Investigation of HIV-TB co-infection through analysis of the potential impact of host genetic variation on host-pathogen protein interactions

Alexa Storme Heekes
HKSALE001
Supervised by: Professor Nicola Mulder

Presented for PhD
Bioinformatics

Faculty of Health Sciences
University of Cape Town

2022

Declaration

I, Alexa Heekes, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I empower the university to reproduce for research either the whole or any portion of the contents in any manner whatsoever.

Signed on 18 January 2022

Abstract

HIV and *Mycobacterium tuberculosis* (*Mtb*) co-infection causes treatment and diagnostic difficulties, which places a major burden on health care systems in settings with high prevalence of both infectious diseases, such as South Africa. Human genetic variation adds further complexity, with variants affecting disease susceptibility and response to treatment. The identification of variants in African populations is affected by reference mapping bias, especially in complex regions like the Major Histocompatibility Complex (MHC), which plays an important role in the immune response to HIV and *Mtb* infection. We used a graph-based approach to identify novel variants in the MHC region within African samples without mapping to the canonical reference genome. We generated a host-pathogen functional interaction network made up of inter- and intraspecies protein interactions, gene expression during co-infection, drug-target interactions, and human genetic variation. Differential expression and network centrality properties were used to prioritise proteins that may be important in co-infection. Using the interaction network we identified 28 human proteins that interact with both pathogens ("bridge" proteins). Network analysis showed that while MHC proteins did not have significantly higher centrality measures than non-MHC proteins, bridge proteins had significantly shorter distance to MHC proteins. Proteins that were significantly differentially expressed during co-infection or contained variants clinically-associated with HIV or TB also had significantly stronger network properties. Finally, we identified common and consequential variants within prioritised proteins that may be clinically-associated with HIV and TB. The integrated network was extensively annotated and stored in a graph database that enables rapid and high throughput prioritisation of sets of genes or variants, facilitates detailed investigations and allows network-based visualisation.

Acknowledgements

I would like to thank my supervisor, Nicola Mulder, for her invaluable guidance, encouragement and support toward the completion of this thesis and other projects in the years before I started. Nicky, thank you especially for your patience in the many months (and even years) where I did not work on my thesis and for somehow managing to keep track of my updates. I would like to thank my boss and mentor, Andrew Boulle, for supporting my academic goals and encouraging me to take time to work on my studies even when it competed with work objectives. I would also like to thank my PHDC colleagues, especially Mariette Smith, Nicki Tiffin, Florence Phelanyane, and Themba Mutemaringa, for all of your guidance and support over the years. Thank you to everyone in the Computational Biology lab for always being willing to help me whenever I needed it, despite my absence from the lab. Thank you to my parents, Allan and Candice Heekes, as well as all my other family members for always supporting me, and believing in me - your confidence in my abilities has always given me strength to keep going. I would like to thank my sisters, Sasha-Lee Heekes and Adiilah Boodhoo, for your unconditional love, encouragement and for being among the few to read my thesis - this would not have been possible without your support. Finally, I would like to thank my husband and best friend, Husain Boodhoo, for always believing in me and for pushing me through this final stretch - I could not have reached the finish line without you by my side.

Dedication

I would like to dedicate this thesis to my grandmother, Valerie de Jongh, and my son, Mikaeel Boodhoo. Nana, for as long as I can remember you have been my number one fan - always there to cheer me on. I think my doctoral thesis submission has been an even bigger dream of yours than of mine, so this is for you. My precious baby Mika, thank you for being the ultimate motivation to put all other priorities aside, work my hardest and carve out time to submit this thesis. One day when we tell you stories of your first year of life I hope that the stories of your adventures, and this one of mine, are so plentiful that stories of the pandemic are mere anecdotes.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Description	Abbreviation	Description
1000G	1000 Genomes Project	MF	molecular function
AIDS	Aquired Immunodeficiency Syndrome	MHC	Major Histocompatibility Complex
ART	Antiretroviral therapy, treatment for HIV	MSL	Mende in Sierra Leone (MSL).
ARV	Antiretroviral, treatment for HIV	Mtb	Mycobacterium tuberculosis
ATC	Anatomic Therapeutic Chemical classification system	NO	Nitric oxide
B cell	Bone marrow-derived lymphocytes	NK cell	Natural killer cell
BAM	Binary version of SAM DNA sequence alignment format (see SAM)	NNRTI	Nonnucleoside reverse transcriptase inhibitor
BLAST	Basic Local Alignment Search Tool	NRTI	Nucleoside reverse transcriptase inhibitor
BP	Biological process	NSF	Non-synonymous frameshift
BWA	Burrow's Wheeler Alignment	NSM	Non-synonymous missense
CC	Cellular component	NSN	Non-synonymous nonsense
ClinVar	Database of clinical variants	OSF	Open Science Framework
CRAM	Highly compressed alternative to the SAM and BAM DNA sequence alignment format (see SAM, BAM)	PAMP	Pathogen-associated molecular pathogen
DAVID	Database for Annotation Visualisation and Integrated Discovery	PCA	Principal components analysis
DNA	Deoxyribonucleic acid	PI	Protease inhibitor
dsRNA	Double stranded ribonucleic acid	PKR	Protein kinase R
eQTLs	Expression quantitative trait loci	PPI	Protein-protein interaction
ER	Endoplasmic reticulum	PPIN	Protein-protein interaction network
ESN	Esan in Nigeria	PRG	Population reference graph
FLASH	Fast Length Adjustment of SHort reads	PRR	Pattern recognition receptor
GATK	Genome Analysis Tool Kit	rpf	Resuscitation promotion factors
GEO	Gene Expression Omnibus	SAHGP	Sothern African Human Genome Programme
GO	Gene ontology	SAM	Sequence Alignment / Mapping
GRC	Genome Reference Consortium	SGDP	Simons Genome Diversity Project
GWD	Gambian in Western Divisions in the Gambia (GWD)	SIV	Simian immunodeficiency virus

Continued…

Abbreviation	Description	Abbreviation	Description
HHPID	HIV-1 human protein interaction database	SNARE proteins	Soluble N-ethylmaleimide sensitive factor attachment protein receptors
HIV	Human Immunodeficiency Virus	SNP	Single nucleotide polymorphism
HLA	Human Leukoctye Antigen	STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
IFN	interferon	T-cell	Thymus-derived lymphocytes
IL	interleukin	TB	Tuberculosis
IRF	interferon regulatory factor	Th cell	T helper cell
LTR	long term repeat	TLR	Toll-like receptor
LWK	Luhya in Webuye, Kenya	TNF	Tumor necrosis factor
ManLam	Mannosylated lipoarabinomannan	TRBP	TAR binding protein
MAPK	Mitogen activated protein kinase	VCF	Variant Call File
MAPQ	Mapping quality score	XDR	Extensively-drug resistant
MDR	Multidrug-resistant	YRI	Yoruba in Ibadan, Nigeria

1. Introduction

Tuberculosis (TB) and human immunodeficiency virus (HIV) co-infection is a lethal symbiosis. For individuals infected with *Mycobacterium tuberculosis* (*Mtb*), the pathogen that causes TB, HIV infection is the strongest risk factor for developing active pulmonary TB. Reciprocally, infection with TB can increase the rate of progression of HIV infection, as well as HIV replication (Deffur et al., 2013; Wells et al., 2007). TB is responsible for the most infection related deaths of people with HIV/AIDS, causing up to half of all AIDS deaths in sub-Saharan Africa (TB Alliance, 2020). Multidrug-resistant (MDR) TB, caused by strains resistant to the first-line drugs isoniazid and rifampicin, has emerged as a global epidemic (Wells et al., 2007). Given that HIV infection has also been associated with MDR-TB outbreaks, it is imperative to understand the relationship between HIV and TB infection in order to facilitate the treatment and control of both diseases. To improve our understanding of the relationship between *Mtb* and HIV, we need to be able to identify commonalities between the relationships of each pathogen with their human host. This chapter presents the literature on the human immune response to pathogenic infection, focusing on the impact of host genetic variation on the immune response, as well as the use of protein-protein interaction networks to further understanding of host-pathogen relationships.

1.1 The human immune response to pathogenic infection

Immunity is an evolutionary trait that protects individuals against exposure to pathogens, including a large array of bacteria and viruses that are pathogenic for humans. The human host has evolved many mechanisms of defence against pathogens, but pathogens have simultaneously evolved mechanisms of evading and manipulating their hosts' immune responses to their own advantage. As such, a complex network of host-pathogen interactions is continuously evolving to maintain a delicate balance of survival amongst host and pathogenic species. In this section, the human immune system and its response to HIV and *Mtb* infection will be briefly introduced.

1.1.1 Key features of the human immune system

The primary roles of the human immune system are to recognise and eliminate foreign antigens, to form immunologic memory, and to develop tolerance toward self-antigens (Goldman and Prabhakar, 1996). The immune response can be separated into three layers: (1) mechanical/chemical immunity, which is continuously operating; (2) innate immune responses, which are rapidly activated upon infection but are often unable to distinguish between pathogens or target pathogen specific traits; and (3) adaptive immune responses, which take longer to develop but are able to distinguish between and target pathogens with high specificity (Lodish, 2008). Thymus-derived lymphocytes (T-cells) are involved in the regulation of the immune response and in cell mediated immunity (Goldman and Prabhakar, 1996). Mature T-cells express antigen-specific T-cell receptors, a CD3 molecule that associates with the T-cell receptors, and a CD4 or CD8 molecule (Goldman and Prabhakar, 1996). The

T-cell receptor and CD3 complex recognise antigens presented by major histocompatibility complex (MHC) molecules on pathogen infected cells (Goldman and Prabhakar, 1996). T helper (Th) cells help to regulate cellular immunity (Th1) and assist bone marrow-derived lymphocytes (B cells) to produce antibodies (Goldman and Prabhakar, 1996). The specific functions of the Th cells depend on the types of cytokines that are generated, for example, interleukin-2 (IL-2) and interferon-γ (IFN-γ), which are generated by Th1 cells and IL-4 and IL-10, which are generated by Th2 cells. Cell mediated immunity is an inflammatory reaction that plays a pivotal role in defence against several intracellular infections such as *Mtb* (Goldman and Prabhakar, 1996). The inflammatory reaction is initiated when Th1 cells recognise specific antigens, resulting in the release of lymphokines, which recruit activated macrophages, which are cells that function to detect, engulf, and destroy pathogens (Goldman and Prabhakar, 1996). The human immune system has evolved many methods to defend the body against pathogens, many of which have not been discussed (see Goldman and Prabhakar (1996), for an in-depth review on the human immune system). In the paragraphs that follow, the ways in which *Mtb* and HIV evade the human immune responses will be discussed.

1.1.2 How the "world's most successful pathogen" - *Mtb* - evades the human immune response

The methods *Mtb* has evolved to evade human immune response has scientists referring to the mycobacterium as the "world's most successful pathogen" (Hingley-Wilson et al., 2003). It is estimated that one third of the human population is infected with *Mtb* in its latent form, which is asymptomatic and difficult to diagnose (Gupta et al., 2012). Most people infected with latent *Mtb* elicit sufficient immune responses to contain *Mtb* and control the *Mtb* population expansion (Urdahl, 2015). In immunocompromised individuals, such as people infected with HIV, this latent form is at high risk of activation and poses great risk to the individual's survival if not detected and treated rapidly. However, even with early detection, multi-drug resistant and even extensively-drug resistant (XDR) *Mtb* strains are developing, making treating tuberculosis increasingly challenging (Gupta et al., 2012).

In the case of pulmonary tuberculosis, *Mtb* is inhaled and passed through the airways to the alveolar space in the lungs where it is phagocytised by macrophages (Gupta et al., 2012). Blood monocytes are recruited by chemokine signals secreted from the infected alveolar macrophages, after which they migrate to the site of infection (Gupta et al., 2012). The monocytes differentiate into macrophages that are often activated to ingest and kill the bacteria by IFN-γ, which is secreted by T-cells within the tissue (Gupta et al., 2012). T-cell activated macrophages are thought to be the dominant effector cells against *Mtb* (Boom et al., 2003). When macrophages encounter *Mtb*, the innate immune response is activated. *Mtb* can replicate within macrophages until the macrophage cell dies and *Mtb* is released and phagocytised by other cells (Urdahl, 2015). Once *Mtb* is phagocytised by dendritic cells it can be transported from the lungs to the mediastinal lymph node (Wolf et al., 2007). CD4+ T-cells are not present in the lung-draining mediastinal lymph node until after *Mtb* has spread to that lymph node. Thus, the adaptive immune response is initiated in the local lymph node despite it being a lung pathogen residing in antigen-presenting cells (Chackerian et al., 2002). The

adaptive immune response is crucial for controlling the infection when innate immunity is insufficient; however, adaptive immune responses are not enough to eliminate *Mtb* (Wolf et al., 2008). In humans, the adaptive immune response takes five to six weeks to be activated after *Mtb* infection, which is longer than the development of adaptive immune responses to other pathogens (Wolf et al., 2008). The adaptive immune response is activated by increasing antigen concentration. The slow growth of *Mtb* means that the antigen concentration builds up slowly, contributing to the delay in adaptive immune response activation (Wolf et al., 2008). In addition, the bacterial burden caused by the buildup in the lungs prior to transport may interfere with the recognition by CD4+ T-cells or effector functions, crippling the ability of the adaptive immune response to eliminate infection (Wolf et al., 2008).

1.1.2.1 How latent *Mtb* is reactivated to its virulent form

As mentioned previously, when the host immune response is weakened, latent *Mtb* can be reactivated to its virulent form. A group of mycobacterial genes called the resuscitation promotion factors (*rpf*) are important in reactivation (Russell-Goldman et al., 2008). If these genes are deleted, the reactivation ability of the bacteria is reduced, even when the host immune system is suppressed (Russell-Goldman et al., 2008). *Mtb* has five *rpf*s (*rpf*A-E) and can survive mutants of these genes, however they are thought to be an integral player in inducing bacteria out of dormancy/latency (Russell-Goldman et al., 2008). Another factor affecting reactivation is CD4+ T-cell count. In a study investigating *Mtb* reactivation in immunodeficient non-human primates infected with Simian immunodeficiency virus (SIV), it was found that immunosuppressed animals with low CD4+ T-cell count show higher reactivation (Diedrich et al., 2010).

1.1.2.2 Host cells make a great hiding place - how *Mtb* delays phagosome maturation

Phagocytosis occurs when macrophages eradicate invading bacteria by ingesting them into a phagosome, a plasma membrane-derived vacuole (Vieira et al., 2002). After phagocytosis, the phagosomes undergo maturation to develop degradative properties (Vieira et al., 2002). *Mycobacteria* use the phagosome to their advantage by gaining access to the inside of cells where they can become intracellular pathogens (Vieira et al., 2002). When matured, phagosomes fuse with lysosomes to form phagolysosomes , which expose the bacteria to hydrolytic enzymes that can kill the bacteria (Gupta et al., 2012). Phagosome maturation begins shortly after phagocytosis, when the phagosome gets RAB5 (a GTPase) that recruits VPS34 protein, which generates phosphatidylinositol 3-phosphate on the cytosolic face of the phagosome (Vieira et al., 2002). The phosphatidylinositol 3-phosphate is a ligand for early endosome antigen 1 (EEA1) which complexes with RAB5 and recruits RAB7, another GTPase that is important for fusion to the lysosomes (Vieira et al., 2002). SNAREs (soluble N-ethylmaleimide-sensitive factor attachment protein receptors) facilitate the fusion (Vieira et al., 2002). *Mtb* can prevent the maturation of the phagosomes, which, in turn, prevents fusion to the lysosomes and the subsequent cell death (Gupta et al., 2012). Evidence suggests that phagosomes containing *Mtb* retain RAB5 for longer than other phagosomes and are unable to acquire RAB7, thereby preventing fusion (Gupta et al., 2012). In addition, *Mtb* cell wall components play a role in *Mtb*'s control over phagosomal maturation (Gupta

et al., 2012). Evidence suggests that mannosylated lipoarabinomannan (ManLAM) can inhibit VPS34, thereby preventing the acquisition of EEA1. In addition, phagosomes containing ManLAM have been shown to poorly recruit SNARE proteins, hydrolases, and EEA1(Gupta et al., 2012). This provides *Mtb* with a comfortable home for intracellular replication within the macrophages.

In addition to delaying phagosome maturation, *Mtb* nests within granulomas as another immune response evasion tactic. Once the dendritic cells present antigens to T-cells in the lymph nodes, the signaling events that follow lead to the formation of granulomas (Silva Miranda et al., 2012). When the immune system is unable to destroy the pathogen, granulomas are formed to contain and prevent further expansion of the pathogen (Silva Miranda et al., 2012). The interaction between IFN-γ secreting lymphocytes, like CD4+ T-cells, as well as tumor necrosis factor (TNF)-α is important for granuloma formation and integrity (Silva Miranda et al., 2012). However, *Mtb* persists in a latent state within the granuloma, often for several years, and reactivates in 10% of latently infected individuals (Silva Miranda et al., 2012). When the infected cells die, a necrotic zone is formed in the centre of the granuloma, which eventually disintegrates and releases the *Mtb* into the lung (Silva Miranda et al., 2012). Thus *Mtb* is able to use the host's immune response to its own advantage, enabling the population to expand and increasing the bacterial burden in the host. As a result, this makes it more difficult for the host immune response to eliminate the pathogen.

1.1.2.3 Resistance to nitric oxides

Another way *Mtb* can evade the host immune response is by resistance to nitric oxides (NOs). Studies have shown that alveolar macrophages that are infected with *Mtb* produce NO, and that there is a negative correlation between NO levels and bacterial growth (Rich et al., 1997). In addition, increased expression of NOS2 has been observed in humans with active pulmonary TB, and genetic mutations in the NOS2A gene have been associated with increased TB susceptibility (Lin and Flynn, 2010). NOs and other reactive nitrogen intermediates (RNIs) are able to attack bacterial DNA and proteins, and can cause enzyme dysfunction (Yang et al., 2009). NO is activated within macrophage phagosomes by the T-cell derived cytokines IFN-γ, TNF-α and IL-1β. The *Mtb* gene *ahp*C (alkyl hydroperoxide reductase subunit C) encodes an enzyme that protects mammalian cells from RNI toxicity (Chen et al., 1998). Other genes involved in RNI resistance in *Mtb* include methionine sulfoxide reductase (*msr*A), *nox*1, and *nox*2 (Gupta et al., 2012).

1.1.2.4 Inhibiting antigen presentation

Antigen presentation is an integral component of activating killing mechanisms in immunity, and happens in one of three ways: (1) *Mtb* antigens are presented on MHC class II molecules to CD4+ T-cells, which kill the infected cell or the bacteria by secreting IFN-γ or TNF-α ; (2) *Mtb* antigens are presented on MHC class I molecules to CD8+ T-cells, which kill the infected cell or the bacteria by secreting toxic granules; and (3) *Mtb* lipid or glycolipid antigens, like ManLAM, are recognised on CD1 molecules by CD8+ T-cells or NK cells, which

kill the infected cell (Gupta et al., 2012). Human *in vitro* models have shown that, during *Mtb* infection, the 19 kDa lipoprotein antigen (LpqH), a *Mtb* cell wall protein and Toll-like receptor 2 (TLR2) ligand, reduces the expression of MHC class II molecules on antigen presenting cells, thereby decreasing presentation to T-cells and enabling pathogen survival (Baena and Porcelli, 2009). MHC class I and class II molecules have different pathways for antigen presentation. MHC class I presentation is cathepsin-S-dependent and endoplasmic reticulum (ER)-independent (Rock and Shen, 2005). In contrast, MHC class II presentation is ER dependent. The phagocytised antigens must undergo degradation into fragments in the phagosomes or phagolysosomes and then secretion with HLA-DM in an endosomal vesicle (Gupta et al., 2012). Parallel to this process, MHC class II molecules are synthesised in the ER, where they have an attached protein (CLIP), which stabilises the molecule by binding to the peptide binding cleft (Gupta et al., 2012). The MHC class II molecule is transported to the cell membrane in an exocytic vesicle and fuses to the vesicle containing peptides. The HLA-DM molecule removes CLIP and the fragmented antigen can bind to the MHC class II molecule, stabilising it so that it may be delivered to the cell surface (Gupta et al., 2012). During MHC class II molecule formation, *Mtb* may inhibit gene expression, as well as inhibiting the upregulation on activation with IFN-γ (Gupta et al., 2012).

Cathepsins, particularly CatS, are cystein proteases that process the MHC class II-associated invariant chain, which is necessary for peptide loading and cell surface expression of MHC class II molecules (Baena and Porcelli, 2009). *Mtb* infection of human macrophages has been shown to reduce CatS activity by inducing IL-10, resulting in reduced expression of peptide loaded MHC class II complexes at the infected cell surfaces (Baena and Porcelli, 2009). A recombinant strain of BCG that secretes CatS, as well as the addition of IL-10 antibodies, have similar effects of restoring the levels of MHC class II molecules at the cell surface (Baena and Porcelli, 2009). The *Mtb* gene *ure*C produces urease, which is also involved in preventing MHC class II molecules from being presented at the cell surface (Baena and Porcelli, 2009).

As demonstrated above, the mechanisms through which *Mtb* evades the immune response are not clear cut, further illustrating how complex this host-pathogen relationship is. The next section will discuss ways in which the viral pathogen, HIV-1, is able to evade the human immune response.

1.1.3 Human immune response to HIV infection and the strategies for escaping antiviral activity

HIV-1 is a pathogenic human virus that is mostly transmitted by sexual exposure through the genital tract or rectal mucosa (McMichael et al., 2010). Acquired Immunodeficiency Syndrome (AIDS), is the stage of HIV infection when the immune system is at its weakest. According to the Statistics South Africa (2017) mid-year population estimates, AIDS-related deaths in South Africa are on the decline but AIDS was still the cause of 25% of deaths in 2017 (decreasing from 49% in 2006 following the increased roll out of antiretroviral therapy (ART)). However, 18% of South African adults aged 15-49 years are estimated to be HIV positive (Statistics South Africa, 2017). As such, the need to prevent new infections is of high priority. In this section, key aspects of the immune response to HIV will be explained and the some of the

tactics used by HIV-1 to evade the immune system will be discussed briefly.

1.1.3.1 Human immune response to HIV infection

During the initial phase of HIV-1 infection, innate immune factors such as inflammatory proteins, including cytokines, chemokines, and antiviral restriction factors, are activated (Guha and Ayyavoo, 2013). These immune factors, particularly IFNs, play an important role in defending the host against the virus by modulating the downstream signaling events, inducing dendritic cell maturation, and by activating macrophages, natural killer (NK) cells, and B and T-cells (Guha and Ayyavoo, 2013). When the pathogen breaches the mucosal barrier (or other physical barrier), they are recognised by pattern recognition receptors (PRRs) expressed either in the cytoplasm or on cell membranes (Iwasaki, 2012). PRRs, such as toll-like receptors, are required to identify pathogen-associated molecular patterns (PAMPs), which are snippets of the virus that contain a conserved viral tag that can be recognised as foreign by the host (Iwasaki, 2012). TLRs recognise nucleic acids from the viral genome. In addition to TLRs, Viral PAMPs are recognised by other PRRs, such as the protein DC-SIGN, which binds to viral envelope glycoproteins (Guha and Ayyavoo, 2013). When viral ligands interact with host receptors, downstream signaling events are activated, which, in turn, activate transcription factors that regulate the expression of genes functioning in innate and adaptive immune responses (Guha and Ayyavoo, 2013). The binding of TLRs to viral PAMPs causes the intracellular part of TLRs to bind to the protein MyD88, which activates mitogen activated protein kinase (MAPK), ultimately activating nuclear factor-κB (NF-κB) (Kawai and Akira, 2007). NF-κB activation promotes the regulation of inflammatory cytokine genes and activates the interferon regulatory factor (IRF) (Kawai and Akira, 2007). IRF, in turn, activates type I IFNs, which have antiviral activity (Yoneyama et al., 1998). However, HIV-1 has evolved several ways to avoid the antiviral effects of the innate and adaptive immune system. This will be discussed in the paragraphs that follow.

