV-1 virions impair viral fusion and spread. Cell host & microbe. 2014;16(6):736-47.
215. Huang X, Lodi S, Fox Z, Li W, Phillips A, Porter K, et al. Rate of CD4 decline and HIV-RNA change following HIV seroconversion in men who have sex with men: a comparison between the Beijing PRIMO and CASCADE cohorts. J Acquir Immune Defic Syndr. 2013;62(4):441-6.
216. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754-60.
217. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078-9.
218. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.
219. McLaren W, European Bioinformatics Institute WTGC, Hinxton, Cambridge, CB10 1SD and 2Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK, Pritchard B, European Bioinformatics Institute WTGC, Hinxton, Cambridge, CB10 1SD and 2Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK, Rios D, European Bioinformatics Institute WTGC, Hinxton, Cambridge, CB10 1SD and 2Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics (Oxford, England). 2017;26(16):2069-70.
220. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols. 2009;4(7):1073-81.
221. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature Methods. 2010;7(4):248-9.
222. Jun G, Flickinger M, Hetrick K, Romm J, Doheny K, Abecasis G, et al. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data. American journal of human genetics. 912012. p. 839-48.
223. Population Structure and Eigenanalysis. 2017.
224. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.
225. Price AL, Weale ME, Patterson N, Myers SR, Need AC, Shianna KV, et al. Long-Range LD Can Confound Genome Scans in Admixed Populations. American journal of human genetics. 832008. p. 132-5.
226. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies.

227. The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. 2017.

228. Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. American journal of human genetics. 892011. p. 82-93.

229. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. American journal of human genetics. 912012. p. 224-37.

230. XUAN Y, University S, WANG LN, University S, LI W, Prevention NMCfDCa, et al. IFITM3 rs12252 T>C polymorphism is associated with the risk of severe influenza: a meta-analysis. Epidemiology & Infection. 2017;143(14):2975-84.

231. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D. Using Extreme Phenotype Sampling to Identify the Rare Causal Variants of Quantitative Traits in Association Studies. Genetic epidemiology. 2011;35(8):790-9.

232. McLaren PJ, Carrington M. The impact of host genetic variation on infection with HIV-1. Nature immunology. 2015;16(6):577-83.

233. Murray NEA, Quam MB, Wilder-Smith A. Epidemiology of dengue: past, present and future prospects. Clin Epidemiol. 52013. p. 299-309.

234. Campagna Dde S, Miagostovich MP, Siqueira MM, Cunha RV. Etiology of exanthema in children in a dengue endemic area. Jornal de pediatria. 2006;82(5):354-8.

235. Guzman MG, Alvarez M, Rodriguez R, Rosario D, Vazquez S, Vald s L, et al. Fatal dengue hemorrhagic fever in Cuba, 1997. International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases. 1999;3(3):130-5.

236. Trung DT, Thao le TT, Hien TT, Hung NT, Vinh NN, Hien PT, et al. Liver involvement associated with dengue infection in adults in Vietnam. The American journal of tropical medicine and hygiene. 2010;83(4):774-80.

237. Anders KL, Nguyet NM, Chau NV, Hung NT, Thuy TT, Lien le B, et al. Epidemiological factors associated with dengue shock syndrome and mortality in hospitalized dengue patients in Ho Chi Minh City, Vietnam. The American journal of tropical medicine and hygiene. 2011;84(1):127-34.

238. Chan M, Johansson MA. The Incubation Periods of Dengue Viruses. PLoS One. 72012.

239. Pouliot SH, Xiong X, Harville E, Paz-Soldan V, Tomashek KM, Breart G, et al. Maternal dengue and pregnancy outcomes: a systematic review. Obstetrical & gynecological survey. 2010;65(2):107-18.

240. Tan PC, Rajasingam G, Devi S, Omar SZ. Dengue infection in pregnancy: prevalence, vertical transmission, and pregnancy outcome. Obstetrics and gynecology. 2008;111(5):1111-7.

241. Kliks SC, Nimmanitya S, Nisalak A, Burke DS. Evidence that maternal dengue antibodies are important in the development of dengue hemorrhagic fever in infants. The American journal of tropical medicine and hygiene. 1988;38(2):411-9.

242. Huang X, Yue Y, Li D, Zhao Y, Qiu L, Chen J, et al. Antibody-dependent enhancement of dengue virus infection inhibits RLR-mediated Type-I IFN-independent signalling through upregulation of cellular autophagy. Scientific Reports, Published online: 29 February 2016; | doi:101038/srep22303. 2016.

243. Boonnak K, Dambach KM, Donofrio GC, Tassaneetrithep B, Marovich MA. Cell type specificity and host genetic polymorphisms influence antibody-dependent enhancement of dengue virus infection. Journal of virology. 2011;85(4):1671-83.

244. Goncalvez AP, Engle RE, St Claire M, Purcell RH, Lai CJ. Monoclonal antibody-mediated enhancement of dengue virus infection in vitro and in vivo and strategies for prevention. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(22):9422-7.