1.1.3.2 HIV-1 can permeate the mucosal barrier

The mucosal tissues present the first line of physical defence against invading pathogens such as HIV-1. However, HIV-1 can evade this barrier by crossing the mucosa and establishing the infection in susceptible host cells (Guha and Ayyavoo, 2013). Exposure to HIV-1 upregulates inflammatory cytokines in the mucosal membrane, which increases the permeability for the virus particle, enabling HIV-1 to use the intercellular spaces to move through the epithelium (Guha and Ayyavoo, 2013). The primary target cells for HIV-1 are cells expressing CD4 on their surface, such as CD4+ T-cells, and cells of myeloid lineage, including monocytes, macrophages, and dendritic cells (Kedzierska et al., 2003). These cells express the chemokine co-receptors for HIV-1 entry, CCR5 and CXCR4. CD4+ memory T-cells and CCR5+ memory T-cells are the first HIV-1 infected cells in cervical tissue (Gupta et al., 2002; Hladik et al., 2007). Early infection of T-cells in the mucosa enables rapid viral expansion (due to the high replication rate of T-cells), and transferal to the dendritic cells (Gupta et al., 2002). Dendritic cells can transport HIV-1 to the lymph nodes where CD4+ T lymphocytes are infected, the prime target of HIV-1 (Gupta et al., 2002). The infected CD4+ T-cells disseminate to the lymphoid tissues with the help of proinflammatory cytokines, enabling the virus to replicate

and spread at a high rate (Guha and Ayyavoo, 2013). In addition to infecting CD4+ T-cells, HIV-1 enters Langerhans cells in the mucosa through endocytosis (Hladik et al., 2007). Intact virions have been observed to remain in epithelial Langerhans cells for at least three days, suggesting that Langerhans cells can house viable HIV-1 for an extended period of time before passing it on to T-cells (Hladik et al., 2007).

This is one of many means by which HIV might evade a protective microbicide, which are drugs that are topically applied to mucosal surfaces to prevent infection (potential ways HIV may evade a microbicide are reviewed in Hladik and Doncel (2010)). The potential advantage of an effective microbicide is demonstrated by Karim et al. (2010). The authors conducted an HIV prevention trial in South Africa, using a microbicide vaginal gel containing 1% of the ARV tenofovir. The tenofovir gel reduced HIV acquisition by an estimated 39% overall, and by 54% in women with high gel adherence (Karim et al., 2010). Thus, while HIV-1 can enter the mucosa, strengthening this first physical barrier by means of microbicides and preventing inflammatory conditions, such as bacterial vaginosis, may assist in preventing new infections.

1.1.3.3 HIV-1 avoids immune recognition

One of the most important immune evasion strategies for HIV-1 is its genetic variability and high mutation rate (Roberts et al., 1988). The high mutation rate of HIV-1 is caused by its error prone viral reverse transcriptase which lacks proofreading activity, enabling random mutations that help the virus to avoid immune recognition and thereby preventing immune response (Roberts et al., 1988). In addition, the PAMPs exposed by HIV-1 are not as easily recognised by PRRs as they are for other viruses, enabling HIV-1 to escape from proinflammatory and antiviral responses (Guha and Ayyavoo, 2013). HIV-1 PAMPs are presented to the cell in the form of capsid structure and nucleic acid (Rustagi and Gale, 2014). When PRRs detect HIV-1, IFN and innate intracellular defences are activated to drive gene expression of restriction factors (Rustagi and Gale, 2014).

Restriction factors are proteins that: (1) exhibit antiviral activity, (2) are often constitutively expressed and sometimes upregulated by interferons, (3) inhibit viral replication, and (4) have co-evolved with viruses resulting in diverse amino acid sequences (Blanco-Melo et al., 2012). Most antiviral restriction factors are constitutively activated, without any requirement for co-factors, binding partners or downstream effector molecules to act upon their viral targets (Blanco-Melo et al., 2012). Unlike T and B cells, which rely on somatic recombination for diversity, restriction factor genes do not undergo recombination events and are constitutively expressed directly from the germline (Blanco-Melo et al., 2012). They are highly diverse - HIV therefore requires multiple mutations that will confer new functions to evade their action (Blanco-Melo et al., 2012).

There are four classes of restriction factors that target HIV-1, namely: APOBEC3 proteins, TRIM5 proteins, Tetherin, and SAMHD1 (Blanco-Melo et al., 2012). One of these restriction factors, TRIM5α, is meant to recognise and break-down the viral capsid, but it has been shown to be inefficient at doing this (Stremlau et al., 2006). In addition, other restriction factor proteins are antagonised by viral proteins, such as Vif, Vpu, Vpx, Nef, and Rnv (Blanco-Melo et al., 2012). For example, Tetherin, which functions to contain viruses in infected cells, is

actively antagonised by the HIV-1 protein Vpu to promote release of the virus from infected cells (Neil et al., 2008). In addition, APOBEC3G, a deaminase that mutates reverse transcribed DNA, is targeted for inhibition by the HIV-1 Vif protein to prevent APOBEC3G from hypermutating the viral genome (Sheehy et al., 2003).

Thus, while the human host has evolved mechanisms to detect and destroy HIV, the fast mutation rate of the virus enables it to evolve much more quickly to its own advantage.

1.1.3.4 HIV-1 infection induces a "cytokine storm"

HIV-1 pathogenesis is greatly influenced by cytokines and chemokines, and HIV exploits the network of cytokines and chemokines throughout its life cycle (Guha and Ayyavoo, 2013). During HIV-1 infection, immune cells are highly activated, which leads to increased production of pro-inflammatory and anti-inflammatory cytokines and chemokines such as IFNs, IL- (IL)-1, -2, -4, -8, -6, -10, -15, and TNF-α (Stacey et al., 2009). Stacey et al. (2009) refer to the increased production of these molecules in acute HIV infection as a "cytokine storm", which in most cases results in the inability of the host to contain the virus. In addition, although some of the cytokines and chemokines produced in acute HIV infection may assist with controlling viral replication, the "cytokine storm" contributes to the early immunopathology of the infection (Stacey et al., 2009).

HIV-1 can supersede cytokine/chemokine networks of the host using a variety of mechanisms, such as mimicking or by modulating certain cytokines. For example, HIV-1 transactivator of transcription (Tat) protein seems to mimic features of CC-chemokines, which are often chemoattractants for monocytes (Albini et al., 1998). CCR5, for example, is a coreceptor for HIV, and mutations of CCR5 have been shown to be associated with resistance to HIV infection, as well as delayed disease progression to AIDS (Dean et al., 1996). Similarly, CCR2, which mediates HIV entry into cells, has also been shown to contain mutations associated with delayed disease progression to AIDS (Smith et al., 1997). Tat has been shown to activate CCR2, which would, in turn, recruit chemokine expressing cells (such as macrophages) to the infected cells promoting the spread of HIV infection (Albini et al., 1998). Another example is HIV-1 Nef and vpr proteins, which are respectively able to upregulate and downregulate proinflammatory cytokine levels, such as IL-1β, IL-12, IL-15, and TNF-α (Guha et al., 2012; Quaranta et al., 2002). Thus, HIV-1 can both mimic and modulate certain cytokines, and can thereby manipulate the cytokine network to enhance its replication and survival.

1.1.3.5 HIV-1 inhibits or reduces interferon production

Interferons stimulate genes that are important for inducing an immune response (Guha and Ayyavoo, 2013). One such gene is protein kinase R (PKR), which binds to double stranded RNA (dsRNA) produced during viral replication, forms a dimer by autophosphorylation, and proceeds to block viral replication by inhibiting translation of alpha subunit of elongation initiation factor two (eIF2α) in infected cells (Guha and Ayyavoo, 2013). HIV-1 replicates in cells with a high level of TAR binding protein (TRBP), where HIV-1 Tat RNA binds to TRBP, which inhibits PKR activation (Sanghvi and Steel, 2011). HIV-1 Tat also acts as a substrate homologue

of eIF2α and competes for PKR mediated phosphorylation. PKR phosphorylates HIV-1 Tat, and the phosphorylation of Tat is necessary for HIV-1 LTR transactivation (Sanghvi and Steel, 2011). This prevents phosphorylation of eIF2α and enhances viral replication (Sanghvi and Steel, 2011).

These examples demonstrate that while the human host has evolved a complex immune system with multiple strategies to prevent and eradicate pathogens, like *Mtb*, HIV-1 is able to evade the immune responses. In the South African context, where an estimated 12.6% of the population are HIV-infected and of those only 53% are receiving anti-retroviral drugs (Statistics South Africa, 2017), the complexity of the host-pathogen relationship is exacerbated by how latent *Mtb* is able to reactivate in immunocompromised individuals. The next section will discuss the complexities of HIV-TB co-infection.

1.1.4 Effects of HIV and *Mtb* on the immunity to each other

Active tuberculosis infection and HIV infection separately are two of the leading causes of death around the world. Together, the pathogens have a more pronounced and often more fatal effect on the host. According to the World Health Organisation (2015), the risk of developing TB is between 16 and 27 times greater in people living with HIV than in people without HIV. In 2015 alone, 11% of 10.4 million new TB cases occurred in HIV-infected individuals (World Health Organisation, 2015). Whilst host immune response towards pathogens is regulated in order to maintain a balance of minimal host damage and pathogen containment, during HIV-1 infection, the host immune system is compromised, providing the perfect environment for other pathogens, such as *Mtb* (Toor et al., 2014).

1.1.4.1 HIV depletes CD4 T-cells

The most important effect of HIV on immunity to TB is the depletion of CD4 T-cells from secondary lymphoid tissues (Kwan and Ernst, 2011). HIV can deplete CD4 T-cells using several mechanisms, however, the induction of apoptosis of directly infected and bystander cells has the most prominent impact on reducing CD4 T-cells (reviewed by Hazenberg et al. (2000)). As stated previously, T-cell activated macrophages are the dominant effector cells against *Mtb* (Boom et al., 2003). As such, reduction of CD4 T-cells results in an inefficient immune response to TB infection.

1.1.4.2 HIV impairs the function of CD4 T-cells

As well as quantitatively reducing the number of CD4 T-cells, HIV infection weakens the qualitative function of the remaining CD4 T-cells (Kwan and Ernst, 2011). HIV infection decreases the surface expression of the CD4 molecule through an effect of the HIV Nef protein which acts by inducing CD4 endocytosis, resulting in its degradation in lysosomes (Aiken et al., 1994). In addition, Chaudhry et al. (2009) have shown that the HIV Nef protein prevents HLA class II molecules from reaching the cell surface in monocytic cells (like macrophages) by delaying the transport of HLA class II molecules to the cell surface and by accelerating endocytic removal of cell surface HLA class II molecules in lysosomes.

Furthermore, Nef has been shown to inhibit HLA class II antigen presentation to T-cells by reducing the surface level of mature HLA class I molecules (Stumptner-Cuvelette et al., 2001). HIV is thus able to disrupt antigen presentation, enabling it to evade the host immune system and to impede host CD4 T-cell recognition peptides from invading pathogens, such as *Mtb* (Kwan and Ernst, 2011).

1.1.4.3 HIV reduces the ability of granulomas to contain *Mtb*

Another way that HIV negatively impacts the immune response to TB is by affecting the granuloma structure (de Noronha et al., 2008). The granuloma functions to contain the *Mtb* infection and prevent the spread of the bacteria to surrounding tissues (de Noronha et al., 2008). Individuals co-infected with HIV and TB have been shown to have altered granulomas, resulting in increased *Mtb* growth and spread, thereby worsening the TB disease (de Noronha et al., 2008; Shankar et al., 2014).

1.1.4.4 HIV replication is increased at sites of TB infection

In addition to HIV affecting TB pathogenesis, TB has been associated with increased HIV replication, accelerated progression of HIV disease, and increased risk of mortality (Badri et al., 2001; Kwan and Ernst, 2011; Whalen et al., 1995). In a prospective cohort study in South Africa of HIV-infected patients residing in an area with high tuberculosis prevalence, Badri et al. (2001) found that TB infection was associated with higher risk of mortality and increased frequency of AIDS-defining illness (i.e. the progression of HIV infection). In addition, active pulmonary TB has been found to be associated with increased HIV replication due to the higher viral load at the sites of TB infection (Toossi et al., 2001). Furthermore, the genetic diversity of HIV harvested from sites of pulmonary TB infection is higher than that from pulmonary sites without TB, which may further indicate higher levels of replication at sites of TB infection (Nakata et al., 1997).

1.1.4.5 Increased concentrations of TNF-α lead to increased TB growth and HIV replication

HIV infection increases the risk of progression to active TB in both primary TB infection and the reactivation of latent TB (Kwan and Ernst, 2011). Imperiali et al. (2001) showed that *Mtb* growth was enhanced in HIV-1 infected monocyte derived macrophages compared to HIV-1 uninfected macrophages. In HIV-1 infected macrophages, they observed higher concentrations of TNF-α after *Mtb* was phagocytised. TNF-α, along with IFN-γ, functions to eliminate *Mtb* (Barnes et al., 1990). However, Imperiali et al. (2001) observed that addition of TNF-α to HIV-1 infected macrophages resulted in increased *Mtb* growth, suggesting that the antimycobacterial activity of TNF-α is inhibited in HIV-1 infected macrophages. Additionally, HIV-1 replication was increased by the addition of TNF-α. These results suggest that TNF-α plays an important role in increasing both HIV-1 expression and *Mtb* growth, and that the effects of TNF-α on *Mtb* growth may be mediated by enhanced HIV-1 expression.

1.1.5 Treating individuals co-infected with HIV and TB

Given the interactions at the molecular level between HIV and *Mtb*, treating individuals infected with both pathogens can be more complex than treating only one of the infections. In South Africa, Lawn et al. (2009) found that in a setting with a high risk of tuberculosis, treatment with and adherence to antiretroviral therapy is associated with decreased TB incidence rates amongst HIV-infected individuals. This was associated with the CD4 cell count of the individuals, whereby individuals with CD4 cell counts higher than 500µl/ml had the lowest incidence rates of TB (Lawn et al., 2009). However, even with ART, the overall rate of TB incidence was almost 10 times higher in HIV-infected patients than for uninfected individuals. In addition, individuals on ART with CD4 cell counts above 500 µl/ml had double the incidence rate of TB than in uninfected individuals (Lawn et al., 2009).

1.1.5.1 Anti-HIV and anti-TB drug interactions

Although adherence to HIV treatment is associated with lower TB incidence rates, commonly used HIV and TB drugs are known to have adverse interactions when treating HIV-TB co-infected individuals. Daily dosing of the first-line TB drug, rifampicin, as part of TB treatment is critical to prevent treatment failure caused by acquired rifampicin resistance in HIV-infected individuals with TB (Khan et al., 2010). However, even with strict adherence to treatment, rifampicin is an inducer of the cytochrome P450 system, and, as such, interacts with several classes of drugs, including antiretrovirals (Kwan and Ernst, 2011). Rifampicin interacts with protease inhibitors (PIs), a class of antiretroviral drugs which are substrates of CYP3A4 (Kwan and Ernst, 2011). Rifampicin induces CYP3A4 to such an extent that, even with ritonavir boosting (a drug commonly issued with PIs to inhibit CYP3A4), it can accelerate the metabolism of PIs, leading to negligible concentrations of PIs in the serum (Kwan and Ernst, 2011). As such, the preferred regimen for the concomitant treatment of HIV and TB is efavirenz-based ART with rifampin-based TB treatment (National Department of Health, 2015).

Nevirapine is the most commonly used nonnucleoside reverse transcriptase inhibitor (NNRTI) in settings that are resource-limited for three reasons (1) it is much cheaper than efavirenz, (2) it is available in fixed dose combinations, and (3) it is not known to have teratogenic effects and is thus much safer than efavirenz for women of child-bearing age (Boulle et al., 2008). In a South African prospective cohort study of HIV-infected individuals, the use of nevirapine with rifampicin was associated with higher risk of elevated viral loads 6 months after starting treatment in individuals with active TB than in individuals without active TB (Boulle et al., 2008). In addition, co-infected individuals developed virological failure sooner than individuals without TB (Boulle et al., 2008). There were no differences in risk of virologic failure between individuals starting efavirenz with or without concurrent TB (Boulle et al., 2008).

1.1.5.2 Human genetic variants and TB/HIV drug susceptibility

In addition to TB and HIV drugs interacting, the pharmacokinetics of drugs can be altered by genetic variants in drug metabolising enzymes or transporters (Kwara et al., 2009). In addition, as with other drugs, there are limited data on the pharmacokinetics of ARV drugs and their

generics in African populations compared to other populations, and even fewer which examine the pharmacokinetics in HIV-TB co-infected individuals (Kwara et al., 2009). Kwara et al. (2009) determined the pharmacokinetics of the ARVs lamivudine, zidovudine, and stavudine in a cohort of HIV-TB co-infected Ghanaian individuals during concurrent TB treatment that included rifampicin. They assayed for the common single nucleotide polymorphism (SNP) UGT2B7*1c in the gene UDP glucuronosyltransferase (UGT) 2B7, and showed higher in vivo clearance and faster hepatic glucuronidation of zidovudine in individuals with this allele (Kwara et al., 2009). Similarly, SNPs of the gene CYP2B6 have been associated with high plasma efavirenz concentrations both in Thai individuals on concurrent HIV and TB treatment with rifampicin (Manosuthi et al., 2013), and in African individuals only on anti-HIV therapy (Wang et al., 2006).

These examples illustrate the potential impact that genetic variation can have on HIV and TB treatment. In the next section, we will present examples from the literature that illustrate the impact of variation on HIV and TB susceptibility, with a particular focus on the human major histocompatibility complex region.

1.2 Human genetic variation and susceptibility to HIV and TB

The human major histocompatibility complex, also called the human leukocyte antigen (HLA) system, contains hundreds of genes related to immune system function, and, as such, has been shown to contain multiple genetic factors influencing both HIV and TB infection (Pérez-Núñez and Martínez-Quiles, 2011; Qidwai et al., 2012). The MHC is the most polymorphic region of the genome, displaying immense inter-population and inter-individual variation. It has been postulated that infectious diseases have inferred selective pressures on the region, resulting in genetic diversity of the region (Lombard et al., 2006). The MHC is located on the short arm of chromosome 6 and there are three major classes of molecules: (1) HLA class I and (2) class II genes, which are both involved in antigen presentation to T-cells, and (3) HLA class III, which include other immune system genes. The MHC molecules, particularly class II, are highly polymorphic, raising the question of whether allele sequence variation affecting the expression of HLA proteins could alter disease susceptibility. HLA alleles have not only been shown to affect HIV progression, but also horizontal and vertical transmission (see Pérez-Núñez and Martínez-Quiles (2011) for an extensive review). Similarly, many HLA alleles have been shown to be associated with susceptibility to TB (see Möller et al. (2010) for an extensive review). In addition, TB strains are hypothesised to have drifted together with human migration, and, as such, adapted to specific populations (Hanekom et al., 2007). This offers a potential explanation for the observation that some of the HLA alleles associated with TB are population specific (Lombard et al., 2006). There is increasing evidence that host genetic factors affect susceptibility to TB and HIV, as well as response to treatment. In the next section, genetic variants that influence disease progression and susceptibility to HIV or TB infection will be discussed.

1.2.1 Variants contributing to TB susceptibility

Lombard et al. (2006) suggest that the study of genetically diverse African genomes may provide better insight into the genetic basis of complex diseases, like TB. Because genes in the MHC region are involved in immunity, they are under immense selective pressure resulting in greater variability than other genes (Lombard et al., 2006). Lombard et al. (2006) investigated the association of HLA-DR and HLA-DQ polymorphisms, as well as vitamin D receptor polymorphisms, in individuals with TB compared with healthy controls in the Venda population in South Africa. The variants within *HLA-DRB1* and *HLA-DQB1* were shown to have significantly higher allele frequency in TB cases than in controls and may predispose this population to TB. Similarly, Barreiro et al. (2012) identified host genetic variants associated with TB susceptibility by performing a genome wide mapping study of loci that are associated with functional variation in immune response to *Mtb* infection. They identified the MAPK phosphatase *DUSP14* as a susceptibility gene for pulmonary TB. Further, a variant rs712039 that affects the expression of *DUSP14* was also identified and found to be associated with the secretion levels of the important cytokines TNF-α and IFN-γ.

1.2.2 HLA variants contributing to HIV susceptibility and disease progression

As with TB, several HLA genes and variants have been associated with HIV susceptibility and disease progression. For example, HLA-B27 has been linked to slow HIV disease progression, whilst HLA-B35 has been shown to play a direct role in rapid HIV progression and is thought to be a marker of disease susceptibility (Pérez-Núñez and Martínez-Quiles, 2011). HLA-G expression levels have been shown to be higher in patients undergoing ARV treatment than in untreated individuals, and different HLA-G alleles have been associated with protection and susceptibility to HIV infection (Pérez-Núñez and Martínez-Quiles, 2011). The progression of HIV-1 infection is primarily ascertained through viral load, which indicates how many copies of the virus are detected per milliliter of blood. A study by Saathoff et al. (2010) showed that viral load was 40% lower in individuals with protective alleles (such as HLA-A*0205, HLA-B*5801, HLA-B*8101, HLA-B*4201, and HLA-B*5703) than those with harmful alleles (such as HLA-B*5802, HLA-B*4501, HLA-B*1801, and HLA-B*1503) or any other neutral alleles. HLA class II allele DRB1*1303 was shown to be associated with low plasma viral load in a South African population infected with HIV-1 clade C, as well as in a European population infected with HIV-1 clade B (Julg et al., 2011). It is possible, although not yet confirmed, that the protective effect of the allele is as a result of CD4+ T-cell responses that are mediated by these genes (Julg et al., 2011). In a study by Carrington et al. (1999) it was found that maximum heterozygosity of HLA-A, -B, and -C genes resulted in delayed progression to AIDS, while homozygosity resulted in rapid progression to AIDS in a Caucasian population.

Mutations in the HIV protein fragments (epitopes), which are normally recognised by CD8+ T-cells after HLA molecule presentation, can result in HIV evading T-cell responses, and thus confer a survival advantage on those strains (Kawashima et al., 2009). These mutations are selected for in the viral population until they are fixed. The "Gag epitope" is an example of such a mutation and is recognised by HLA-B27, which usually binds to peptides that contain an arginine at the second position (Phillips et al., 1991). Patients with HLA-B27 progress more

slowly to AIDs than other HIV-infected individuals. The Gag epitope contains a mutation that changes this arginine so that it is no longer recognised by T-cells (Phillips et al., 1991). The mutation does, however, have a fitness cost which results in the virus reverting to the wild-type sequence unless the mutation is selected for. The association between escape variants and HLA types is strong evidence of HIV adaptation to HLA at the population level (Kawashima et al., 2009).

The diversity of the MHC region, coupled with its widespread association with immune response, makes characterising the variation in individuals of diverse ancestry important for investigating disease susceptibility, immune response to infection, and evolution of the MHC locus. Both HIV and *Mtb* are likely to have undergone selection by, and, in turn, influenced selection on, the MHC. Both of these infectious diseases are responsible for particularly high rates of mortality in Africa relative to any other continent. With an increase in available African and specifically Southern African whole genome sequences, the prospects for accurately and fully characterising variation in this region are promising (Choudhury et al., 2017).

1.3 Challenges of identifying variants in ancestrally diverse populations

High throughput sequencing and variant calling techniques have enabled the sequencing and identification of millions of human sequence variants in ancestrally diverse populations by projects such as the 1000 Genomes project (Consortium et al., 2012) and the Simons Genome Diversity project (www.simonsfoundation.org), which also provide publically available data. When human genome sequences are assembled, the sequencing reads are aligned to a reference genome sequence to ensure coordinate consistency and computational efficiency. The current reference genome (GRCh38) is "linear"; in other words, each genetic element is uniquely describable by its coordinate position along the genome. Initially it was thought that this single non-redundant haploid path of the genome would sufficiently represent the human genome (Kent and Haussler, 2001). This fits the prediction that single nucleotide polymorphisms would be the most prevalent source of variation between populations (Church et al., 2011). However, this notion excludes structural complexities and sequences that do not closely map to the reference chromosome sequences. The Genome Reference Consortium (GRC) aimed to build upon this by representing structurally complex regions, such as the MHC, with multiple paths and incorporating one path into the assembly and the other branching paths into an additional file (Church et al., 2011). Paten et al. (2014) suggest, however, that the downfall of this linear reference genome is that it incorporates little variation information, which is instead separated across various data sources in different formats; for example, 1000 Genomes (1000 Genomes Project Consortium, 2012), dbSNP (Sherry et al., 2001), and dbVar (Lappalainen et al., 2013). This makes all science biased toward the reference, which is an issue as the quality of the reference impacts the ability to characterise regions of high structural or sequence diversity (Dilthey et al., 2015). In addition, there is no standard method for alignment to a reference sequence. As such, different mapping procedures can produce different mapping results, not only in terms of overall read placement but also in terms of the placing of bases within a read, which can result in

erroneous determination of variants (Paten et al., 2014).

1.3.1 Biases in read alignment and variant classification

Because HLA loci are the most polymorphic in the human genome, they are incredibly prone to mapping bias and miscalculations of allele frequency. In addition, HLA genes are often highly paralogous, which increases the chance that a read will be mapped to many locations and be discarded by most alignment platforms due to multi-mapping (Treangen and Salzberg, 2012). To investigate the effects of variant calling from sequences aligned to a single reference genome, Brandt et al. (2015) compared variants that had been discovered by whole genome sequencing after mapping to the single reference (1000 Genomes Project Consortium, 2012) to variants called from Sanger sequencing of the same samples (PAG2014) in five classical HLA genes. Their results showed that, on average, 18.6% of genotypes were mismatched between 1000 Genomes and PAG2014, and that higher nucleotide diversity tended to result in higher mismatches. In addition, most of the genotype mismatches were found to be due to miscalling an alternative allele as a reference allele. Likewise, they found that deviations in allele frequency estimates were in the direction of an overestimation of reference allele frequencies in the 1000 Genomes data. Mapping bias can result in failure to identify true variants because they are present in haplotypes that diverge from the reference, resulting in reads generated from these regions being unmapped (Brandt et al., 2015). One method that has been suggested to avoid missing identifying variants due to reference-mapping bias, is to *de novo* assemble sequencing reads and identify variants from the *de novo* assembled reads relative to the reference genome (Iqbal et al., 2012). Identifying variants in this manner, to supplement existing collections of variants, may improve the reliability of the variants in predictions of downstream functions.

1.3.2 From a linear reference towards a reference graph

As mentioned previously, the linear reference genome fails to capture inter-population variation in highly polymorphic regions of the genome, such as the MHC (Dilthey et al., 2015). Thus, the diversity of the MHC region makes assembling sequences of individuals who are highly divergent from the reference a difficult and error prone task (Dilthey et al., 2015). Given that there is no current African reference sequence, the problem for characterising highly diverse regions, such as the MHC, within African populations is apparent. Paten et al. (2014) propose a new reference genome structure, which they describe as a reference graph that includes a reference genome assembly, as well as large scale and small scale common human variation. They suggest that a reference hierarchy constructed from human genomes from diverse populations and subpopulations would efficiently create a detailed representation of human genetic variation. This may have important applications in both population genetics and medical genomics. They also suggest that each human genome would comprise of a separate context-driven reference structure that could then be grouped by subpopulations – where each subpopulation is represented by a merged sequence graph that includes all of the variation in the subpopulation's genomes. This would result in something like a subpopulation-specific reference genome, which may enable the discovery of haplotypes specific to these regions (Paten et al., 2014).

Some positions in an input genome could be novel to that genome, such as virally inserted DNA or highly mutated stretches of DNA that cannot be mapped to even a general reference graph. Using the Paten et al. (2014) structure, these unmapped positions will form an unmapped subgraph, consisting of unmapped components of the genome. Each of these unmapped components is adjacent to a set of mapped positions in the genome, allowing approximate mapping to a neighbourhood of possible regions. If a variant occurs frequently in new genomes it may be a human genetic variant that could be added to the reference graph (Paten et al., 2014). The use of a human reference graph will also alleviate the need to remap coordinates whenever the genome assembly is updated, as the identifier for each position is a specific and permanent position in the sequence graph (Paten et al., 2014).

1.3.2.1 An example of a population reference graph of the MHC region

Dilthey et al. (2015) describe an approach to represent variation within a reference, known as a population reference graph (PRG), and tested the approach on the MHC region. They define a PRG as a directed, acyclical model for genetic variation that incorporates information about known allelic relationships between sequences. The first step in constructing the graph involves multiple sequence alignment of the reference sequences. Thereafter, by collapsing aligned regions with sequence identity over a defined length, a graph structure is generated from the alignment. Finally, additional SNP information is added to paths in the graph with matching sequence at those positions. To develop their graph, Dilthey et al. (2015) used the REF and ALT sequences of GRCh37, SNPs from the 1000 Genomes project (1000 Genomes Project Consortium, 2012), as well as the set of HLA allele sequences from the International Immunogenetics Information System (Lefranc et al., 2009). Using the PRG, they intend to enable inference of the diploid path that most closely resembles the two haplotypes of a sample. This diploid path is a merging sub-graph of the PRG where heterozygous sites result in bubbles in the graph, enabling the identification or decomposition of two paths. Then, to detect novel variation in the sample, reads are mapped to the two resulting paths and a standard variant caller is used to discover new alleles.

In highly diverse regions of genome sequences, such as the HLA alleles, mapping to a single reference followed by variant calling is prone to errors due to the density of mismatches to the reference. Dilthey et al. (2015) compared the per base diploid genotypes inferred by mapping and PRG approaches to the results expected from sequence-based typing of highly polymorphic regions. In regions of low diversity and high sequence coverage, the accuracy of all approaches is very high. However, when coverage is low or divergence is high, the PRG approach is more successful than mapping. Dilthey et al. (2015) suggest that the PRG could incorporate weights along the various paths according to population frequency. They also suggest that linkage disequilibrium information could be used to improve the genome inference.

1.3.2.2 Using a reference graph to identify variation in *Plasmodium falciparum*

A similar method to the PRG, was implemented to identify variation in the malaria parasite *Plasmodium falciparum* (Miles et al., 2015). The methods used in this study are of particular

relevance to the issues faced when trying to understand the MHC region, due to the evolutionary complexity and vast variability of the *Plasmodium falciparum* genome. The 23Mbp genome of *Plasmodium falciparum* has biased nucleotide composition, as well as highly repetitive regions and high levels of diversity between genes (Miles et al., 2015). Knowledge of variation in non-coding regions of the genome is limited, and, like the MHC, knowledge of complex variation in haplotypes highly divergent from the reference is also very limited (Miles et al., 2015). Miles et al. (2015) use combinations of methods for variant discovery, including alignment to a reference using a Burrow's Wheeler Alignment (BWA) and Genome Analysis Tool Kit (GATK), as well as reference-free sequence assembly using Cortex (Iqbal et al., 2012), to build a map of genome variation within *Plasmodium falciparum*.

1.3.2.3 The advantages of reference graphs for African populations

While African populations have the most genetic diversity and the highest per capita health burden according to the WHO, they are underrepresented in large scale genome studies of disease association (Choudhury et al., 2017). This genetic diversity poses both a challenge and an opportunity for biomedical research, as well as the hope that Africans will benefit from genomic medicine in the future (Choudhury et al., 2017). Creating reference genome sequence graphs for African populations would reduce the reference mapping bias, as observed in the current linear reference, and in so doing may uncover novel variation that could have implications for disease management. Marschall et al. (2016) discuss the potential advantages of a pan-genomic reference graph structure; with the new structures that are developing we should expect substantial improvements in characterising parts of the genome that are currently not easily accessible with current sequencing technologies, as well as the detection of complex structural variants, particularly when long-read sequencing data is incorporated. Furthermore, because the use of genomic sequencing data in a clinical setting is expected to rise, improving the accuracy of genetic variation identification in diverse populations will be a great advantage (Marschall et al., 2016).

This section has focused on the need to improve methods to identify genetic variation in diverse populations. To fully understand the impact of this variation on the individual's biology, immune responses and susceptibility to disease, networks of molecular interactions at various levels need to be considered. To this end, in the next section, human-pathogen protein interaction networks will be discussed.

1.4 Human-pathogen protein interaction networks

Owing to the complexity of the immune system, considering the effect of individual alleles on single proteins is increasingly being extended to consider the impact of alleles on gene-gene and gene-strain interactions (Deffur et al., 2013; Möller et al., 2010). Many studies have tried to explore these interactions by applying a reductionist approach, that is considering the effects of small sets of alleles on small sets of genes. Other studies have made use of high throughput technologies to extend these approaches. For example, a genome-wide screening approach was used to identify human sequence variants associated with gene expression levels (expression quantitative trait loci, eQTLs) in *Mtb* infected compared to *Mtb*

uninfected dendritic cells (Barreiro et al., 2012). Systems biology approaches to understanding the relationship between HIV, *Mtb*, and their human host may enable the prediction of emerging patterns that could complement reductionist approaches (Deffur et al., 2013). Interactions between the human and pathogen proteins are essential for the survival and viability of the pathogen within the host. As such, human-pathogen protein-protein interaction networks of human-*Mtb* proteins (Huo et al., 2015; Rapanoel et al., 2013), as well as human-HIV-1 proteins (Fu et al., 2009), have been constructed, which may further understanding of the molecular mechanisms of survival and pathogenicity.

1.4.1 Protein interactions and interaction networks

The function of a protein cannot be fully understood by studying the protein in isolation, and interactions between proteins can be used to describe and narrow down a protein's function. Some of the functions of protein-protein interactions (PPIs) are to: modify the kinetic properties of enzymes, enable substrate channeling, create new binding sites for small effector molecules, inactivate or suppress a protein, change the binding specificity of a protein, and to regulate downstream or upstream processes (Rao et al., 2014). PPIs can be detected using a variety of methods, including *in vitro* methods, such as affinity chromatography and protein arrays, as well as *in vivo* methods, such as yeast two-hybrid systems. Moreover, PPIs can be predicted using *in silico* methods, such as sequence-based approaches, gene expression-based approaches, and structure-based approaches (Rao et al., 2014). Protein interactions can be described as physical or functional interactions; whereby physical interactions involve physical contact between proteins (e.g. binding) and functional interactions are indirect associations between proteins (e.g. involvement in the same biological process). Because interactions can be discovered or predicted using a wide variety of approaches, it is necessary to be aware that each approach is subject to its own biases and shortcomings and may incorrectly classify interactions. By integrating data from multiple sources, triangulation provides improved confidence in the interactions, whereby interactions from different sources can be weighted based on their reliability and a confidence score can be assigned based on the number of sources that report that interaction (Szklarczyk et al., 2014). Furthermore, to reduce the likelihood of false positives and false negatives, a reliability threshold can be applied to this score to ensure that only high confidence interactions are included (Mazandu and Mulder, 2011). A protein-protein interaction network (PPIN) contains all known and predicted protein interactions in an organism or between organisms. By incorporating all physical and functional interactions with appropriate thresholds into a functional PPIN, it is possible to illustrate a more complete picture of the biological context in which the protein functions (Mazandu and Mulder, 2011).

Within a PPIN, proteins are depicted as the nodes and the interactions between them are depicted as edges. PPINs are constructed as network graphs, and, as such, concepts and algorithms from graph theory can be used to identify subnetworks of important proteins and interactions within the complete network (Mulder et al., 2014). Network representation of PPIs assists visualisation and identification of potentially biologically important proteins at a high-throughput scale. Various databases exist containing intraspecies PPIs, such as STRING (Szklarczyk et al., 2014) and Reactome (Croft et al., 2014) (for a comprehensive list refer to the

review by Rao et al. (2014)). PPINs tend to be modular in structure and exhibit a scale-free property in which many proteins are involved in few interactions, and few proteins are involved in many interactions. As such, the relative importance of a protein within the network can be determined by using centrality measures, such as how many proteins it interacts with, how often it falls on the shortest path between two other protein interactions, and the average path distance to it and every other protein.

1.4.2 Host-pathogen interactions

Host-pathogen PPINs provide an opportunity to study the complexities of the relationship between the host and the pathogen during infection (Mulder et al., 2014). Experimental detection of host-pathogen PPIs is limited, however methods in computational prediction of host-pathogen interactions are increasing (Mulder et al., 2014).

Host-pathogen interactions can be predicted using interologues, which are conserved interactions between two proteins that have orthologues that interact in another organism (Rapanoel et al., 2013). After prediction, these interactions need to be filtered so that they have biological relevance, as many of the interactions will be unlikely to occur *in vivo* (Mulder et al., 2014; Rapanoel et al., 2013). For example, Rapanoel et al. (2013) used interologues to predict interactions between *Mtb* and human proteins, which were then filtered based on differential gene expression in microarray data.

1.4.2.1 Using protein structure to predict host-pathogen protein interactions

In addition, since protein domains determine the structure and function of proteins and interactions between two proteins are mediated by these domains (Mulder et al., 2014), host-pathogen interactions have be predicted by using protein domains (Dyer et al., 2007). Human-pathogen PPIs have also been predicted based on structural and sequence similarity after applying appropriate filters for biological relevance (Davis et al., 2007; Doolittle and Gomez, 2010; Mulder et al., 2014)

1.4.2.2 Prediction of host-pathogen protein interactions using machine learning techniques

Machine learning techniques have also been used to predict human-pathogen PPIs (Mulder et al., 2014; Tastan et al., 2009). For example, Tastan et al. (2009) trained a random forest classifier to classify HIV-1 and human protein pairs as interacting or not, by using features such as human gene expression during HIV infection, as well as gene ontology features and sequence similarity.

1.4.2.3 Assessing the relevance of predicted interactions

Several computational methods have been used to predict host–pathogen interactions. However, because few known host–pathogen interactions are documented in the literature, assessing these prediction methods is difficult (Mulder et al., 2014). For example, Rapanoel et al. (2013) identified 47 experimentally observed human–*Mtb* interactions in the literature,

but there was little to no overlap between these known interactions and predicted interactions in their dataset and in a similar dataset of predicted human-*Mtb* compiled by Huo et al. (2015). Human-HIV protein interactions are far better documented due to the HIV-1 Human Interaction Database, but even interactions in this database may need to be filtered to avoid false-positives (MacPherson et al., 2010), which will be discussed in more detail in chapter 2.

1.4.2.4 Using host-pathogen interactions to understand immune responses

Knowledge of host-pathogen PPIs may help to understand how pathogens attack the host and how the host's immune system responds. For example, predicted interactions between human-*Mtb*, have been used to identify potential drug targets in *Mtb* by filtering on proteins with high network centrality measures of *Mtb* proteins in an *Mtb*-*Mtb* PPIN (Mazandu and Mulder, 2011). The *Mtb* network displays scale-free and small world properties, which make the system vulnerable against targeted attack and facilitate easy network navigation, regardless of the network size (Mulder et al., 2014). In such networks, few proteins are essential for the survival of the system, allowing proteins with high centrality measures to be considered as potential drug targets (Mulder et al., 2014).

1.5 Aims

To our knowledge, very little work has been done in which human-*Mtb* and human-HIV protein interaction networks have been combined, along with individual species PPINs. Such an approach may uncover pathways that play an important role in host responses to HIV-TB co-infection.

Mapping sequence variation onto a human-pathogen protein interaction network will enable high-throughput analyses of the impact of alleles on the individual proteins and their underlying pathways. The impact of a variant will not only be analysed by its position relative to the gene and the sequence change it confers, but also by the network properties of the proteins it maps to and interacts with. Analysing the network properties of the proteins encoded by the MHC, and the proteins that interact with pathogens along with their variants, may give novel insight into the impact of these variants on the human-pathogen interactions associated with these proteins. In particular, due to the shared genetic association with HIV and TB and immune response, focusing on human proteins encoded by the MHC region and proteins that interact with MHC-encoded proteins may improve the understanding of immune response in HIV-TB co-infected individuals.

This study aimed to investigate the potential impact that genetic variation could have on HIV and TB co-infection, and to provide resources for further analysis of human-pathogen protein interactions, as well as genetic variation in African populations.

The three broad aims of the study are listed below, along with the specific research questions and hypotheses that were formulated to address these aims.

1. **To generate and interrogate a network of functional PPIs containing human, HIV and**

Mtb proteins.

1.1. Which human proteins are likely to be facilitating interactions between HIV-1 and *Mtb* proteins, and what are the roles of these human proteins?
Hypothesis 1: The network characteristics of human MHC proteins suggest that they are more important than other human proteins in the network, and are thus likely to be facilitating interactions between HIV-1 and *Mtb* proteins.
Hypothesis 2: The human proteins that functionally interact with both HIV-1 and *Mtb* proteins are enriched for immunological functions compared to other human proteins that do not functionally interact with both pathogens.

2. **To construct a sequence graph of human chromosome 6 that will enable identification of genetic variation in the MHC region in African populations, and may act as a reference for alignment of new sequencing reads.**

2.1. Are previously unmapped reads incorporated into the assembly when using *de novo* assembly methods?
Hypothesis: Previously unmapped reads are incorporated into the reference graph.

2.2. Does the sequence graph reveal variants that are undetected by current variant calling methods?
Hypothesis 1: New variants are identified in the sequence graph that were not previously identified by calling variants against the linear reference genome.
Hypothesis 2: Fewer variants are identified when calling variants between individuals of similar ancestry in the de Bruijn graph than were previously identified by calling variants against the linear reference genome.

3. **To overlay genetic variation and gene expression data on the human-pathogen functional protein interaction network and to describe the possible impact of genetic variation on the interactions.**

3.1. Do the genes that are differentially expressed in HIV-TB co-infected individuals compared to uninfected individuals exhibit network properties that suggest they facilitate interactions between the host and pathogens?
Hypothesis: Genes that are differentially expressed in HIV-TB co-infected individuals compared to uninfected individuals have high network importance in the functional host-pathogen PPIN.

3.2. Do the human proteins that are likely to facilitate interactions between HIV-1 and *Mtb* proteins have variants in their corresponding gene sequences that could affect disease progression?
Hypothesis 1: Variants with known clinical consequences for the progression of the two diseases are found in human proteins with high network importance.
Hypothesis 2: Human proteins that functionally interact with both HIV-1 and *Mtb* proteins have additional variants that change gene structure, expression or function, which have not yet been shown to have clinical consequences.

1.6 Thesis structure

This thesis consists of five chapters. The second chapter describes the human-pathogen protein interaction network that was constructed for this thesis by combining functional PPIs between the human host proteins, and *Mtb* and HIV proteins. The third chapter presents the sequence graph of chromosome 6 that was constructed by *de novo* assembly of sequence reads from African individuals. In the fourth chapter, the outputs of the second and third chapter are integrated by presenting the host-pathogen protein interaction network overlaid with genetic variants identified in chapter two and gene expression data from HIV and TB co-infected individuals and uninfected individuals. The fifth and final chapter concludes the thesis and describes the contributions of this study, the limitations, as well as directions for future research.

2. Constructing a human-pathogen functional protein-protein interaction network

2.1 Introduction

Protein interactions make up biological pathways and processes, which, in turn, drive the functioning of biological systems. As such, the study of proteins within the context of their interaction networks provides more meaningful insight than studying proteins in isolation. In the same way that protein interactions drive the biological processes within an organism, interactions between the human and pathogen proteins are essential for the survival of the host and the viability of the pathogen during infection. In this section the available host-pathogen protein interaction networks for HIV-1 and human proteins, *Mtb* and human proteins, as well as for intra-species interaction networks will be introduced. Thereafter, network analysis as applied to the analysis of protein-protein interactions (PPIs) will be introduced before describing the aims and hypotheses.

2.1.1 Human-human PPIs

Human-Human protein-protein interactions are well defined and detailed interaction networks are publicly available (Bossi and Lehner, 2009; Croft et al., 2014; Szklarczyk et al., 2014). STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a database of the known and predicted protein interactions from five sources: (1) primary interaction databases, which contain curated experimental evidence for interactions; (2) manually curated pathway databases; (3) text-mining to identify links between proteins based on entries in abstracts and articles; (4) predicted interactions based on co-expression and genomic information; and (5) interologues (http://string-db.org). It should be noted that STRING includes protein-coding gene loci, and, as such, alternatively spliced or post-translationally modified proteins are collapsed to the gene level. Each source is allocated a confidence score, whereby interactions supported by experimental evidence have higher confidence scores. These scores are then combined into a unified score according to the following formula:

$$S = 1 - \prod_{i=1}^{n}(1 - s_i)$$

where s_i is the score from method i and n is the number of methods used. In this way, the interactions with several types of evidence have higher scores and, thus, higher confidence.

In addition to the STRING network, Bossi and Lehner (2009) constructed a network of human PPIs from 21 different sources (outlined in Table 7.1). Rapanoel et al. (2013) has previously combined the Bossi and Lehner (2009) network and STRING PPIN to supplement the STRING PPIN with additional interactions and increase the confidence in any interactions that

overlapped. Rapanoel et al. (2013) assigned scores to each of the 21 sources incorporated in the Bossi and Lehner (2009) network based on the quality of the data source, whereby data sources including experimental evidence and/or manual curation were assigned higher scores than other data sources. Thereafter, these scores could be combined into a unified score using the STRING scoring formula.

The third source of human PPIs available at the time this analysis was conducted is Reactome. Reactome (http://www.reactome.org) is a manually curated database of human pathways and reactions, which sources interactions from peer-reviewed research articles (Croft et al., 2014). In this analysis, these three human PPINs will be combined.

2.1.2 HIV-1-human PPIs

HIV-1 and human protein interactions are probably the most well documented host-pathogen interactions due to NCBI's HIV-1 human protein interaction database (HHPID) (http://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/) (Ako-Adjei et al., 2014). The HHPID is a centralised and categorised resource for all publications citing associations between HIV-1 and human proteins, which has been used, for example, to identify potential drug targets (Li et al., 2013). Each set of publications in the database is annotated with a set of keywords describing an interaction, for example 'binds to' or 'induces acetylation of'. However, within the HHPID, an interaction between the same pair of HIV-1 and human proteins may be recorded multiple times in the database, with different citations or different keywords.