245. Stephens HAF, Institute of Urology and Nephrology UCL, UK, Department of Transfusion Medicine SHaMS, Mahidol University, Bangkok, Thailand, Klaythong R, Department of Transfusion Medicine SHaMS, Mahidol University, Bangkok, Thailand, Sirikong M, et al. HLA-A and -B allele associations with secondary dengue virus infections correlate with disease severity and the infecting viral serotype in ethnic Thais. Tissue Antigens. 2017;60(4):309-18.

246. Thiel S, Frederiksen PD, Jensenius JC. Clinical manifestations of mannan-binding lectin deficiency. Molecular immunology. 2006;43(1-2):86-96.

247. Zhu X, He Z, Yuan J, Wen W, Huang X, Hu Y, et al. IFITM3-containing exosome as a novel mediator for anti-viral response in dengue virus infection. Cellular microbiology. 2015;17(1):105-18.

248. Chan YK, New England Primate Research Center DoMaI, Harvard Medical School, Southborough, Massachusetts, United States of America, Huang I-C, Farzan M, New England Primate Research Center DoMaI, Harvard Medical School, Southborough, Massachusetts, United States of America. IFITM Proteins Restrict Antibody-Dependent Enhancement of Dengue Virus Infection. PLOS ONE. 2012;7(3).

249. Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics. 2008;9(1):540.

250. Delaneau O, Marchini J. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun. 2014;5:3934.

251. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. American journal of human genetics. 2013;93(4):687-96.

252. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nature Reviews Genetics. 2011;12(10):703-14.

253. Phuong HL, de Vries PJ, Nagelkerke N, Giao PT, Hung le Q, Binh TQ, et al. Acute undifferentiated fever in Binh Thuan province, Vietnam: imprecise clinical diagnosis and irrational pharmaco-therapy. Tropical medicine & international health : TM & IH. 2006;11(6):869-79.

254. Cao XT, Ngo TN, Wills B, Kneen R, Nguyen TT, Ta TT, et al. Evaluation of the World Health Organization standard tourniquet test and a modified tourniquet test in the diagnosis of dengue infection in Viet Nam. Tropical medicine & international health : TM & IH. 2002;7(2):125-32.

255. Y Z, S M-N, L Q, Y Z, T Z, L W, et al. Interferon-induced transmembrane protein-3 rs12252-C is associated with rapid progression of acute HIV-1 infection in Chinese MSM cohort. 2015.

256. Smith L, Watson M, Gates S, Ball D, Foxcroft D. Meta-analysis of the association of the Taq1A polymorphism with the risk of alcohol dependency: a

HuGE gene-disease association review. American journal of epidemiology. 2008;167(2):125-38.

257. Bustamante CD, Burchard EG, De La Vega FM. Genomics for the world: Medical genomics has focused almost entirely on those of European descent. Other ethnic groups must be studied to ensure that more people benefit, say. Nature.475(7355):163-5.

258. Johnston HR, Hu Y-J, Gao J, O'Connor TD, Abecasis GR, Wojcik GL, et al. Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. Scientific Reports. 2017;7.

259. Kumar V, Cheng S-C, Johnson MD, Smeekens SP, Wojtowicz A, Giamarellos-Bourboulis E, et al. Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia. Nature Communications, Published online: 8 September 2014; | doi:101038/ncomms5675. 2014.

260. Williams DE, Wu WL, Grotefend CR, Radic V, Chung C, Chung YH, et al. IFITM3 polymorphism rs12252-C restricts influenza A viruses. PLoS One. 2014;9(10):e110096.

261. Nakatani H. Global Strategies for the Prevention and Control of Infectious Diseases and Non-Communicable Diseases. J Epidemiol. 262016. p. 171-8.

262. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2014;517:327-32.

263. Consortium SWGotPG. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421-7.

264. Power RA, Institute of Psychiatry P, and Neuroscience, King's College London, London, Tansey KE, MRC Centre for Neuropsychiatric Genetics and Genomics IoPMaCN, School of Medicine, Cardiff University, Cardiff, United Kingdom, Buttenschøn HN, Lundbeck Foundation Initiative for Integrative Psychiatric Research i, Aarhus University, Aarhus, Denmark, et al. Genome-wide Association for Major Depression Through Age at Onset Stratification: Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Biological Psychiatry. 2017;81(4):325-35.

265. Levinson DF, Mostafavi S, Milaneschi Y, Rivera M, Ripke S, Wray NR, et al. Genetic studies of major depressive disorder: Why are there no GWAS findings, and what can we do about it? Biol Psychiatry. 2014;76(7):510-2.

266. Kublin JG, Patnaik P, Jere CS, Miller WC, Hoffman IF, Chimbiya N, et al. Effect of Plasmodium falciparum malaria on concentration of HIV-1-RNA in the blood of adults in rural Malawi: a prospective cohort study. Lancet (London, England). 2005;365(9455):233-40.

267. Abu-Raddad LJ, Patnaik P, Kublin JG. Dual infection with HIV and malaria fuels the spread of both diseases in sub-Saharan Africa. Science (New York, NY). 2006;314(5805):1603-6.

268. Gao Y, Deng L, Yan Q, Gao Y, Wu Z, Cai J, et al. Single molecule targeted sequencing for cancer gene mutation detection. Scientific Reports, Published online: 19 May 2016; | doi:101038/srep26110. 2016.