Despite the obvious value of having a highly detailed, centralised database of interactions between HIV-1 and human proteins, the resource is not void of caveats. MacPherson et al. (2010) point out three prominent problems with the dataset: (1) the wide range of publications, in terms of date range, methods used, and varied data quality, makes it difficult to estimate the number of false-positive interactions recorded; (2) the HHPID is manually curated, which lends itself toward inconsistencies; (3) the interactions are largely redundant, as the human and HIV-1 protein pair is not a unique identifier for an interaction. In order to account for these problems when using the HHPID, MacPherson et al. (2010) organised the interaction types provided by the HHPID into a hierarchy. The hierarchy divided interactions according to the following categories: regulatory interaction, gene regulation, protein regulation, physical interaction, binding interaction, degradation event, and modification event. Uninformative interactions are defined as interactions with 'interacts with' as the only keyword or interaction description. This categorisation and filtering of interactions was identified as a useful technique for the present study.

2.1.3 *Mtb*-human PPIs

Human and *Mtb* interactions are poorly documented in comparison to HIV-human PPIs. Due to the limited availability of experimentally verified host-pathogen interactions, computational prediction of host-pathogen interactions has been employed (Mulder et al., 2014). At least two separate groups have predicted human and *Mtb* interactions using

computational methods (Huo et al., 2015; Rapanoel et al., 2013). The network constructed by Huo et al. (2015) contains predicted interactions between *Mtb* lab strain H37RV and human proteins. They first identified host-pathogen interactions by using the *Basic Local Alignment Search Tool* (*BLAST*) (Altschul et al., 1990) to identify similar sequences in human and *Mtb* to a set of sequences of interacting proteins derived from HPIDB and BIPS. Thereafter, potentially interacting proteins were filtered where there was a domain-domain interaction prediction. Finally, the credibility of the interaction was assessed based on the UniProt protein annotation (i.e. the subcellular location, tissue specificity, and GO ontology).

The interactions predicted by Rapanoel et al. (2013) were predicted for the *Mtb* clinical isolate CDC1551 (Oshkosh), by means of interologues. Human-pathogen interactions identified using intraspecies interologues were filtered based on whether they referred to known interactions (from the literature) or whether they were expressed during infection (identified from microarray data) (Rapanoel et al., 2013). In addition, experimentally verified interactions between human and bacteria were filtered based on whether the bacterial protein had an orthologue in *Mtb*, and whether both proteins where differentially expressed in their microarray dataset (Rapanoel et al., 2013). This set of interactions was later used to identify potential drug targets by overlaying *Mtb* proteins that have important network properties in the *Mtb*-*Mtb* PPIN on the human-*Mtb* PPIN, demonstrating an important use case for host-pathogen PPINs (Mulder et al., 2014).

2.1.4 Drug-target interactions

When analysing host-pathogen interactions, it is valuable to consider the impact that treatment with drugs could have on these interactions. DrugBank is a drug and drug target database that serves as a bioinformatics resource for drug and target analyses (Wishart et al., 2006). DrugBank contains comprehensive molecular information about drugs, mechanisms of action and drug-target interactions, including protein, DNA and RNA targets.

2.1.5 Network analysis of PPINs

Network analysis is a way of studying the relationships between individual and groups of points in large datasets by representing the points as nodes and the relationships between them as edges connecting the nodes. In the case of PPIN analysis, the network is represented as an undirected graph, $G(V, E)$, in which V is a set of nodes (proteins) and E is a set of edges that connect the proteins (the protein or drug-target interactions). The relative importance of nodes and edges in a network can be determined through calculations of measures that describe volume, distance, and mediation (Borgatti and Everett, 2006). Four frequently used centrality measures are degree, closeness, betweenness and shortest path distance. These measures will be introduced, along with a fifth, and less frequently used, measure called bridging centrality.

2.1.5.1 Degree centrality

The degree centrality measures the number or volume of nodes that a given node is connected to. In the case of PPINs, the degree of the protein indicates how many proteins it interacts with. The formula for calculating degree is:

$$D(p) = \sum_{i=1}^{n} u(p, p_i) \text{ for } i = 1, 2, ...n,$$

where $u(a_1, a_i)$ is the Kronecker Delta function, and p is a protein in the network G and n is the total number of proteins in the network. This formula can be normalised by dividing it by the maximum possible degree $n - 1$.

$$D'(p) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} u(p, p_i) \text{ for } i = 1, 2, ...n$$

2.1.5.2 Shortest path distance

The shortest path distance between nodes is the minimal amount of edges traversed to reach a destination node from a given node. In protein terms, it is how many proteins a source protein must interact with to interact with a target protein.

2.1.5.3 Closeness centrality

Another measure of distance is closeness, which is defined as the total distance from a given node to every other node (Freeman, 1979). It is a measure of how quickly information is transmitted through a network (Valente et al., 2008). The closeness centrality is calculated as the reciprocal of the sum of the shortest paths from a given node to all other nodes, and normalised by the sum of the minimum possible distances, $n - 1$. In this way, interpretation of the centrality measure follows the same pattern as the others - that is that the higher the closeness, the nearer the node is to other nodes in the network. The formula is as follows:

$$C(p) = \frac{n-1}{\sum_{q=1}^{n-1} d(q, p)},$$

where $d(p, q)$ is the shortest path distance between proteins p and q, and n is the number of proteins in the network.

2.1.5.4 Betweenness centrality

In order to measure a node's importance for mediation, the betweenness centrality has been defined (Freeman, 1979). Betweenness (of nodes) is the number of shortest paths between any two nodes that need to pass through the given node. As such, it can be described as a

measure of potential influence of a node on both direct and indirect pathways in a network (Valente et al., 2008). The betweenness centrality of a node p is calculated as the sum of the fraction of all the pairs of shortest paths that pass through a protein p:

$$B(p) = \sum_{s,t \epsilon G} \frac{\sigma(s,t|p)}{\sigma(s,t)},$$

where G is the set of all proteins, $\sigma(s,t)$ is the number of shortest paths between any proteins s and t, and $\sigma(s,t|p)$ is the number of shortest paths between any proteins s and t that pass through p, and $s \neq p$. This formula is normalised to fall within a range of 0 to 1 as follows:

$$B(p) = \frac{2}{(n-1)\cdot(n-2)} \cdot \sum_{s,t \epsilon G} \frac{\sigma(s,t|p)}{\sigma(s,t)},$$

where n is the number of proteins in the network.

2.1.5.5 Bridging centrality

High degree nodes tend to have high betweenness, because the nodes with the highest frequency tend to fall in highly connected, core areas of the network resulting in the nodes having several shortest paths passing through them (Hwang et al., 2006). Betweenness is commonly used as a measure for how important a node is for information flow between other nodes; however, as explained above, it is really a measure of global importance. To measure the local importance of a node, bridging centrality was proposed by Hwang et al. (2006) to identify nodes that are important for bridging submodules of the network. Bridging nodes, unlike nodes with high values for other centrality measures, can cause network disruption without dismemberment (Hwang et al., 2008). Hwang et al. (2008) define a bridge as a node or an edge that connects modules in a graph, and the bridging centrality of a node as the product of its global importance and its bridging coefficient. The global importance of a node or edge is calculated as its betweenness, and the bridging coefficient of a node p is the average probability of leaving the set of nodes directly connected to p. The bridging coefficient of an edge is the product of the weighted average of the bridging coefficients for the two nodes whose connection creates the bridge, and the reciprocal of the number of common direct neighbour nodes of the two nodes. In mathematical terms, the bridging coefficient is calculated as follows:

$$BC(p) = \frac{D(p)^{-1}}{\sum_{i \epsilon N(p)} \frac{1}{D(i)}},$$

where $D(p)$ is the degree of a node p, and $N(p)$ is the set of the neighbours of a node p. Simplified, the bridging centrality of a node p is:

$$Br(p) = BC(p) \times B(p),$$

where $B(p)$ is the betweenness centrality and $BC(p)$ is the bridging coefficient of a node p.

Nodes with high bridging centrality will be more important in the network and connect more densely connected modules to one another. In addition, deletion of a node with high bridging centrality will cause similar disruption to the path length distribution as deleting a node with high betweenness. It will also result in fewer singleton nodes created as the average size of the isolated modules will be larger than betweenness (Hwang et al., 2008).

As this study is interested in the human proteins that play an important role in HIV-TB co-infection, the bridging centrality was adjusted to calculate a new measure, *"pathogenicity bridging centrality"*. This will be described in the methods section.

2.1.6 Interpreting proteins and PPIs using gene ontology enrichment analysis

Gene ontology (GO) terms are a defined and structured vocabulary used to describe a gene or protein's molecular function (MF), biological process (BP), and cellular component (CC). GO term enrichment analysis searches for similar or related GO terms describing a set of genes relative to the rest of the genes in the genome. GO term analysis can be used to complement analysis of interacting proteins by helping to identify biologically relevant interactions, as well as improve the interpretation of the interactions. DAVID, the Database for Annotation, Visualisation and Integrated Discovery, offers a tool for functional annotation (Huang et al., 2008, 2009, 2007). Using DAVID, it is possible to cluster lists of genes based on the similarity of the GO terms that those genes share. If a set of genes share similar properties, such as processes or cellular component, it is possible that they are involved in similar pathways. The algorithm used by the functional annotation tool within DAVID uses *kappa* statistics, a chance-corrected measure of co-occurrence between two categorised data sets, to determine the similarity of terms (Huang et al., 2007). For each pair of genes in the set, it computes pair-wise comparisons of a given terms to all the other terms and sorts the similarity, resulting in a *kappa* score between 0 and 1 (Huang et al., 2007). For a pair of genes, a *kappa* score greater than 0.7 indicates strong similarity. Thereafter DAVID applies a clustering algorithm to identify clusters of related genes. To determine the relative importance of the identified clusters, a modified Fishers Exact score is calculated and multiple testing correction applied (Huang et al., 2007).

2.1.7 Aim and hypotheses

The aim of this chapter was to generate and interrogate a network of functional PPIs containing human, HIV-1, and *Mtb* proteins that may yield insight into HIV-TB co-infection and provide a resource for future analysis. To this end, a network of functional inter- and intraspecies PPIs was constructed for human and it's often co-infecting pathogens HIV and *Mtb* using intraspecies PPIs from STRING (Szklarczyk et al., 2014) and human-pathogen PPIs from the aforementioned resources (Ako-Adjei et al., 2014; Huo et al., 2015; Rapanoel et al., 2013). We hypothesise that (1) the human MHC proteins have network centrality measures that suggest they are more important than other human proteins for facilitating interactions between HIV-1 and *Mtb*, and (2) proteins with higher importance in the network are enriched for immunological functions compared to other human proteins. By interrogating the produced network based on the network properties of the proteins and the biological

importance of the interactions, a subnetwork of proteins was identified that may yield insight into HIV-TB co-infection.

2.2 Methods

2.2.1 Human-human PPIN dataset

Before generating the human PPIN, a non-redundant human proteome was generated, which contains one protein per gene identifier. In order to do this, the human proteome was downloaded from UniProt (Consortium et al., 2017) (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebasereference_proteomes/Eukaryota/UP000005640_9606.gene2acc.gz). The 91 618 proteins and protein fragments in this dataset were entered into UniProt's web-based tool for identifier mapping, in order to annotate the proteins with the gene name, gene description, and gene ontology for biological process, molecular function, and cellular component. This resulted in 69 693 annotated proteins and protein fragments. For each gene in the UniProt proteome, if it mapped to a protein that had been reviewed, the protein was added to the non-redundant set and any unreviewed proteins it mapped to were excluded. If the gene only mapped to unreviewed proteins, it was also added to the list, without any indexing suffixes (e.g. '-1'), which are assigned to protein fragments. This resulted in a non-redundant human protein set containing 21 712 proteins. The human PPIN was constructed by combining the interactions from three datasets, namely STRING (Szklarczyk et al., 2014), Reactome (Croft et al., 2014), and the interaction network constructed by Bossi and Lehner (2009).

2.2.1.1 Interactions from STRING

STRING version 10 of the human protein interaction network was downloaded from the user interface of STRING (http://string-db.org/newstring_download/protein.links.detailed.v10/9606.protein.links.detailed.v10.txt.gz, Accessed: 2015-08-12). The STRING human PPIN contained 8 548 002 interactions between 19 247 proteins (with the proteins labelled with Ensembl protein identifiers). Thereafter, the Ensembl protein identifiers provided by STRING were mapped to their equivalent UniProt identifiers in the non-redundant set using the identifier mapping file provided by UniProt (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640_9606.idmapping.gz). This resulted in 7 510 084 interactions between 17 650 proteins.

2.2.1.2 Interactions from Bossi and Lehner (2009)

To supplement the STRING PPIN with additional interactions or evidence, the interaction network constructed by Bossi and Lehner (2009) was downloaded. This interaction network contains interactions from 21 different sources (outlined in Table 7.1), which together include 80 992 physical interactions between 10 229 human proteins. The interactions are recorded using Ensembl protein identifiers, which were mapped to UniProt identifiers in the

non-redundant set using the same mapping file that was used for the STRING interactions. This resulted in 75 221 interactions between 9771 proteins. Each of the data sources was assigned a score based on the quality of the data source, whereby data sources including experimental evidence and/or manual curation have higher scores than other data sources. Scores were assigned following the methods of Rapanoel et al. (2013). These scores were combined into a unified score using the STRING scoring formula.

2.2.1.3 Interactions from Reactome

The interactions from Reactome in the above network were updated with the most recent version of the database (version 46). Reactome (http://www.reactome.org) is a manually curated database of human pathways and reactions. The version downloaded contains 2 134 007 reactions between 5550 human proteins. After mapping to non-redundant UniProt IDs, this network contained 1 953 815 interactions between 6378 proteins. Reactome acquired this data from research articles in PubMed. Because the database is manually curated, protein interactions acquired from this database were assigned a high score as observed in Table 2.1.

2.2.1.4 Combining the sources of interactions

The three data sources were combined into a single non-redundant human-human PPIN by combining the unified STRING score with the scores that Rapanoel et al. (2013) used to score the network by Bossi and Lehner (2009) and Reactome, ensuring that the Reactome score was only counted once. It is generally accepted that interactions with scores greater than or equal to 0.7 have high confidence Mazandu and Mulder (2011). The network was thus filtered on interactions with scores greater than or equal to the aforementioned threshold in order to ensure that any downstream predictions made based on the network are supported. The final human PPIN contained 407 996 interactions between 15 774 proteins.

2.2.2 *Mtb*-*Mtb* PPIN dataset

A *Mtb*-*Mtb* PPIN was constructed for the strain CDC1551 using interactions from STRING and interactions predicted by shared domains (domain-domain interactions) using data from InterPro. Confidence scores for STRING interactions were extracted from STRING, while those for shared domains were computed using the methodology described in Mazandu and Mulder (2011). The string network contained 487 440 interactions between 4163 proteins. The CDC1551 proteome was downloaded and annotated using UniProt, and the network was filtered to only contain non-redundant proteins. The datasets were formatted, combined and scores calculated using *Python* code written by Mazandu and Mulder (2011). Once filtered on high confidence PPIs, the final set of *Mtb*-*Mtb* PPINs contained 15 172 high confidence interactions between 3364 proteins.

2.2.3 *Mtb*-human PPIN dataset

The human-*Mtb* PPIN was constructed by combining two datasets of recently published interactions between *Mtb* and human (Huo et al., 2015; Rapanoel et al., 2013). The

Table 2.1 Scoring of the protein interaction data sources used by Bossi and Lehner (2009)

Source	Source Description	Score
Reactome	Manually curated by experts	0.95
CORUM	Manually curated, experimentally verified mammalian complexes (excluding high-throughput data)	0.9
PC/Ataxia	Stringent Yeast two-hybrid, acquired from literature	0.85
HPRD	Manually curated, including experimental evidence and Y2H	0.85
INTACT	Interactions from literature and direct submissions	0.85
MDC	Y2H and verification	0.85
Unilever	Y2H, mass spec and literature	0.85
BIND	Literature curated	0.85
BIOGRID	Literature curated	0.85
CCSB-curated	Stringent Y2H	0.78
DIP	Experimental evidence and manually curated by experts	0.85
MIPS	Manually curated by experts	0.85
IntNetDB	PPI predictions from data integration	0.75
BIOVERSE	Data integration	0.75
OPHID	Known, experimental, and predicted PPIs	0.75
HOMOMINT	Human orthologues of experimentally verified PPIs in model organisms	0.7
Co-citation	Co-citation	0.65
Co-expression	Co-expression	0.65
Interologue	Inferred from interactions between orthologues in another organism	0.7
Ottowa	Affinity purification mass spectrometry	0.7
Transcription complexes	Affinity purification mass spectrometry	0.7

interactions in the network generated by Rapanoel et al. (2013) were predicted for the *Mtb* clinical isolate CDC1551 (Oshkosh) by means of interologues, using orthologue data from Integr8 (Pruess et al., 2005), and intraspecies interactions from the Database of Interacting Proteins (DIP) (Xenarios et al., 2000). Rapanoel et al.'s (2013) interactions are divided amongst four datasets: (1) interologue-DIP-array, which contains 78 interactions between 35 and 47 human and *Mtb* interologues respectively, in which both proteins are differentially expressed in microarray analysis during infection; (2) interologue-DIP-known, which contains three interactions between three human and three *Mtb* interologues respectively that have been reported in the literature; (3) interologue-HPI-array, which contains 109 interactions between 85 and 53 human and *Mtb* interologues respectively, in which both proteins are differentially expressed in microarray analysis during infection; and (4) known-interactions, which contains 47 interactions between 14 and 25 human and *Mtb* proteins respectively that have been reported in the literature. For this analysis, these four datasets were combined into a single dataset and scored using the same scoring as described in Table 2.1, whereby interologues received a score of 0.7, DIP received a score of 0.85, known interactions received a score of 0.85, array received a score of 0.65 and domain-domain interactions received a score of 0.65. The scores were combined using the STRING scoring formula.

The network constructed by Huo et al. (2015) contains predicted interactions between *Mtb*

lab strain H37RV and human proteins. The H37RV protein identifiers were mapped to their orthologues in CDC1551 using Tuberculist (Lew et al., 2011). Thereafter, the CDC1551 gene names were mapped to their corresponding UniProt identifiers using UniProt's identifiers mapping tool. Since these interactions were predicted using domain-domain interactions, they were assigned a score of 0.65. This PPIN contained 118 interactions between 43 human proteins and 48 *Mtb* proteins.

The two *Mtb*-human PPINs were combined resulting in 339 interactions between 172 human proteins and 136 *Mtb* proteins. It should be noted that because the interactions had already undergone filtering for biological relevance, these *Mtb*-human interactions were not filtered based on high-confidence scores.

2.2.4 HIV-human PPIN dataset

The HIV-1 and human protein interaction dataset was downloaded from the HHPID. The network contained 16 141 interactions between 23 HIV-1 protein fragments and 3576 human protein identifiers. This network makes use of RefSeq identifiers, which were subsequently mapped to their corresponding UniProt identifiers using the relevant ID mapping files from UniProt (for HIV-1: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Viruses/UP000002241_11706.idmapping.gz, for human: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640_9606.idmapping.gz). The HIV-1 proteome contained 10 proteins. The HIV-1 proteins were annotated using UniProt, to extract the gene name, gene description, and gene ontology for biological process, molecular function, and cellular component. Once these were mapped, the results were filtered to only contain human proteins in the non-redundant set. This resulted in a network containing 3972 interactions between 2548 human proteins and 7 HIV-1 proteins. It should be noted that the drastic decrease in interactions was primarily due to the observation that, within the HHPID, an interaction between the same pair of HIV-1 and human proteins may be recorded multiple times in the database, with different citations or different keywords.

2.2.4.1 Categorising HIV-human interactions to filter the HHPID

The interactions were categorised using the hierarchy suggested by MacPherson et al. (2010), which was collapsed into seven categories for our network, as depicted in Figure 2.1. The categories are regulatory interaction, gene regulation, protein regulation, physical interaction, binding interaction, degradation event, and modification event. Uninformative interactions are defined as interactions with 'interacts with' as the only keyword. For each interaction, the number of publications was counted and added as a score, which could be used as a crude measure of confidence. In addition, a binary score column was added for each of the seven possible categories, where a score of 0 in the physical interaction column means that the interaction has no evidence of being physical, while a score of 1 would indicate evidence of a physical interaction between the two proteins. The categories were used to identify a high confidence set of interactions, by restricting to interactions that were reported by at least two publications, and were annotated as being a physical interaction, binding interaction or

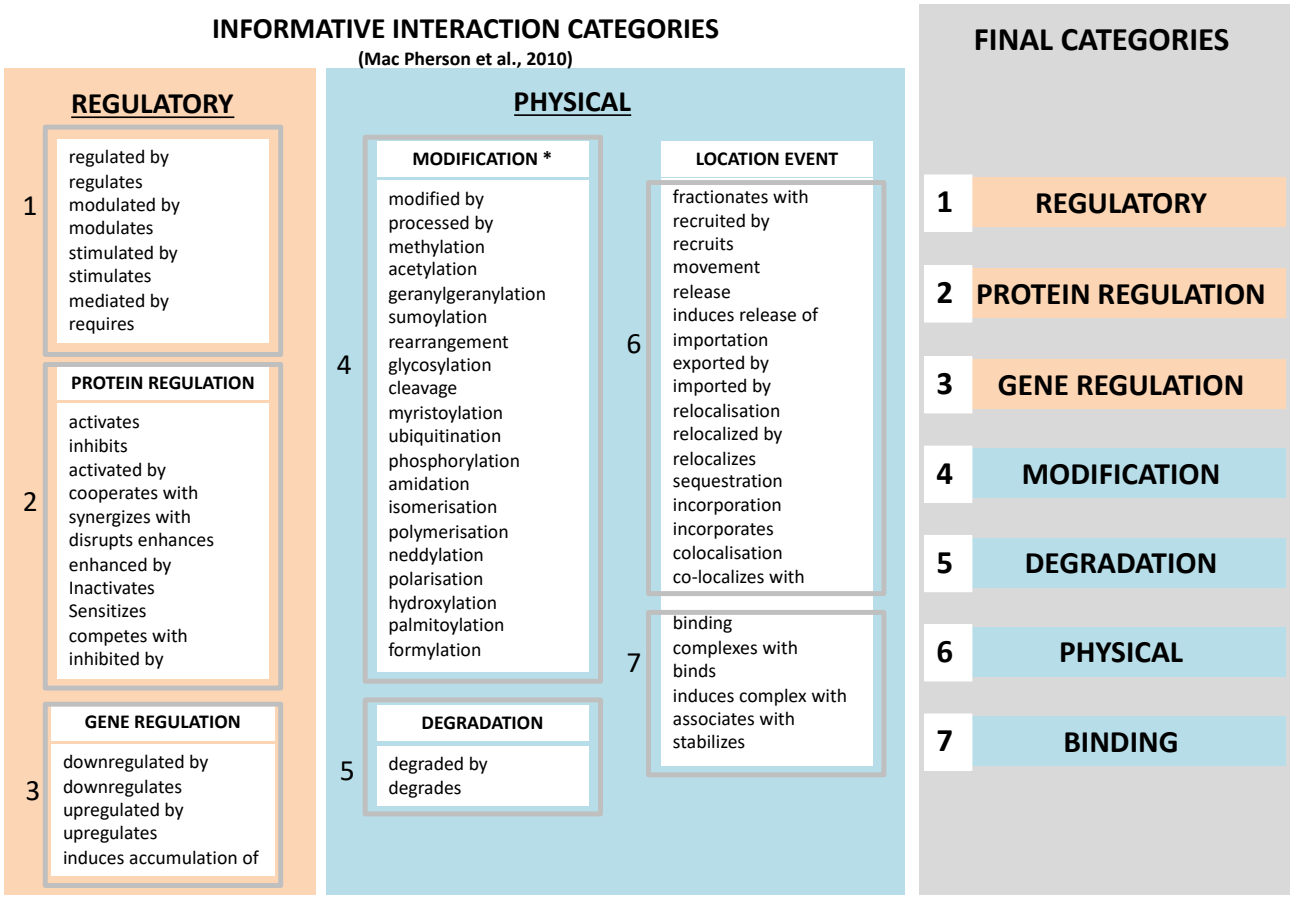


Figure 2.1 Categories used to classify interactions recorded in the HHPID

protein regulation interaction. This resulted in 890 interactions between 701 human and 7 HIV-1 proteins.

2.2.5 Drug-target interactions

In order to enable analyses with practical application to understanding HIV and TB co-infection, drug-target interactions were included in the network. A list of anti-HIV and anti-TB drugs was compiled using the World Health Organisation's Anatomic Therapeutic Chemical (ATC) classification system for drugs. The final list contained 16 Anti-TB drugs and 20 Anti-HIV drugs, of which all the Anti-TB drugs are currently used in South Africa, while 10 of the Anti-HIV drugs are used in South Africa (National Department of Health, 2015). The interactions between these drugs and their targets were ascertained by querying DrugBank (V4.3), a drug and drug target database that serves as a bioinformatics resource for drug and target analyses (Wishart et al., 2006). In total, 239 interactions between the 36 drugs and 95 targets were included. Using this resource, it was possible to determine the target gene name, UniProt identifier, organism, whether the interaction was reported to have pharmacological action, as well as the type of target for each drug of interest. The target type could be a target, enzyme, transporter or carrier (coded as 1, 2, 3, and 4 respectively). The pharmacological action could be yes, no or unknown (coded as 1, 2, and 3 respectively). The target action could be inhibitor, inducer, substrate, or any combination of the 3 (coded as 1, 2,

3, 4 for inhibitor and inducer, 5 for substrate and inducer, 6 for substrate and inhibitor, and 7 for substrate and inhibitor and inducer). These codes were used as score types for the interaction so that one could filter the interactions based on a particular target type or target action. When the target provided by DrugBank was non-human, or not belonging to HIV-1 or *Mtb*, if an orthologue could be found using UniProt this target was included, otherwise it was excluded. The drug names are listed in Table 2.2 and Table 2.3, for TB and HIV respectively. The drug and target pairs included in the analysis are listed as supplementary tables in Table 7.2 and Table 7.3, for TB and HIV respectively.

Table 2.2 Anti-TB drugs included in the analysis

Generic name	Category	Used in South Africa	DrugBank	ATC Code
Isoniazid	first line	Y	DB00951	J04AC01
Rifampicin	first line	Y	DB01045	J04AB02
Pyrazinamide	first line	Y	DB00339	J04AK01
Streptomycin	first line	Y	DB01082	J01GA01
Ethambutol	first line	Y	DB00330	J04AK02
Ethionamide	second line	Y	DB00609	J04AD03
Kanamycin	second line	Y	DB01172	J01GB04
Aminosalicylic Acid	second line	Y	DB00233	J04AA01
Ofloxacin	second line	Y	DB01165	J01MA01
Levofloxacin	second line	Y	DB01137	J01MA12
Amikacin	second line	Y	DB00479	J01GB06
Cycloserine	second line	Y	DB00260	J04AB01
Clofazimine	second line	Y	DB00845	J04BA01
Moxifloxacin	third line	Y	DB00218	J01MA14
Imipenem	third line	Y	DB01598	
Linezolid	third line	Y	DB00601	J01XX08

Table 2.3 Anti-HIV drugs included in the analysis

Generic name	Category	Used in South Africa	DrugBank	ATC Code
Abacavir	NRTI	Y	DB01048	J05AF06
Atazanavir	PI	Y	DB01072	J05AE08
Efavirenz	NNRTI	Y	DB00625	J05AG03
Emtricitabine	NRTI	Y	DB00879	J05AR06
Lamivudine	NRTI	Y	DB00709	J05AF05
Lopinavir	PI	Y	DB01601	J05AR10
Nevirapine	NNRTI	Y	DB00238	J05AG01
Ritonavir	PI	Y	DB00503	J05AE03
Tenofovir	NRTI	Y	DB00300	J05AF07
Zidovudine	NRTI	Y	DB00495	J05AF01
Darunavir	PI	N	DB01264	J05AE10
Elvitegravir	II/FI/EI	N	DB09101	J05AX11
Enfuvirtide	II/FI/EI	N	DB00109	J05AX07

Continued…

Generic name	Category	Used in South Africa	DrugBank	ATC Code
Etravirine	NNRTI	N	DB06414	J05AG04
Fosamprenavir	PI	N	DB01319	J05AE07
Maraviroc	II/FI/EI	N	DB04835	J05AX09
Raltegravir	II/FI/EI	N	DB06817	J05AX08
Rilpivirine	NNRTI	N	DB08864	J05AG05
Saquinavir	PI	N	DB01232	J05AE01
Tipranavir	PI	N	DB00932	J05AE09

Notes. II/FI/EI=Integrase inhibitors, fusion inhibitors, and entry inhibitors; NRTI=nucleoside/nucleotide reverse transcriptase inhibitors; NNRTI=non-nucleoside reverse transcriptase inhibitors; PI=protease inhibitors

2.2.6 Human-pathogen PPIN dataset

Finally, the four aforementioned PPINs and Drug Target interactions were combined into a single human-pathogen interaction network, containing 424 397 functional interactions between 19 155 proteins, and 239 drug-target interactions between 36 drugs and 95 protein targets. This contains a subset of 15 774 human proteins, 3370 *Mtb* proteins and 11 HIV-1 proteins. There are 36 drugs included in the network. The PPIN is comprised of 407 996 human-human PPIs, 339 *Mtb*-human PPIs, 890 HIV-human PPIs, 15 172 *Mtb*-*Mtb* PPIs, and 239 drug-target interactions.

2.2.7 Network analysis

The aforementioned centrality measures of betweennes, closeness, degree and bridging centrality were calculated using *NetworkX*, a *Python* language software package for complex network analysis (Hagberg et al., 2008). Betweenness was calculated with the betweenness_centrality function, degree with the degree_centrality function, and closeness with the closeness_centrality function (Hagberg et al., 2008). These centrality measures were calculated for the human-human PPIN and *Mtb*-*Mtb* PPIN separately. This was done to separate the importance of proteins within the organism's network from their importance between the networks. In addition, there is no HIV-HIV protein interaction network as HIV relies on interactions with host proteins in order to function.

In addition to these measures, the shortest path distance and average shortest path distance to MHC proteins, drugs, and pathogens was calculated using *NetworkX*. The function shortest_path was used on each protein against every protein in the set of targets. The shortest distance to every target protein between which a path existed was calculated and added to a list, from which both the minimum and average shortest path distance to proteins in the target set could be determined. This gives an indication of how many proteins a particular protein needs to interact with before interacting with any protein in one of the respective target groups.

2.2.7.1 Pathogenicity bridging centrality

As this study is interested in the human proteins that play an important role in HIV-TB co-infection, a *"pathogenicity bridging centrality"* measure was proposed. The measure consists of three mutually exclusive subgroups of nodes: start-points, end-points, and connectors. As the network is an undirected graph, the start-points and end-points are interchangeable. In this case the start-points are viral proteins, end-points are bacterial proteins, and the connectors are human proteins. The pathogenicity bridging centrality measures how important a human protein is for connecting viral and bacterial proteins. It is defined as the product of the betweenness between two distinct subsets of nodes and the bridging coefficient of a node:

$$PBr(p_h) = \sum_{v,b \epsilon G_{path}} \frac{\sigma(v,b|p_h)}{\sigma(v,b)} \times BC(p_h),$$

where p_h is a human protein, v is a viral protein, b is a bacterial protein, and G_{path} is the set of all human-human and human-pathogen interactions.

The first part of the pathogenicity bridging centrality was calculated using the *NetworkX* function betweenness_centrality_subset, with the sources set to be the HIV proteins, the targets set to be the *Mtb* proteins. The calculation was normalised (Hagberg et al., 2008). The rest of the pathogenicity bridging centrality was calculated using the degree function of *NetworkX* in a *Python* script and multiplied by the betweenness_centrality_subset.

2.2.8 Protein annotation

In addition to annotations regarding gene names and gene ontology obtained from UniProt, BioMart was used in order to annotate the gene's chromosomal location and position within the chromosome on GRCh38 (Smedley et al., 2015). Using this annotation, it was possible to determine which proteins were within the MHC region (chromosome 6, position 28 500 000 to 33 490 000). An additional column 'Region' was added to the annotations file, in which proteins within the MHC region where labelled as MHC. In addition, the network properties were added to the annotation file. Proteins were also annotated with which pathogens they interact with: 'NONE', 'HIV and MTB', 'HIV' or 'MTB'.

Human and *Mtb* proteins were annotated with betweenness, closeness, and degree within their intraspecies network and within the complete network, whereas HIV-1 proteins were only annotated with the three network properties in the complete network as there are no intraspecies interactions. Proteins were also annotated with the pathogenicity bridging centrality measure. The centrality measures were not normalised in the annotation file as this would narrow the range of visible variability within the measures on the visualisation tool. Furthermore, fields for rank within the network were added to the annotation, for which the upper quartile and 95^{th} percentile was calculated for each measure (both in the inter- and intraspecies PPINs). Proteins that fell within the range of either of these thresholds and the maximum were labelled as 'upper quartile' or '95^{th} percentile' respectively. Intuitively, falling

above the '95^{th} percentile implies that the result lies above the upper quartile and, as such, proteins were not labelled twice. It should be noted that because many of the proteins had a pathogenicity bridging centrality of zero, as there are few human-pathogen interactions, zero values were excluded when calculating the upper quartile and 95^{th} percentile for this measure.

2.2.9 Statistical methods for comparing network properties

In order to compare significant differences between the network properties for subsets of proteins, a Wilcoxon rank-sum test was performed. The Wilcoxon rank-sum test is a non-parametric alternative to the two-sample t-test that is less sensitive to outliers (Wilcoxon, 1945). The null hypothesis of the test is that the distribution of measurements between two samples is identical. The Wilcoxon rank-sum test looks for shifts in location by assigning a rank to each of the observations of the combined sample, followed by summing the ranks for each sample and comparing the sum of the ranks. The *p* value is calculated by comparing the summed ranks to the distribution of ranks if the null hypothesis were true. This test was performed as a two-tailed test, as no *a priori* assumptions were made. A *p* value ≤ 0.05 was considered significant. To determine which sample had measurements that were greater or less than the other sample, the medians were calculated for each sample. The test was executed using the built in *Python* function Scipy.stats.ranksums, which takes two arrays (one for each sample) as input and returns the two-sided *p* value of the test.

2.2.10 GO term enrichment using DAVID

GO term enrichment analysis was performed on the 28 human proteins that were identified to functionally interact with at least one protein from both pathogenic species compared to the entire human genome using the functional annotation tool provided by the Database for Annotation, Visualisation and Integrated Discovery (DAVID) v6.7 (Huang et al., 2008, 2009, 2007).

2.2.11 GO semantic similarity comparisons of interacting pairs of human and pathogen proteins

Using the 28 human proteins that were identified to interact with at least one protein from both pathogenic species, the GO terms for MF, BP, and CC were compared for semantic similarity using an *R* package called GOSemSim (Yu et al., 2010b). For each of MF, BP, and CC, and for each trio of human, HIV, and *Mtb* proteins, every human protein's GO term was compared to every pathogen's GO term and every HIV GO term was compared to every *Mtb* GO term.

This was particularly important for cellular component, as if any of the cellular components were highly similar this may indicate that the interaction is taking place. It should be noted that using semantic similarity for terms describing different species was not useful, and most of the results were ignored. Instead, related biological processes and cellular localisations were searched manually. For example, the cellular location 'host-cell cytoplasm' for a HIV

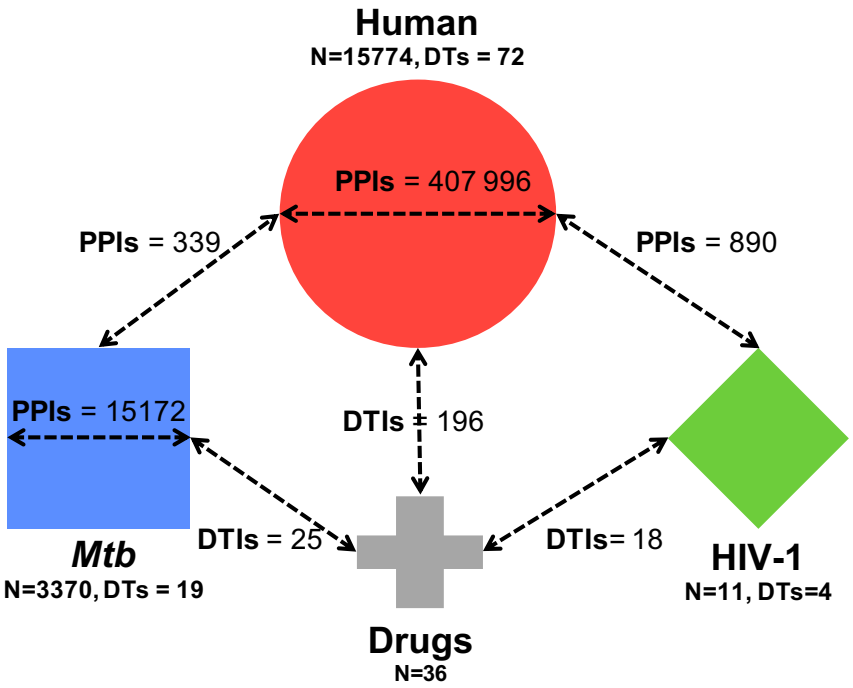


Figure 2.2 Counts of proteins and interactions in the human-pathogen PPIN.

gene received a low score when compared to the cellular location of cytoplasm in a human protein, but this represents a meaningful similarity in this context.

2.3 Results

The final human-pathogen PPIN contained 424 397 functional protein interactions and 239 drug-target interactions. The network contained 15 774 human proteins, and 407 996 functional human-human PPIs. There were 3370 *Mtb* proteins, and 15 172 functional *Mtb*-*Mtb* PPIs and 339 *Mtb*-human PPIs. There were 11 HIV-1 proteins, of which seven were involved in the 890 HIV-1 and human PPIs, while four were unreviewed proteins that were also listed as drug targets. The network contained 239 drug-target interactions between 36 drugs and 95 targets. 72 of the targets were human proteins, 19 of the targets were *Mtb* proteins, and four of the targets were unreviewed HIV-1 proteins. A detailed diagram of the protein counts and interactions is presented in Figure 2.2.

2.3.1 Comparison of the network properties of MHC proteins with other proteins in the network

Out of the 15 774 human proteins in the final network, 145 were within the MHC region. Of these 145, 15 fell within the upper quartile for pathogenicity bridging centrality. This included 10 HLA class II proteins (HLA-DRB1, HLA-DQA1, HLA-DQA2, HLA-DRA, HLA-DQB2, HLA-DRB5, HLA-DRB1*1, HLA-DRB1*14, HLA-DPA1, HLA-DPB1); one HLA class I protein (HLA-A); DAXX (Death domain-associated protein 6); SYNGAP1 (Synaptic Ras

GTPase-activating protein 1); VARS (Valine–tRNA ligase); and LSM2 (Small nuclear ribonuclear protein D homologue). Three fell within the 95^{th} percentile for pathogenicity bridging centrality. This included one HLA class I protein (HLA-C), one HLA class II protein (HLA-DRB1*16), and TNF-α (tumor necrosis factor α). There were no predicted interactions between MHC proteins and *Mtb* proteins; however, 23 MHC proteins were involved in a physical, binding or protein regulation interaction with at least one HIV-1 protein.

In order to test whether proteins within the MHC region are significantly more important in the network relative to proteins that aren't in this region, a Wilcoxon rank-sum test was performed on the proteins (see Table 2.4). Betweenness and closeness were significantly lower in MHC proteins compared to non-MHC proteins, although only marginally so ($p<0.047$ and $p<0.002$ for betweenness and closeness respectively). The median degree and pathogenicity bridging centrality was higher for MHC proteins than non-MHC proteins, but not significantly so. This suggests that MHC proteins, although biologically important, are not in their own right significantly more or less important in the high confidence network than other human proteins. In addition, the average shortest path distance and minimum shortest path distance to HIV and *Mtb* proteins were not significantly lower in MHC proteins compared to non-MHC proteins, suggesting that the MHC proteins themselves are not significantly more important for connecting the pathogens than non-MHC proteins in the network.

Table 2.4 Wilcoxon rank-sum test comparing network properties of MHC and non-MHC proteins.

Network property	MHC proteins (n=145) median (IQR)	Non-MHC proteins (n=15 629) median (IQR)	S	p-value
Betweenness	1201.06 (10.72, 8169.28)	1754.71 (51.19, 14 386.97)	1.99	0.05
Closeness	0.27 (0.26, 0.29)	0.28 (0.26, 0.31)	3.03	0.002
Degree	23 (3, 91)	16 (4, 58)	-0.91	0.36
Pathogenicity Bridging	9.476 e-06 (0, 0.008)	0 (0, 0.005)	-1.62	0.11
Minimum distance to *Mtb* proteins	3 (2, 3)	3 (2, 3)	-0.82	0.41
Average distance to *Mtb* proteins	5.51 (5.25, 5.97)	5.46 (5.17, 5.85)	-1.79	0.07
Minimum distance to HIV-1 proteins	2 (2, 3)	2 (2, 3)	2.86	0.004
Average distance to HIV-1 proteins	2.71 (2.43, 3.29)	2.86 (2.57, 3.14)	1.39	0.16

Notes. S is the Wilcoxon rank-sum test statistic.

Because proteins in the MHC region are known to play an important role in immunity towards pathogens, a second approach was taken to investigate their importance in the network. Instead of looking at the importance of MHC proteins compared to non-MHC proteins in the network, the distance from MHC proteins to proteins that were found to be important for bridging pathogens was calculated. Three sets of proteins were compared: (1) proteins that interact with both pathogens versus proteins that interact with any or neither pathogen, (2) proteins that have bridging centrality in the 95^{th} percentile versus all other proteins, and (3) proteins that have bridging centrality in the upper quartile versus all other proteins. In

each case, there was a significantly shorter minimum and average distance to MHC proteins in the subset of proteins important for connecting pathogens compared to the remaining proteins (see Table 2.5). This suggests that although the MHC proteins are not necessarily more important for connecting pathogens in the high confidence network, the proteins that are important for connecting pathogens are closer to the MHC proteins. Thus, the MHC proteins, through the guilt by association property, likely play an important role in mediating the interactions between the bridging proteins and the pathogens.

Table 2.5 Wilcoxon rank-sum test comparing the distance to MHC proteins from subsets of proteins identified as potentially important for connecting pathogens versus the remaining human proteins in the network.

Network Properties	Group 1	Group 2	S	*p*-value
	Proteins interacting with both HIV-1 and *Mtb* (*n*=28)	All other human proteins (*n*=15746)		
Shortest distance to MHC	1 (1, 2)	2 (2, 2)	-4.63	<0.001
Average distance to MHC	2.65 (2.52, 2.77)	3.23 (2.99, 3.61)	-8.34	<0.001
	Proteins with bridging centrality in the 95th percentile (*n*=345)	All other human proteins (*n*=15429)		
Shortest distance to MHC	1 (1, 2)	2 (2, 2)	-14.37	<0.001
Average distance to MHC	2.72 (2.57, 2.90)	3.25 (3.01, 3.62)	-23.04	<0.001
	Proteins with bridging centrality in the upper quartile (*n*=1795)	All other human proteins (*n*=13979)		
Shortest distance to MHC	1 (1, 2)	2 (2, 2)	-31.76	<0.001
Average distance to MHC	2.81 (2.70, 2.93)	3.32 (3.04, 3.66)	-48.41	<0.001

2.3.2 Network properties of drug target proteins

Functional PPINs have been used in various studies to identify potential drug targets based on the network centrality measures of proteins in the PPIN (Mazandu and Mulder, 2011; Mulder et al., 2013). In such studies, the common thread is that higher network centrality implies higher biological importance, enabling a narrowing down of potential drug targets. To confirm whether or not this principle holds in the PPIN created here, the network centrality measures of proteins targeted by either anti-TB or anti-HIV medication were compared to non-drug target proteins. Proteins targeted by drugs had significantly higher betweenness and degree compared to non-drug targets ($p<0.001$); however, closeness was not significantly different (see Table 2.6).

Table 2.6 Wilcoxon rank-sum test comparing network properties of proteins targeted by anti-HIV or anti-TB medication to non-drug targets.

Network property	Drug target proteins (*n*=97) median (IQR)	Non-target proteins (*n*=19 060) median (IQR)	S	*p*-value
Betweenness	19 321.80 (3984.76, 47 624.12)	1754.71 (36.12, 16 656.79)	6.93	<0.001
Closeness	0.28 (0.24, 0.31)	0.27 (0.24, 0.31)	0.32	0.75
Degree	27 (11, 63.5)	12 (3, 45)	4.13	<0.001
Pathogenicity Bridging	0.0 (0.0, 0.006)	0.0 (0.0, 0.003)	1.13	0.26

2.3.3 28 human proteins that functionally interact with both pathogens

There were 28 human proteins that functionally interacted with both pathogens (termed "bridging proteins" for convenience). These are listed in Table 2.7. A web-based visualisation created for these interactions can be viewed in Figure 2.3 and in a higher quality, interactive viewer at the Network Data Exchange (NDEx) https://www.ndexbio.org/#/network/d612b4a1-763f-11ec-b3be-0ac135e8bacf?accesskey=ebd0c3311a55add48895b218e43a901c82e03effbb26f8bebcafb05f3346eb38. In Figure 2.3, the red circles are the human proteins, the green diamonds are HIV-1 proteins, and the blue squares are *Mtb* proteins. Interactions with *Mtb* proteins and HIV-1 proteins are indicated by blue and green lines from human proteins respectively, while interactions between the 28 bridging proteins are coloured red.

Table 2.7 Human proteins interacting with both pathogens.

UniProt ID	Protein Full Name	Protein Name	Minimum distance to MHC	Minimum distance to HIV/TB drug
Q99683	Mitogen-activated protein kinase 5	MAP3K5	1	3
Q03135	Caveolin-1	CAV1	1	2
P07339	Cathepsin D	CTSD	1	2
P62330	ADP-ribosylation factor 6	ARF6	2	2
P07814	Bifunctional glutamate/proline–tRNA ligase	EPRS	1	2
O94992	Hexamethylene bis-acetamide-inducible protein 1	HEXIM1	2	3
P02751	Fibronectin (FN)	FN1	2	2
P26038	Moesin (Membrane-organizing extension spike protein)	MSN	1	3
P06241	Tyrosine-protein kinase Fyn	FYN	1	3
P62158	Calmodulin (CaM)	CALM1	1	3
P60953	Cell division control protein 42 homologue (G25K GTP-binding protein)	CDC42	2	2
Q9NNX6	Dendritic cell-specific ICAM-3-grabbing non-integrin 1	CD209	2	3
P10809	60 kDa heat shock protein, mitochondrial	HSPD1	1	2
Q04759	Protein kinase C theta type	PRKCQ	1	2
P55209	Nucleosome assembly protein 1-like 1	NAP1L1	2	3
P14625	Heat shock protein 90 kDa β member 1	HSP90B1	1	3
P51149	Ras-related protein Rab-7a	RAB7A	1	3
O43390	Heterogeneous nuclear ribonucleoprotein R (hnRNP R)	HNRNPR	1	3
P24723	Protein kinase C eta type	PRKCH	2	3
P19838	Nuclear factor NF-κ-B p105 subunit (DNA-binding factor KBF1)	NFκB1	1	2
P06239	Leukocyte C-terminal Src kinase	LCK	1	2
P15153	Ras-related C3 botulinum toxin substrate 2	RAC2	2	2
P40763	Signal transducer and activator of transcription 3	STAT3	1	2
P19525	Interferon-induced, double-stranded RNA-activated protein kinase	EIF2AK2	2	2
P67809	(Y-box-binding protein 1	YBX1	1	2
P04406	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	GAPDH	1	2
Q13951	Core-binding factor subunit β (CBF-β)	CBFB	2	3
P31751	RAC-β serine/threonine-protein kinase	AKT2	1	2

2.3.3.1 Comparing the centrality measures of the 28 bridging proteins to other human proteins

The network centrality measures of these 28 proteins were compared to all other human proteins in the PPIN. The bridging proteins had significantly higher betweenness, closeness, and degree than non-bridging proteins ($p<0.001$). The results of the Wilcoxon rank-sum tests are displayed in Table 2.8. The median degree of the bridging proteins was 183.5 (99.5, 275.5), indicating that these proteins functionally interact with hundreds of other proteins, which is suggestive to their biological importance.

Table 2.8 Wilcoxon rank-sum test comparing network properties of the 28 bridging proteins to non-bridging proteins.

Network property	Bridging proteins (*n*=28) median (IQR)	Non-bridging proteins (*n*=15 748) median (IQR)	S	*p*-value
Betweenness	373 589.72 (235 304.15, 900 873.55)	1754.71 (50.10, 14 220.20)	8.83	<0.001
Closeness	0.35 (0.33, 0.36)	0.28(0.26, 0.31)	8.16	<0.001
Degree	183.5 (99.5, 275.5)	16 (4, 58)	6.77	<0.001

2.3.3.2 Comparing the centrality measures of the *Mtb* proteins that interact with the bridging proteins to other *Mtb* proteins

In addition to the 28 human bridging proteins having significantly higher network centrality than non-bridging human proteins, the *Mtb* proteins that functionally interact with the human bridging proteins were found to have significantly higher betweenness and degree in the *Mtb*-*Mtb* functional PPIN than other *Mtb* proteins. This indicates that *Mtb* proteins that interact with human proteins, which also interact with HIV genes, may be more biologically important than other *Mtb* proteins.

Table 2.9 Wilcoxon rank-sum test comparing network properties of the *Mtb* proteins interacting with the 28 bridging proteins to other *Mtb* proteins.

Network property	Bridging *Mtb* proteins (*n*=47) median (IQR)	Non-bridging *Mtb* proteins (*n*=3323) median (IQR)	S	*p*-value
Betweenness	4738.73 (648.99, 15 718.22)	1037.25 (0.0, 5643.55)	3.70	<0.001
Closeness	0.22 (0.20, 0.24)	0.201(0.18, 0.23)	1.51	0.13
Degree	13 (6.5, 18)	4 (2, 10)	4.36	<0.001

2.3.4 Enriched GO terms and pathways in the 28 proteins interacting with both pathogens

Enriched GO terms in the subnetwork of proteins were determined using functional annotation clustering, which clusters related enriched terms in the dataset. For each cluster, a

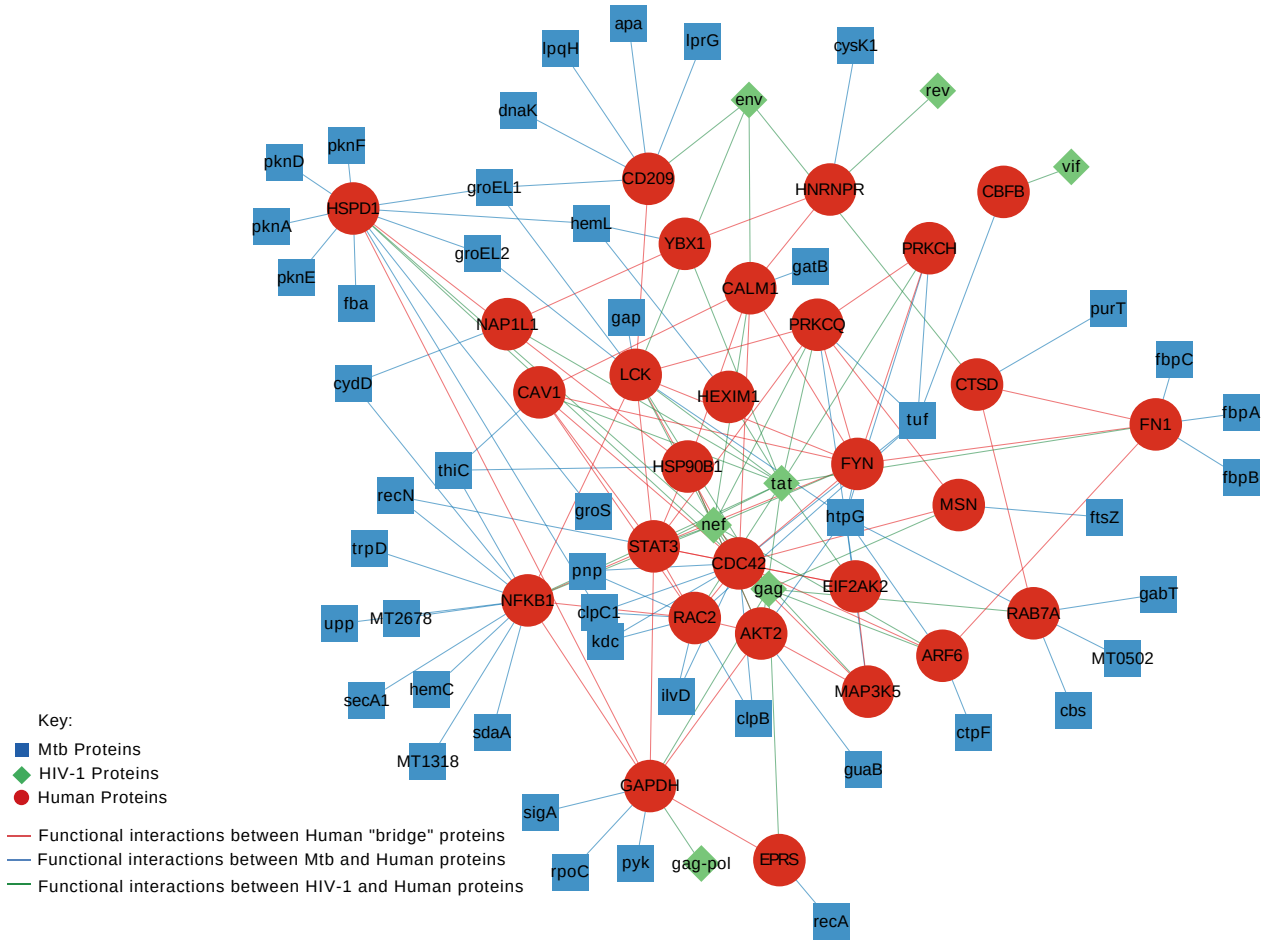


Figure 2.3 Subnetwork of PPIs between 28 human proteins that interact with both HIV-1 proteins and *Mtb* proteins.

The red circles are the human proteins, the green diamonds are HIV-1 proteins, and the blue squares are *Mtb* proteins. Interactions with *Mtb* proteins and HIV-1 proteins are indicated by blue and green lines from human proteins respectively, while interactions between the 28 bridging proteins are coloured red. To access the interactive viewer at NDEx use the following link: https://www.ndexbio.org/#/network/d612b4a1-763f-11ec-b3be-0ac135e8bacf?accesskey=ebd0c3311a55add48895b218e43a901c82e03effbb26f8bebcafb05f3346eb38

group enrichment score is derived by calculating the geometric mean (in -log scale) of the group member's *p* values in order to rank the clusters by biological significance. 29 clusters were suggested. The five clusters that stood out as relevant in the context of this study were: (1) nucleotide binding and T-cell receptor signaling pathway (terms enriched between 25 and 57% in the subset relative to the human genome), (2) T-cell activation and immune development (10.7 to 14.3% enriched), (3) regulation of T-cell activation (10.7 to 25% enriched), (4) GTPase activity and phagocytosis (10.7 to 14.3% enriched), and (5) Cell death (17.9 to 25% enriched). The terms associated with these clusters that were significantly enriched in the subset are listed in Figure 2.4. Alongside each cluster name in square brackets is the enrichment score for the cluster and the number of proteins that contained terms that were significantly enriched (p value ≥ 0.05) (before Benjamini-Hochberg multiple-testing correction). Terms that were not significantly enriched were excluded from the diagram and counts. The terms that were not significant after Benjamini-Hochberg multiple-testing correction are in italics.

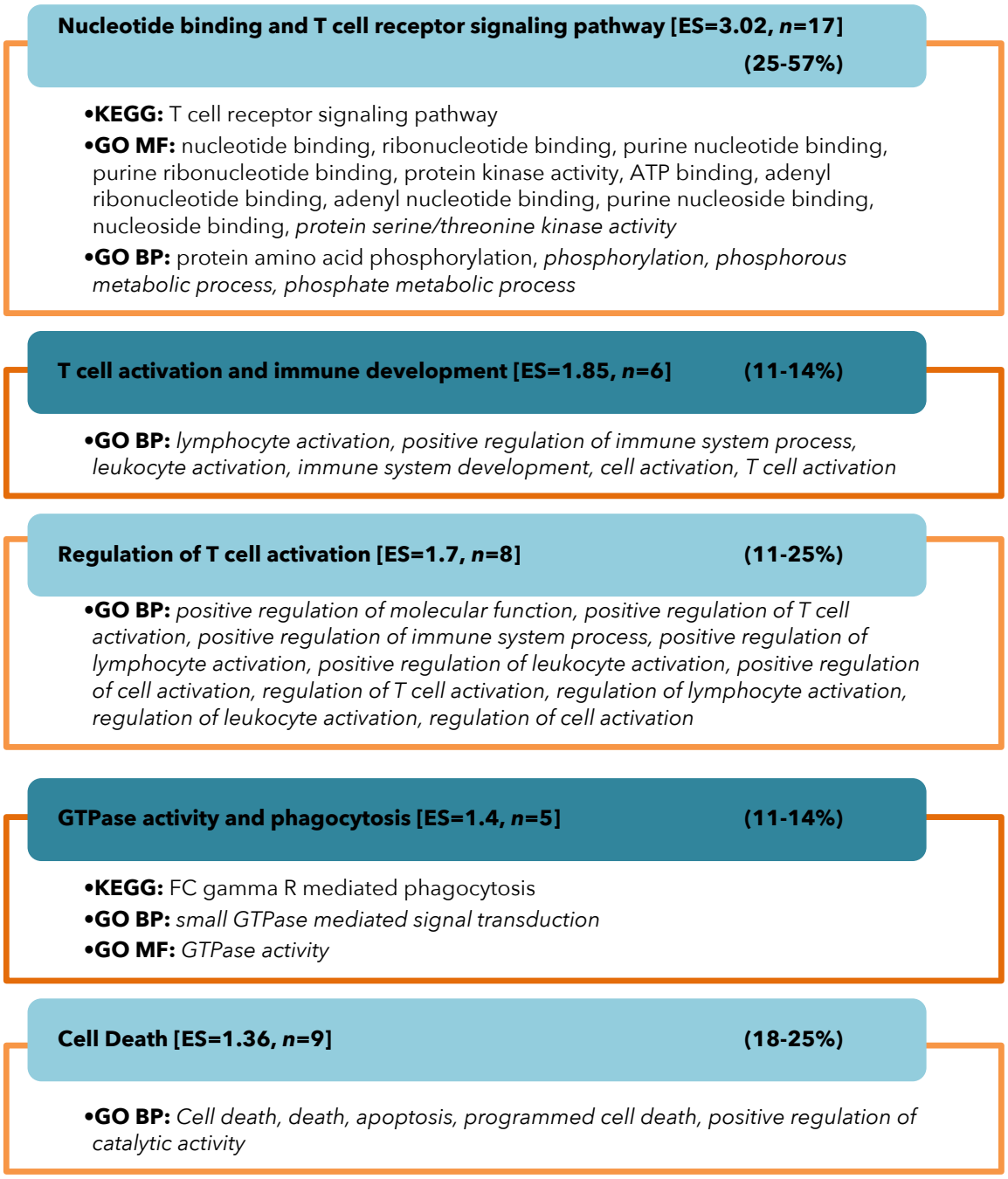


Figure 2.4 Clusters of enriched GO terms and KEGG pathways in the subnetwork of 28 human proteins functionally interacting with both HIV and *Mtb* proteins.
Five clusters of significantly enriched Gene Ontology (GO) terms for the 28 bridging proteins are represented in the blue rectangles, along with the enrichment score (ES) and number of proteins containing the terms in square brackets, and the percentage enrichment in round brackets. Below the cluster name within the orange framed boxes are some of the relevant terms divided into the categories of GO BP (biological process), GO MF (molecular function), and KEGG pathways.

2.3.5 The distance between human proteins important in the PPIN and existing anti-HIV and anti-TB drugs

The shortest distance and average distance to an anti-HIV and an anti-TB drug was calculated for each protein in the network. Out of the 1795 proteins in the upper quartile for pathogenicity bridging centrality, only one protein is targeted by both an existing anti-TB and anti-HIV drug. The protein was Serum Albumin (P02768), and is targeted by the anti-TB drug rifampicin and the anti-HIV drug saquinavir (as a carrier). Eleven proteins in the upper quartile for pathogenicity bridging centrality directly interact with a drug targeting at least one of the pathogens, and interact with at most two other proteins before interacting with a drug targeting the other pathogen.

None of the 28 proteins that interact with both pathogens are listed as targets in drug bank for either disease, and the 28 proteins have a shortest path distance of between two and three other proteins to a drug target. In other words, the proteins that interact with both pathogens must interact with at least one to two other proteins before being targeted by a drug that targets either pathogen.

2.4 Discussion

In this chapter, a network of functional protein interactions between human and it's co-infecting pathogens, HIV and *Mtb*, which also included drug-target interactions for anti-HIV and anti-TB drugs, was constructed. In addition to defining a centrality measure that may identify important human proteins that act as "bridges" between the pathogens, 28 human proteins were found to functionally interact with both HIV and *Mtb* proteins. Furthermore, although the network centrality measures of the MHC proteins did not show that they are more important than other human proteins for facilitating interactions between HIV-1 and *Mtb*, we did observe that MHC proteins had a significantly shorter path distance to the 28 human proteins that interact with both pathogens. The network was constructed from other networks of experimentally verified and computationally predicted interactions, and this work did not attempt to predict interactions. In this discussion, a literature review of the plausibility of the host-pathogen interactions predicted by the sources integrated in this network will be provided. In addition, the biological significance of these 28 proteins will be discussed in detail. Thereafter, the findings regarding the importance of the MHC proteins within the network will be discussed.

2.4.1 28 human proteins interact with both pathogens

As previously mentioned, 28 human proteins that functionally interact with at least one HIV-1 protein and at least one *Mtb* protein were identified. Compared to other human proteins in the network, these proteins had significantly higher network centrality measures, suggesting their biological importance and potential use as drug targets. In addition, these proteins were found to be significantly enriched for immunological functions, including: T-cell receptor signaling pathway, T-cell activation and immune development, GTPase activity and phagocytosis, and cell death. In order to try understand how these host-pathogen

interactions could impact HIV-TB co-infection, the literature was searched for (1) each protein's function, (2) evidence of the human protein playing a role in HIV or TB pathogenesis, and, where possible, (3) evidence of the specific interaction. Because the HIV-1-human interactions were determined based on text-mining, there was far more literature describing the specific interactions than there was for *Mtb*-human interactions, which were largely predicted based on homology (Huo et al., 2015; Rapanoel et al., 2013).

With the exception of one human protein, EPRS, the remaining 27 human proteins all seemed to play a likely role in co-infection. EPRS, Bifunctional glutamate/proline-tRNA ligase, is involved in biological processes such as tRNA aminoacylation for protein translation, cellular response to interferon-γ and negative regulation of translation. It was predicted to interact with *Mtb* gene *rec*A by Rapanoel et al. (2013) and with HIV gene *Gag*. The HHPID annotated that *Gag* complexes with EPRS. However, there is minimal evidence in the literature that indicates that EPRS plays a role in the pathogenesis of either disease. The potential relevance of each of the remaining 27 human proteins and the pathogenic proteins they interact with is discussed in the next section.

2.4.1.1 Regulation of apoptosis

Apoptosis, programmed cell death, is an important defence mechanism used by the immune system to eradicate invading pathogens. Induction of apoptosis is a mechanism used by HIV to deplete CD4 T-cells, which are incidentally the dominant effector cells against *Mtb* (Boom et al., 2003), and, as such, reduction of CD4 T-cells weakens the immune response to TB infection. In addition to promoting apoptosis of CD4 T-cells during co-infection, our results suggest that HIV may inhibit apoptosis in an AKT dependent manner, leading to persistent *Mtb* growth in macrophages. According to Cooray (2004), Protein Kinase B, or AKT signaling, is an important mechanism for inhibiting apoptosis during viral infection. AKT regulates apoptosis through transcriptional control of genes that activate and inhibit apoptotic genes (Cooray, 2004). For example, by phosphorylating *FKHR*, a gene expressed in the nucleus that regulates the transcription of genes needed for apoptosis and cell proliferation, AKT promotes the export of FKHR from the nucleus to the cytosol where it is inhibited by other proteins (Brunet et al., 1999).

In addition, AKT2 regulates cell survival by phosphorylating MAP3K5, which is a kinase that releases signals to induce apoptosis (Kumawat et al., 2010). TNF-α activates a pathway involving the kinase MAP3K, which promotes apoptosis (Kumawat et al., 2010). Furthermore, *Mtb* induced TNF-α depends on MAP3K5 (Kumawat et al., 2010). Kumawat et al. (2010) showed that HIV-1 Nef inhibits MAP3K5 activation during co-infection, such that Nef reduces TNF-α production and apoptosis in macrophages infected with *Mtb* by inhibiting MAP3K5. According to Diedrich and Flynn (2011), macrophage apoptosis is an important immune response to *Mtb* during HIV-TB co-infection. As such, Kumawat et al. (2010) propose that HIV-1 Nef may be able to exacerbate tuberculosis infection by reducing *Mtb*-induced TNF-α production. Tachado et al. (2008) found that there was less TNF-α released by the HIV+ macrophages. Lower levels of TNF-α reduces the apoptotic response to *Mtb*, thereby exacerbating the tuberculosis infection (Tachado et al., 2008).

Based on the network generated in this study, AKT2 was predicted to functionally interact with the HIV-1 proteins Nef and Tat (Fu et al., 2009). According to the GO cellular component annotation, Nef is located in the host plasma membrane, and AKT2 is located in the plasma membrane of human cells. *AKT* genes had higher levels of phosphorylation in HIV+ macrophages, induced by Nef (Tachado et al., 2008). Similarly to Nef, Tat shares a cellular component with AKT2. According to the GO cellular component annotation, Tat is located in the host cell nucleus and host cell nucleolus, and AKT2 is located in the human nucleus and nucleoplasm. In a study investigating Kaposi's sarcoma (a common HIV-1 malignancy), it was observed that after treatment with Tat, *AKT* activity increased, thus enabling the infected cells to persist given that *AKT* is an anti-apoptotic gene (Deregibus et al., 2002). As in the study by Deregibus et al. (2002), Van Grol et al. (2010) found that incubation of non-infected monocyte-derived macrophages with HIV-1 Tat resulted in higher levels of phosphorylation in *AKT*. In addition, they checked whether *AKT* was required for autophagy by siRNA silencing, and found that the inhibition of autophagy was *AKT*-dependent (Van Grol et al., 2010).

AKT2 interacted with two *Mtb* proteins, guaB (Inosine-5'-monophosphate dehydrogenase) and htpG (Chaperone protein HtpG). The *Mtb* protein GuaB (IMP dehydrogenase), plays an important role in regulating cell growth. GuaB's interaction with AKT2 was inferred by interactions determined by Huo et al. (2015). GuaB is found in both human and *Mtb*, along with several other species. Inhibiting GuaB reduces guanine nucleotides, and, since GTP is needed to covert Inosine monophosphate (IMP) to adenosine monophosphate (AMP), adenylate pools are reduced. This results in the interruption of DNA and, RNA synthesis and, as such, cytotoxicity (Shu and Nair, 2008). Inhibiting H37Rv GuaB resulted in reduced growth of *Mtb* (Usha et al., 2011). Although, to our knowledge, there is no evidence in the literature of the interaction between GuaB and AKT2 in *Mtb*, the P13K/AKT pathway has also been shown to play a role in *Mtb* survival in macrophages (Maiti et al., 2001). ManLAM, has been shown to phosphorylate *BAD* (a Bcl-2 family member that induces apoptosis) in a P13/AKT pathway dependent manner, which, in turn, increases cell survival (Maiti et al., 2001). GuaB does not have any interaction (high or low confidence) with ManLam in our network.

It is clear that separately these host-pathogen interactions with AKT2 and MAP3K5 are beneficial to the pathogens' survival, and thus, together in the context of HIV-TB co-infection, it is plausible that the effects of the interactions are amplified by enabling infected cells to escape apoptosis. The host-pathogen interactions with AKT2 that may inhibit apoptosis are illustrated in Figure 2.5.

2.4.1.2 Increasing viral infectivity

Moesin (MSN) is the membrane-organizing extension spike protein. MSN has been predicted to interact with *Mtb* cell division protein FtsZ by Rapanoel et al. (2013). FtsZ is a bacterial homologue of tubulin that is necessary for initiation of bacterial cell division. It is highly conserved amongst prokaryotes and not present in humans, making it a potential antimicrobial target. Dziadek et al. (2002) showed that elevated levels of FtsZ affect growth and interfere with cell division.

MSN is a F-actin binding protein that acts as a linker between membrane proteins and the

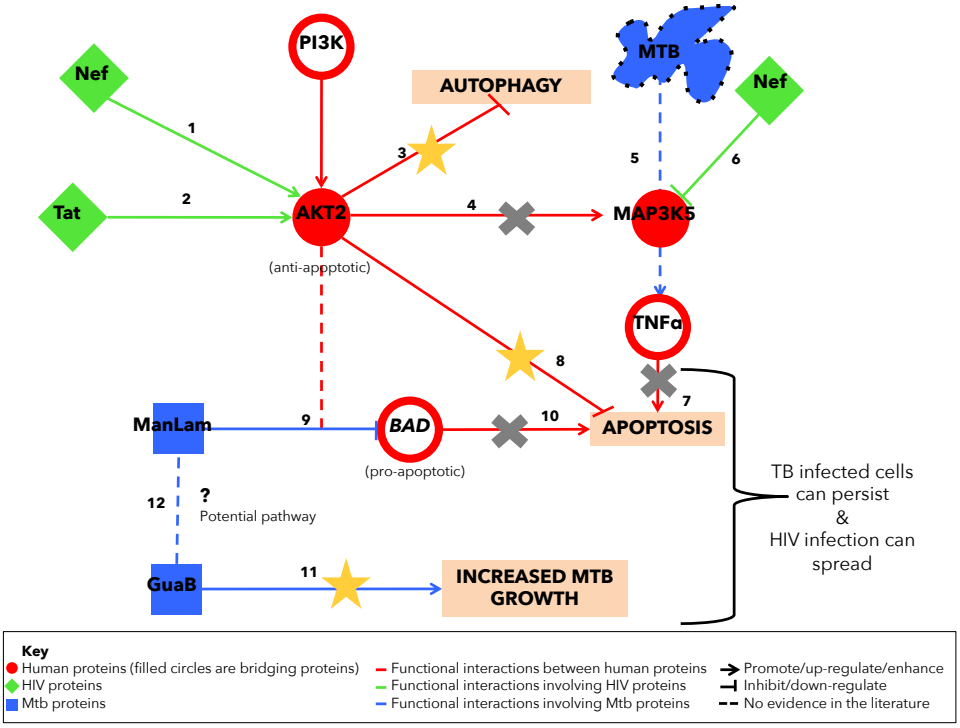


Figure 2.5 Host-pathogen interactions with human protein AKT2 that inhibit apoptosis.
AKT2 is an anti-apoptotic gene (Cooray, 2004). (1) Nef induces phosphorylation of AKT2 (in macrophages) (Tachado et al., 2008). (2) at induces phosphorylation of AKT2 and increases activity (Van Grol et al., 2010). (3) AKT2 inhibits autophagy – this is increased by its Nef and Tat mediated enhanced activity (Van Grol et al., 2010). (4) AKT2 activates MAP3K5 to signal apoptosis. (5) *Mtb* induces TNF-α in a MAP3K5 dependent way, which signals apoptosis - this is blocked by Nef (Kumawat et al., 2010). (6) Nef inhibits MAP3K5. (7) Apoptosis via MAP3K5/TNF-α is reduced due to reduced MAP3K5 activity (Kumawat et al., 2010). (8) By enhancing activity of anti-apoptotic gene AKT2, apoptosis inhibition is enhanced. (9) *Mtb* ManLAM inhibits pro-apoptotic gene *BAD* by phosphorylating in a PI3K/AKT dependent manner, which reduced apoptosis (10) (Maiti et al., 2001). (11) guaB increases *Mtb* growth (Usha et al., 2011). (12) Since the PI3K/AKT pathway is also involved in *Mtb* growth, perhaps guaB is involved in a signaling pathway with ManLAM. The reduced apoptosis and autophagy means that *Mtb* infected cells can persist and HIV infection can spread.

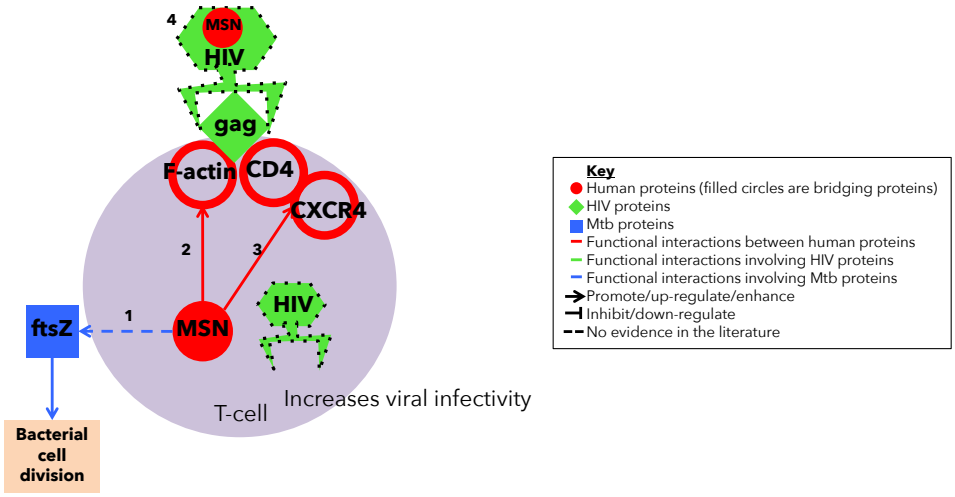


Figure 2.6 Interactions between human protein MSN and HIV and *Mtb* proteins that promotes infectivity and bacterial cell division.
(**1**) MSN is predicted to interact with MTB protein FtsZ, which is important for bacterial cell division (Dziadek et al., 2002; Rapanoel et al., 2013). (**2**) MSN promotes F-actin redistribution, as well as (**3**) CD4 and CXCR4 clustering which enhances viral infection by making the T-cells permissive for HIV (Barrero-Villar et al., 2009). (**4**) MSN is also incorporated into virions (Ott et al., 1996).

cytoskeleton, and is expressed in T-cells (Roy et al., 2014). Barrero-Villar et al. (2009) proposed that activated MSN promotes F-actin redistribution and CD4-CXCR4 clustering, and is also required for efficient HIV-1 infection in permissive lymphocytes. MSN is also known to be incorporated into HIV-1 virions (Ott et al., 1996). The cytoskeletal proteins ezrin, MSN, and cofilin are incorporated into HIV-1 particles, presumably through their interaction with actin which binds to the nucleocapsid domain of HIV-1 Gag. Roy et al. (2014) found that HIV-1 Gag co-localises with ezrin-radixin-moesin proteins at polarised HIV-1 assembly sites in human T-cells. These mechanisms by which MSN might promote HIV and TB infectivity are illustrated in Figure 2.6.

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) plays a role in glycolysis and nuclear events including transcription, RNA transport, DNA replication and apoptosis, as well as regulating viral replication. GAPDH interacts with HIV-1 proteins Gag and Gag-pol, which according to GO annotation are both located with GAPDH in the nucleus and plasma membrane. Many host proteins are found in virions, for example, LysRS is included due to the interaction with Gag or Gag-pol during assembly, which is important for Gag protein folding and tRNALys3 packaging (Kishimoto et al., 2012). Kishimoto et al. (2012) have shown that GAPDH is also incorporated into HIV-1 virions. They further showed that in GAPDH-defective virions, infectivity increased likely due to higher efficiency in reverse transcription (Kishimoto et al., 2012). This higher efficiency was due to the increased efficiency of tRNALys3 and LysRS packaging into the virions by virtue of GAPDH competing to bind to Gag and Gag-pol (Kishimoto et al., 2012). Similarly, in GAPDH enhanced virions, infectivity decreased (Kishimoto et al., 2012).

GAPDH was predicted to interact with three *Mtb* proteins, namely, *pyk*, *rpo*C and *sig*A (Rapanoel et al., 2013). GAPDH is also a *Mtb* protein (UniProt ID=P9WN82 for CDC1551). In

the glycolysis pathway that *Mtb* GAPDH functions in, *pyk* (Pyruvate kinase) is involved in step five of the subpathway of carbohydrate degradation (Consortium et al., 2017). According to the STRING *Mtb*-*Mtb* PPIN *pyk* interacts with *Mtb* GAPDH with high confidence (0.95), while *sig*A and *rpo*C interact with *Mtb* GAPDH with low confidence (0.3 and 0.23 respectively). The predicted interactions between GAPDH and *sig*A and *rpo*C could be explained due to their interaction with the *Mtb* homologue.

NF-κB (NFκB1) is a transcription factor that plays a role in immunity, and is involved in the T-cell receptor signaling pathway. It was predicted to interact with two HIV proteins, Nef and Tat (Fu et al., 2009). Niederman et al. (1992) showed that Nef inhibits induction of NFκB1 DNA-binding activity by T-cell mitogens. They additionally showed that the effect of Nef on HIV-1 transcription depends on an intact NF-κB-binding site (Niederman et al., 1992). *In vitro*, HIV-1 protein Tat, which is found in the host cell nucleus with NFκB1, has been shown to induce SOCS3 expression in human and murine macrophages in a NF-κB-dependent manner, which increases viral replication (Akhtar et al., 2010).

Rapanoel et al. (2013) predicted that NFκB1 interacted with 11 *Mtb* proteins, including CydD, HemC, MT1318, MT2678, RecN, SdaA, SecA1, ThiC, TrpD, and Upp. Although there was no evidence of these interactions in the literature, NFκB1 sites have been shown to mediate IL-6 induction in response to LAM and may thus act as mycobacterial response elements (Zhang et al., 1994). HIV-1 replication is notably upregulated in alveolar macrophages during pulmonary tuberculosis infection, which is associated with the activation of nuclear factor NF-κB (Hoshino et al., 2002). If NFκB1 is activated by HIV proteins such as Tat, this may negatively impact the response to *Mtb* infection. Furthermore, CDC1551 was shown to induce HIV-1 replication, which was shown to be associated with significantly increased levels of TNF and IL-6, and of NFκB (Ranjbar et al., 2009).

2.4.1.3 Enabling persistent infection and pathogen survival

Caveolin-1 (CAV1) is involved in the costimulatory signal that is required for T-cell receptor mediated T-cell activation. It interacts with HIV-1 protein Tat, and is predicted to interact with *Mtb* protein ThiC (Rapanoel et al., 2013). HIV-1 protein Tat is located in the host cell cytoplasm, and CAV1 is expressed in the perinuclear region of cytoplasm. In addition, Tat is involved in the apoptotic process, while CAV1 is involved in the apoptotic signaling pathway. Lin and Flynn (2010) showed that Tat mediates upregulation of CAV1 expression in HIV-infected human monocyte-derived macrophages, THP-1 macrophages and CD4 cells. In cells over-expressing CAV1, HIV production is reduced, suggesting that CAV1 may enable persistent infection in macrophages (Lin et al., 2010). *Mtb* ThiC is a phosphomethylpyrimidine synthase, also known as Thiamin (Vitamin B1). Thiamin is a necessary cofactor for all organisms, and Thiamin biosynthetics have been suggested as potential drug targets for new antimicrobial agents (Du et al., 2011). Many important enzymes are dependent on the active form of Thiamin, ThDP (thiamin diphosphate). In prokaryotic cells, thiamin biosynthesis involves formation of a thiazol moiety THZ-P, and a pyrimidine moiety HMP-PP (Du et al., 2011). ThiC catalyses aminoimidazole ribotide to form HMP-P, which is phosphorylated to HMP-PP by ThiD. ThiC is essential for *Mtb in vitro*, however this has not been shown *in vivo*

(Du et al., 2011).

HNRNPR, heterogeneous nuclear ribonucleoprotein, plays an important role in processing precursor mRNA in the nucleus. It is predicted to interact with *Mtb* protein CysK1 (O-acetylserine sulfhydrylase) (Rapanoel et al., 2013). CysK1 is an enzyme in the cystein biosynthesis pathway (Poyraz et al., 2013). The long term survival of *Mtb* within granulomas is directly related to the availability of cystein (Poyraz et al., 2013). In addition, cystein biosynthesis pathway is not apparent in humans, making it an attractive target for anti-TB drugs (Poyraz et al., 2013).

The HIV-1 Rev protein interacts with HNRNPR. Rev regulates HIV-1 gene expression, and acts as an adaptor protein for nuclear export of HIV RNAs (Hadian et al., 2009). Rev has been shown to bind specifically to HNRNPR, and expression levels of hnrNPs influence HIV replication (Hadian et al., 2009).

2.4.1.4 Enhancing viral cellular location for transmission and assembly

Studies suggest that HIV-1 assembly may occur in either of the two cellular components that Gag is targeted to – the plasma membrane or the multivesicular bodies (late endosome) (Ono et al., 2004). While assembly occurs at the plasma membrane in T-cells, it occurs in the late endosome in macrophages (multivesicular bodies), and mutations in the matrix (MA) domain of the Gag sequence have been shown to affect Gag targeting and membrane binding ability (Ono et al., 2004). In the human-pathogen PPIN presented here, Gag was found to functionally interact with ADP-ribosylation factor 6 (ARF6). ARF6 is a GTPase that localises at the cell periphery and moves between the plasma membrane and endosome dependent on guanine nucleotide concentration (Dana et al., 2000). GTPase activating proteins (GAPS) enable ARF catalysis of GTP and are therefore important for inactivating ARFs. From the GO annotation, we found that HIV-1 Gag shares cellular components with ARF6; specifically, ARF6 is found in the plasma membrane and Gag is found in the host plasma membrane. In addition, Gag is found in host multivesicular bodies, and ARF6 is found in the endosome, early endosome, endocytic vesicle, and in the recycling endosome membrane. Furthermore, one of the biological processes ARF6 is involved in is vesicle-mediated transport, which relates to the biological process that Gag is involved in, intracellular viral transport. This further supports the plausibility of the functional interaction between the Gag and ARF6 proteins.

Ono et al. (2004) investigated whether the host cell protein phosphatidylinositol (4,5) bisphosphate (PI(4,5)P2), a protein that is regulated by ARF6, is involved in Gag targeting and virus production. PI(4,5)P2 is a member of the phosphoinositide family of lipids, which localise preferentially to specific subcellular membranes, which therefore influences the target of the proteins they bind to (Ono et al., 2004). Ono et al. (2004) found that expression of ARF6 defective for GTPase activity notably reduced virus production, and induced endosomal vesicles enriched with PI(4,5)P2, to which Gag was localised, thereby reducing release from the cell surface. Ono et al. (2004) highlight the importance of this, as the location of HIV-1 assembly may enhance viral transmission through cell-cell contact and thus will likely affect replication and pathogenesis. García-Expósito et al. (2011) found that inactive mutants of ARF6 inhibited HIV-1 envelope induced membrane fusion and infection of T-cells, as well as

inhibited cell-to-cell HIV-1 transmission of primary CD4+ T-cells.

In addition to functionally interacting with HIV protein Gag, ARF6 functionally interacts with Nef, which like Gag is also located in the host plasma membrane. Nef plays an important role in immune evasion by down-modulating the expression of MHC class I genes, as well as the proteins CD4 and CD28 at the plasma membrane (Larsen, 2004). Blagoveshchenskaya et al. (2002) showed that over-expression of ARF6 defective for GTPase activity is able to block Nef mediated MHC class I down-modulation. They proposed that Nef targets the endosome compartment, which first activates PI3K. This then activates a guanine nucleotide exchange factor that lastly activates *ARF6* at the plasma membrane, enabling MHC class I internalisation. However, Larsen et al. (2004) showed that MHC class I down-modulation was not affected by other *ARF6* mutants defective in GDP-GTP cycling, and that inhibiting PI3K, which activates *ARF6*, also had no effect on MHC class I internalisation. In addition, Larsen et al. (2004) showed that the role of ARF6 in Nef mediated MHC class I down-modulation is due to membrane trafficking. Because of the differences in these results, others have continued to test the role of *ARF6* in Nef mediated MHC class I down-modulation. Wonderlich et al. (2011) found that Nef mediated MHC class I down-modulation was ARF1 dependent, but not ARF6 dependent. An inhibitor of ARF1 blocked the ability of Nef to recruit AP-1 to MHC-1, whereas active ARF1 mutants stabilised the Nef-MHC-I-AP-1 complex.

The two *Mtb* proteins, HtpG and CtpF, were identified as potential interactors with ARF6 based on the network by Huo et al. (2015). The *Mtb* protein CtpF is a P-type ATPase that has been shown to exhibit early, strong and sustained induction during infection of human macrophages with *Mtb* (Botella, 2010). In addition, CtpF is one of the most over-expressed transporters when *Mtb* is treated with toxic substances, such as first-line drug isoniazid. There was no evidence to our knowledge that CtpF or HtpG interact with ARF6; however, the observation that CtpF is strongly induced during macrophage infection and is over-expressed when treated with isoniazid suggests that CtpF plays a role in pathogen survival in the host. In addition, ARF6 mutants defective in GTP binding or hydrolysis inhibit the Fc-γ receptor, a receptor that induces phagocytosis, thereby reducing phagocytosis (Dana et al., 2000). Thus, ARF6 may play a role in phagocytosis of *Mtb* enabling CtpF to be expressed within macrophages and promote pathogenic survival.

Another functional interaction that may affect viral cellular location is the interaction between the human nucleosome assembly protein (NAP1L1) and HIV protein Tat. NAP1L1 is involved in important biological processes such as nucleosome assembly, regulation of cell proliferation, and DNA replication. NAP1L1 interacts with HIV-1 protein Tat, which, according to GO cellular component annotation, is located in the host cell nucleus and NAP1L1 is also localised in the nucleus. Nucleocapsid is a protein that packages viral genomic RNA and is encoded by *Gag*. In cells co-infected with nucleocapsid and Tat, NAP1L1 is downregulated (Lee and Park, 2009). Vardabasso et al. (2008) showed that Tat binds to NAP1L1 *in vivo* and *in vitro* and that this binding regulates Tat-mediated activation of viral gene expression. De Marco et al. (2010) showed that NAP1L1 interacts with Tat at the nuclear rim, and proposed that NAP1L1 is required to transport Tat within the nucleus. Thus, like ARF6, NAP1L1 may facilitate transport of viral proteins within the cell.

NAP1L1 was also predicted to interact with *Mtb* protein CydD by Rapanoel et al. (2013). CydD is involved in cysteine transport (GO) and the long term survival of *Mtb* within granulomas is directly related to the availability of cystein (Poyraz et al., 2013). As such, it is possible that NAP1L1 is involved in regulating the survival of *Mtb* in granulomas.

Based on their functional interactions with pathogenic proteins, NAP1L1 and ARF6 both seem to play a role in cell-cell transmission of viral proteins, and may potentially be involved in regulating the survival of *Mtb*. These interactions are depicted in Figure 2.7

2.4.1.5 Regulation of phagosome maturation

Phagosome maturation is marked by the fusion of the phagosome with the lysosome, which confers degradative properties to the phagosome. The fused phagolysosomes expose the bacteria to hydrolytic enzymes, which can kill the bacteria (Gupta et al., 2012). IL-10 is secreted from *Mtb* infected macrophages and has been shown to inhibit phagosome maturation (O'Leary et al., 2011) and thereby prevent killing of the bacteria. This inhibition was shown to be mediated by Signal transducer and activator of transcription 3 (STAT3). STAT3 plays a role in viral processes and other processes related to immunity. STAT3 was predicted to interact with *Mtb* DNA repair protein recN (Rapanoel et al., 2013). Inhibition of STAT3 resulted in enhanced phagosome maturation in *Mtb* infected macrophages (O'Leary et al., 2011). In addition, STAT3 interacts with HIV-1 proteins Tat and Nef. Van Grol et al. (2010) showed that HIV-1 inhibits autophagy in bystander macrophages by acting through counter-regulators of this process, namely SRC, AKT, and STAT3. Furthermore, HIV-1 Tat upregulates IL-10, similarly to what is observed in *Mtb* infected macrophages. There is evidence that IL-10 is required to inhibit autophagy (Van Grol et al., 2010). Silencing of STAT3 prevented inhibition of autophagy by IL10 and Tat (Van Grol et al., 2010). Nef has been shown to activate STAT3 in human macrophages (Percario et al., 2003). In HIV-TB co-infected patients, levels of IL-6 are high (Toossi et al., 2012). STAT3 is an IL-6 inducible transcription factor that forms a complex with CDK9, which facilitates the replication of HIV in mononuclear cells (Toossi et al., 2012).

In addition to STAT3, Calmodulin (CALM1) is involved in phagosome maturation. It has been suggested that calmodulin and a calmodulin dependent protein kinase II may regulate phagosome and lysosome fusion (Vieira et al., 2002). *Mtb* inhibits phagosome maturation, for which a mechanism has been proposed where *Mtb* inhibits sphingosine kinase (SK), which results in a block in CA2+ dependent phagosome maturation (Malik et al., 2003). The *Mtb* protein GatB was predicted to interact with CALM1 by Rapanoel et al. (2013). GatB is the Aspartyl/glutamyl-tRNA(Asn/Gln) amidotransferase subunit B (Asp/Glu-ADT subunit B). The biological process it is involved in is translation and its molecular functions are ATP binding and glutaminyl-tRNA synthase (glutamine-hydrolyzing) activity. The two HIV-1 proteins that were predicted to interact with CALM1 were Env and Nef (Fu et al., 2009). Nef is expressed during early HIV infection, and can alter T-cell signaling pathways. Nef has been shown to bind to CALM1 in HIV-infected cells (Matsubara et al., 2005). In addition, under the influence of Ca2+ signaling, the interaction between Nef and CALM1 was enhanced (Matsubara et al., 2005). It is possible that this interaction may relate to the mechanism for inhibiting phagosome maturation by *Mtb*.

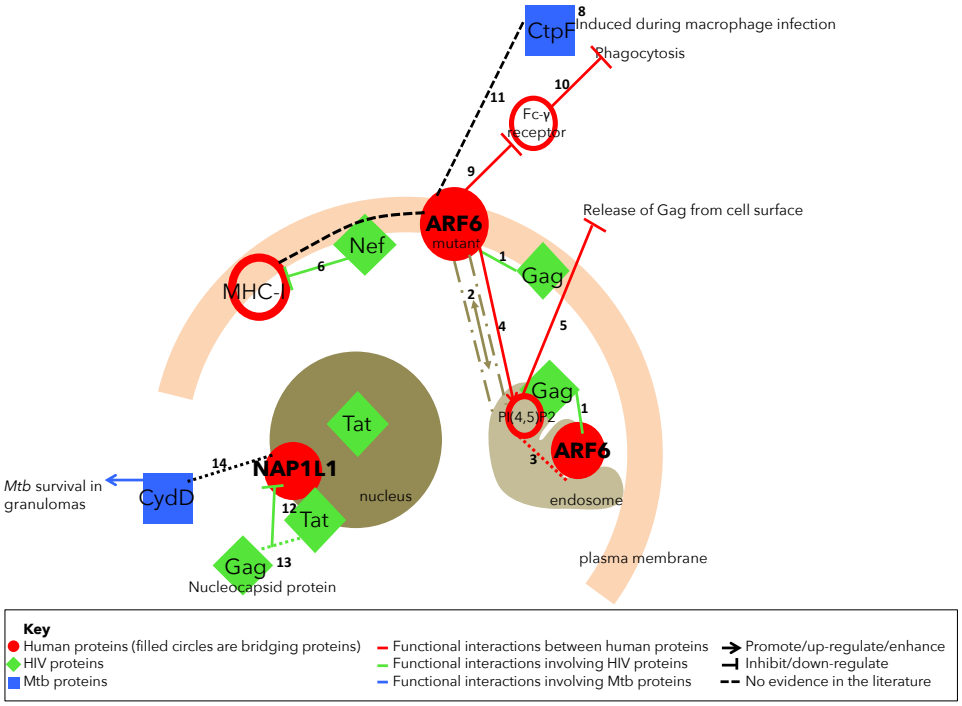


Figure 2.7 Functional host-pathogen interactions that may be involved in enhancing viral cellular location for transmission and assembly.
(**1**) AFR6 is likely to interact with Gag at the plasma membrane and the endosome. (**2**) ARF6 moves between the plasma membrane and the endosomal compartment, enabling it to be involved in vesicle mediated transport. (**3**) ARF6 regulates PI(4,5)P2. (**4**) Mutant ARF6 defective for GTPase activity leads to enriched PI(4,5)P2 in endosomal vesicles, causing Gag to co-localise with PI(4,5)P2 in the endosome. (**5**) This inhibits release of Gag from the cell surface. (**6**) Nef down modulates the expression of MHC class I molecules at the plasma membrane. (**7**) Mutant ARF6 may block Nef mediated down modulation or play a role in membrane trafficking. (**8**) *Mtb* protein CtpF is induced during macrophage infection. (**9**) ARF6 mutant inhibits binding to the Fc-γ receptor, which reduces phagocytosis (**10**) . (**11**) ARF6 was predicted to functionally interact with CtpF, which could be due to its involvement in regulation of phagocytosis. (**12**) NAP1L1 binds to Tat at the nucleus. (**13**) In the presence of the Nucleocapsid protein encoded by Gag, Tat can downregulate NAP1L1 which may enable Tat to be transported into the nucleus. (**14**) CydD was predicted to interact with NAP1L1, and is involved in cystein transport at the long term survival of *Mtb* in granulomas.

CALM1 is recognised by different CALM1 activated enzymes based in amphipathic helical segments (Ishikawa et al., 1998). The carboxyterminal domains of HIV-1 Env contains regions that are able to fold into similar helical segments resembling those found in CALM1 activated enzymes. Env gp160 can bind to CALM1 in the presence of Ca2+ and in low concentration (Ishikawa et al., 1998). CALM1 bound to Env results in enhanced CALM1 activity which increases the level of Ca2+ and this is followed by DNA fragmentation and apoptosis (Ishikawa et al., 1998). In the HIV-TB co-infected host where CALM1 regulates the phagosome and lysosome fusion (Vieira et al., 2002), *Mtb* blocks Ca2+ in the host in order to inhibit phagosome maturation (Malik et al., 2003). By blocking Ca2+, the binding of CALM1 to Env is diminished, reducing apoptosis (Ishikawa et al., 1998).

Ras-related protein RAB7A is an important regulator in endo-lysosomal trafficking. It also plays a role in the maturation of phagosomes and phagosome-lysosome fusion. One of the reasons why *Mtb* can survive and replicate inside macrophages is the failure of phagosome-lysosome fusion (Via et al., 1997). Via et al. (1997) showed that in *M. bovis* BCG phagosomal compartments, RAB5 is retained by RAB7 and does not associate with the phagosome, inhibiting phagosome lysosome fusion and phagosome maturation. Seto et al. (2009) showed that RAB7 is recruited to the phagosome, but it is later dissociated from the phagosome. This inhibits fusion of the phagosome with cathepsin D, thus blocking of phagosome maturation and phago-lysosome fusion (Seto et al., 2009). RAB7A was predicted by Huo et al. (2015) to interact with four *Mtb* proteins, Cbs, GabT, HtpG, and MT0502. GabT is involved in the pathway 4-aminobutanoate degradation, which is part of Amino-acid degradation. MT502 is an uncharacterised oxidoreductase. Cbs, putative cystathionine beta-synthase, has been found to be exclusively expressed in *Mtb* grown intra-phagosomally. Overexpression of RAB7 has been shown to inhibit HIV-1 gene expression (Vidricaire and Tremblay, 2005). HIV-1 Gag mainly remains in endosomes and co-localises with endosomal protein RAB7 (Vidricaire and Tremblay, 2005).

The host-pathogen protein interactions that are involved in regulating phagosome maturation are displayed in Figure 2.8.

2.4.1.6 Increasing viral infectivity and reduced T-cell recognition

CBFβ, (core-binding factor, beta subunit), binds to many enhancers and promoters including T-cell receptor enhancers, LCK (another one of the 28 "bridging" human proteins), and IL3. The HHPID indicated that CBFβ binds to HIV-1 protein Vif. Vif, is the viral infectivity factor protein that is required by HIV-1 to neutralise APOBEC3 restriction factors in the human host. Four members of the APOBEC family of restriction factors combine to restrict Vif deficient HIV-1 replication. Vif, on the other hand, neutralises APOBEC3 proteins by recruiting a ubiquitin ligase complex to polyubiquitinate the APOBEC3 restriction factors and targets them for degradation (Jäger et al., 2012). CBFβ has been found to be associated with a HIV-1 Vif, E3 ubiquitin ligase complex, where, *in vivo*, CBFβ knockdown resulted in lower levels of Vif, as well as reduced APOBEC3 degradation and decreased viral infectivity (Jäger et al., 2012). Jäger et al. (2012) thus propose that Vif is able to 'hijack' CBFβ and take it to the E3 ubiquitin ligase complex required for APOBEC3 polyubiquitination and degradation. This

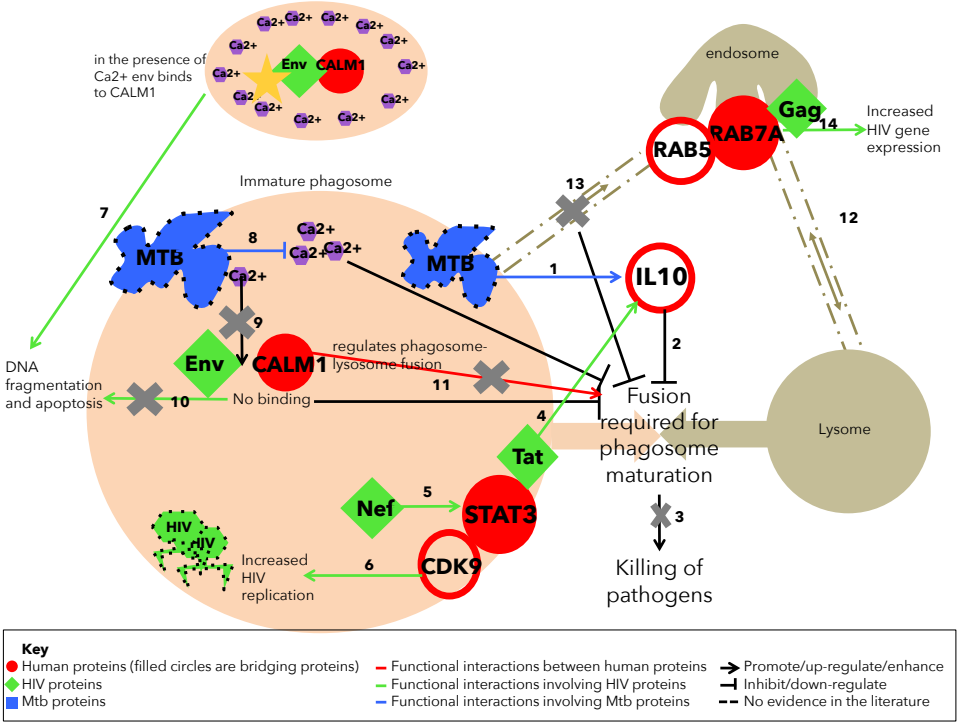


Figure 2.8 Regulation of phagosome maturation through host-pathogen protein interactions during HIV-TB co-infection.

(**1**) IL10 is secreted from *Mtb* infected macrophages. (**2**) IL-10 inhibits phagosome maturation, which blocks the killing of pathogens (**3**) . (**4**) HIV protein Tat functionally interacts with STAT3, and upregulates IL10. (**5**) Nef activates STAT3, which complexes with CDK9 facilitating replication of HIV in mononuclear cells (**6**) . (**7**) In the presence of Ca2+, Env binds to CALM1, which promotes DNA fragmentation and apoptosis. (**8**) *Mtb* blocks Ca2+ in the host to inhibit phagosome maturation. (**9**) The reduced levels of Ca2+ prevents Env and CALM1 binding. (**10**) The reduced binding prevents DNA fragmentation and apoptosis. (**11**) CALM1 regulates phagosome-lysosome fusion, but, due to immature phagosomes, fusion is reduced. (**12**) RAB7A is involved in endolysosomal trafficking. (**13**) During *Mtb* infection, RAB7 retains RAB5, preventing it from associating with the phagosome, which inhibits phagosome maturation. (**14**) Gag colocalises with RAB7 at endosomes which may enhance HIV gene expression.

enables Vif to increase viral replication and infectivity in the host, as well as reduce T-cell receptor recognition.

The interaction with *Mtb* protein tuf and CBFβ was predicted by Rapanoel et al. (2013). The tuf gene encodes the Ef-Tu protein, an elongation factor that plays a role during translation and is a GTPase (Lathe and Bork, 2001).

Like CBFβ, CD209 interacts with LCK. CD209, also known as DC-SIGN, is a calcium dependent C-type lectin pathogen recognition receptor that is expressed in the surface of immature dendritic cells and is involved in primary immune response (Tailleux et al., 2003). *Mtb* is able to infect dendritic cells by ligation of CD209 by lipoarabinomann (LAM) (Tailleux et al., 2003). In addition, alleles in CD209 have been shown to affect susceptibility to tuberculosis in different populations (Barreiro et al., 2006; Vannberg et al., 2008). Vannberg et al. (2008) suggest that this is likely explained by the binding of CD209 to *Mtb* LAM to prevent pro-inflammatory immune response.

CD209 was predicted to interact with five *Mtb* proteins by Huo et al. (2015), namely: LprG, LpqH, GroEL1, DnaK, and Apa. LprG is a lipoprotein, predicted to be localised in the cell wall or periplasm (Drage et al., 2010). Knocking out LprG resulted in reduced growth of *Mtb* and reduced *Mtb* survival (Bigi et al., 2004). In *M. smegmatis*, deleting the LprG operon resulted in altered cell morphology (Farrow and Rubin, 2008). Drage et al. (2010) propose that LprG contributes to virulence by acting as a carrier of glycolipids during their trafficking and delivery to the mycobacterial cell wall. In addition, they showed that LprG binds to LAM, potentially indicating its relationship with CD209. Toll-like receptor 2 (TLR2) plays an important role in innate immune recognition of *Mtb* (Drage et al., 2010). LprG has been shown to inhibit MHC class II antigen processing in human macrophages in a TLR2 signaling dependent manner, which may, in turn, prevent recognition by CD4+T-cells (Gehring et al., 2004). LpqH is another *Mtb* lipoprotein that interacts with the mannose receptor to promote phagocytosis, and behaves as a TLR2 agonist that downregulates antigen presentation to T-cells (Sánchez et al., 2012). Sánchez et al. (2012) show that lpqH triggers TLR2 activation, which induces macrophage apoptosis. GroEL1, also known as Cpn60.1 is a molecular chaperone. GroEL1 has been shown to be non-essential for *Mtb* viability; however, knockouts exhibited slower growth and showed less granulomatous inflammation (lower levels of pro-inflammatory cytokines, e.g. TNF-α, IFN-γ, IL6, IL12) in the lungs of mice (Hu et al., 2008). DnaK, also known as Hsp70, has been shown to induce dendritic cell maturation, which relates to the binding of LAM to dendritic cells. Pitarque et al. (2005) proposed that *Mtb* recognition by CD209 depends on the difference of accessibility of LAM in the envelope and the binding of other ligands, possibly including the Apa (alanine/proline-rich antigen) - one of the proteins predicted by Huo et al. (2015) to interact with CD209.

The HIV-1 protein CD209 interacts with is Env. At the level of GO annotation, CD209 is located in the plasma membrane and Env is located in the host cell plasma membrane. Some of the biological processes that CD209 is involved in that relate to viral processes included virion attachment to host cell, viral genome replication, and modulation by virus of host morphology or physiology. CD209 has been shown to preferentially bind to HIV-1 Env glycoprotein gp120, indicating that it is a specific dendritic cell surface receptor for Env

(Geijtenbeek et al., 2000). The binding with Env also plays an important role in the propagation of HIV-1 in dendritic cell and T-cell co-cultures (Geijtenbeek et al., 2000). However, CD209 was shown to not be able to substitute CD4 or CCR5 in HIV-1 entry into target cells. They further found that CD209 expressed at the surface of cells can sequester the HIV-1 in a way that can subsequently infect cells that are HIV-1 permissive, and enhance CD4-CCR5 mediated HIV-1 entry (Geijtenbeek et al., 2000). Thus, CD209 both enables the infection of dendritic cells with *Mtb* and enhances the infection of target cells with HIV-1.

Cathepsin D (CTSD) is an aspartic protease that participates in proteolysis and antigen presentation via MHC class II to CD4 and CD8 T-cells. Like CD209, CTSD interacts with HIV protein Env (gp120). Yu et al. (2010a) defined highly conserved cleavage sites within Env gp120 that are recognised by cathepsin D (as well as cathepsin S and L), which are important for antigen processing and presentation via MHC class II. Yu et al. (2010a) suggest that these cleavage sites are yet another mechanism by which HIV-1 escapes immune response, by inserting protease cleavage sites in regions for receptor binding. In addition to potentially enabling HIV-1 to escape immune response, there is evidence suggesting that CTSD may modify the conformation of Env gp120, allowing direct interaction with CXCR4 co-receptor, which, in turn, increases viral infectivity and entry into cells (El Messaoudi et al., 2000). CTSD was predicted to interact with the *Mtb* protein PurT by Rapanoel et al. (2013), which has been associated with slow growth rate in tuberculosis (Beste et al., 2009), which is further associated with delayed detection of infection and delayed immune response.

Interactions with CTSD, CD209, and CBFβ are displayed in Figure 2.9.

2.4.1.7 Regulating attachment to immune recognition cells

Fibronectins play a role in cell adhesion and cell motility. Huo et al. (2015) predicted interactions between fibronectin (FN1) and *Mtb* fibronectin binding proteins Fbpa, Fbpb, and Fbpc - all four of which are located in the extracellular region/extracellular space. The earliest phase of primary infection in pulmonary TB is marked by the attachment of *Mtb* to alveolar macrophages. Pasula et al. (2002) have shown that fibronectin binds to the cell-binding domain of alveolar macrophages, and to the heparin binding domain of *Mtb* organisms, facilitating the attachment to alveolar macrophages. *Mtb* proteins Fbpa, Fbpb, and Fbpc are all fibronectin-binding proteins (also known as the antigen 85 complex), and play an important role in maintaining the bacterial cell envelope and thus in the pathogenesis of tuberculosis. FbpC is the most important of the three, and inactivation of it alters the permeability of the *Mtb* cell envelope (Puech et al., 2002).

Fibronectin has also been shown to play a role in HIV, and interacts with HIV-1 protein Tat. Pre-treating HIV-1 infected cells with human fibronectin increased HIV-1 infectivity when the virus is in low concentration (Pugliese et al., 1996). The endothelium function is reduced by HIV-1 infection, due to the changes in cytokine levels and Tat, which has been shown to inhibit cell proliferation in the presence of fibronectin (Cavallaro et al., 1997). During HIV-1 infection, CD4+ T-cells are depleted. The thymic epithelium plays an important role in T-cell maturation. Fibronectin has been shown to be involved in enhancing T-cell maturation driven by the thymic epithelium. Fibronectin contains the RGD amino acid sequence, which is recognised

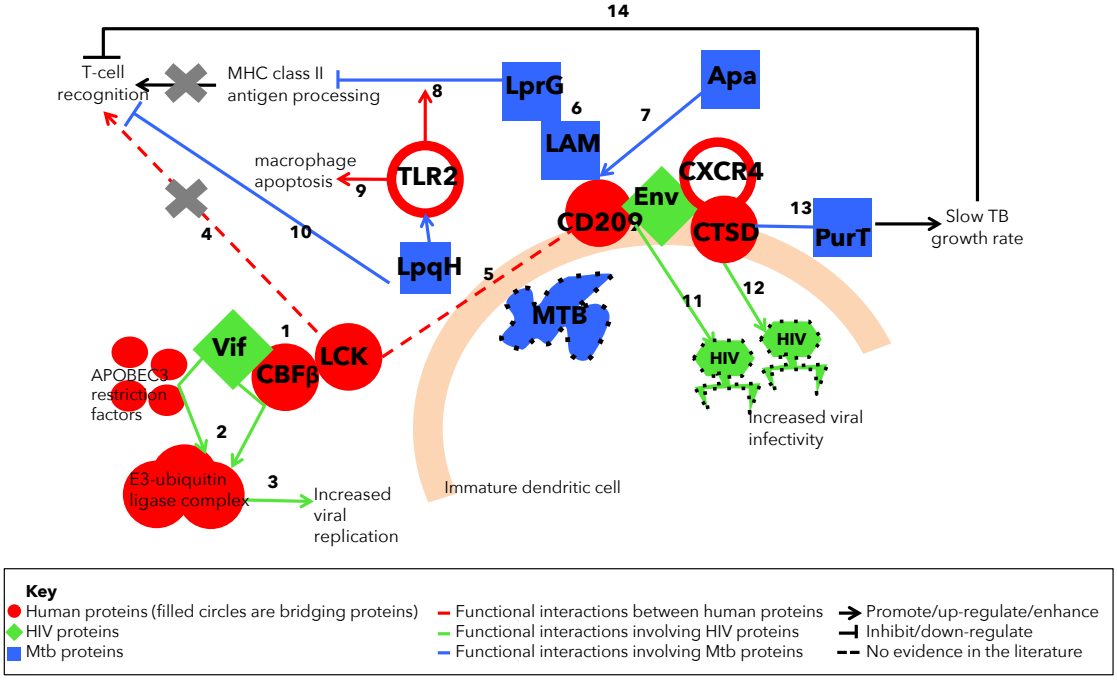


Figure 2.9 Functional host-pathogen interactions that increase viral infectivity and reduce T-cell recognition.
(1) CBFβ binds to the T-cell receptor enhancer LCK, as well as to HIV-1 protein Vif. (2) Vif hijacks CBFβ and takes it with the APOBEC3 restriction factors to the E3-ubiquitin ligase complex to be degraded. (3) Vif can increase viral replication. (4) T-cell receptor recognition may be reduced by the degradation of CBFβ. (5) LCK interacts with CD209, which is located on the surface of immature dendritic cells. (6) lprG binds to LAM, which binds to CD209 to prevent pro-inflammatory immune response and enable MTB to infect dendritic cells. (7) Recognition of *Mtb* by CD209 depends on ligand binding such as Apa. (8) LprG inhibits MCH class II antigen processing in a TLR2 dependent manner. (9) LpqH triggers TLR2 activation which induces macrophage apoptosis. (10) LpqH downregulates antigen presentation to T-cells. (11) CD209 binds to HIV protein Env on the cell membrane, which plays a role in the propagation of HIV in the cell. (12) CTSD modifies the conformation of Env enabling it to interact directly with CXCR4, which increases viral infectivity and entry into cells. (13) CTSD interacts with *Mtb* PurT which has been associated with slow growth rate in tuberculosis. (14) The slow growth rate of *Mtb* is related to the delayed T-cell recognition.

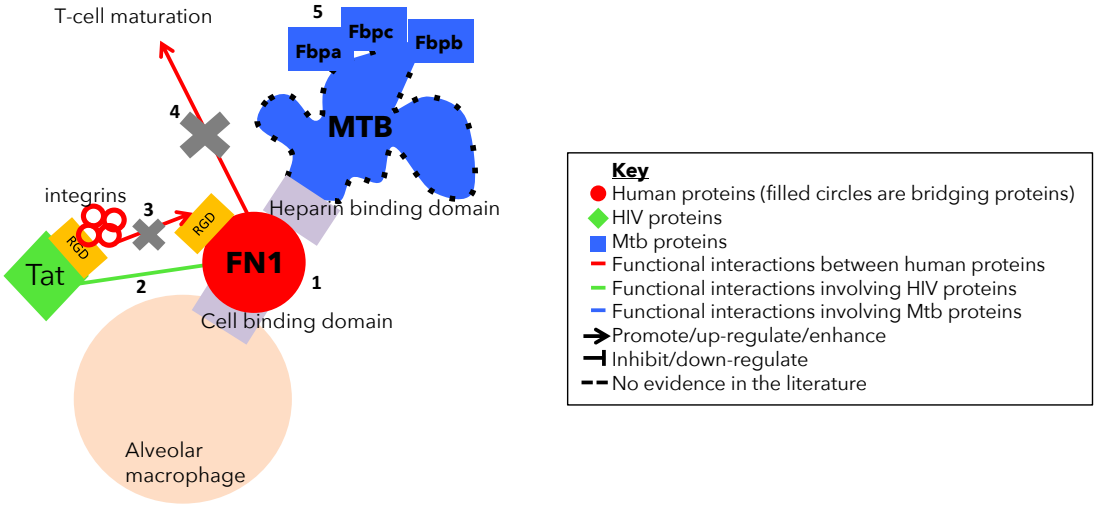


Figure 2.10 Fibronectin enables the attachment of *Mtb* to macrophages and reduces T-cell maturation. (**1**) FN1 binds to the cell binding domain of alveolar macrophages and to the heparin binding domain of *Mtb*, facilitating the attachment of *Mtb* to macrophages. (**2**) FN1 interacts with HIV-1 protein Tat. (**3**) Tat and FN1 both have the RGD amino acid sequence, which is recognised by integrins. Tat competes with FN1 for integrin binding. (**4**) This reduces T-cell maturation, which is enhanced by FN1 in the absence of HIV infection. (**5**) *Mtb* fibronectin binding proteins were predicted to interact with FN1, and are involved in maintaining the bacterial cell envelope.

by integrins as a cell attachment sequence (Maroder et al., 1996). HIV-1 Tat also contains the RGD cell adhesion site, which is also able to bind to integrin receptors (Maroder et al., 1996). Maroder et al. (1996) have shown that, in a murine thymic epithelial cell line infected with Tat, the T-cell maturation is reduced, and fibronectin is increased and redistributed relative to Tat expression. By competing with integrins for fibronectin binding, Tat is able to reduce T-cell maturation, and thus able to reduce T-cell recognition of viral proteins. The host-pathogen interactions are illustrated in Figure 2.10.

2.4.1.8 Regulating dendritic cell maturation

Cell division control protein 42 homologue (CDC42) is a plasma membrane-associated small GTPase that is active when GTP bound. According to GO annotation, CDC42 is involved in biological processes such as innate immune response, T-cell co-stimulation, and phagocytosis. CDC42 is located in the cytoplasm, which is the same cellular location as interacting HIV-1 protein Nef (located in the host cell perinuclear region of the cytoplasm). VAV is a guanine nucleotide exchange factor that activates CDC42. Quaranta et al. (2003) showed that in immature dendritic cells, Nef is able to target VAV, thereby enhancing the activation of CDC42 and another GTPase RAC1, which, in turn, results in morphological changes that make them more like mature dendritic cells. This increased the ability for dendritic cells to form clusters with T-cells.

CDC42 was predicted to interact with six *Mtb* proteins by Huo et al. (2015), namely: ClpB, ClpC1, IlvD, Kdc, Pnp, and Tuf. The *tuf* gene encodes the Ef-Tu protein, an elongation factor

that plays a role during translation and is a GTPase (Lathe and Bork, 2001). This interaction was likely predicted based on the similarity of function for the two proteins. Chopra et al. (2004) found that the *Mtb* protein Ndk acts as a GTPase activating protein for CDC42, and interacts directly with the GTP bound form to convert it to the inactive GDP bound form. This downregulation of CDC42 activity may aid in the pathogenesis of *Mtb* by slowing down dendritic cell maturation and enabling *Mtb* to persist and grow. Ndk was predicted to interact with the Tuf protein in the *Mtb*-*Mtb* PPI network, with evidence from co-expression and text-mining. Given its low confidence (STRING score = 0.27), the interaction was not included in the final high confidence functional host-pathogen PPIN. Ndk was also predicted to interact with *Mtb* protein Pnp, with high confidence, suggesting that pnp and Tuf may be involved in the similar pathways as Ndk and, through this protein, may interact with CDC42. Interestingly, Ndk was not found in either Huo et al. (2015) or Rapanoel et al.'s 2013 networks as a *Mtb* protein that interacts with human proteins.

Three of the other *Mtb* proteins predicted to interact with CDC42 have been shown to be important for regulating cell growth. ClpC1 is a regulatory ATPase that regulates the Clp protease and is essential for bacterial growth (Schmitt et al., 2011). Cyclomarin A, a naturally occurring antibiotic, has been shown to modulate the activity of ClpC1 by binding specifically and with high affinity, and having a bactericidal effect (Schmitt et al., 2011). The expression of ClpC1 and ClpB is regulated by the same regulator ClgR, and the induction of these genes has been shown to be essential for successful replication of *Mtb* in macrophages (Estorninho et al., 2010). Ilvd (dihydroxyacid dehydratase) has been shown to affect growth of *Mtb* in the lungs of infected mice, whereby downregulated IlvD resulted in reduced growth (Singh et al., 2011).

CDC42 interacts with RAC2, a plasma membrane-associated small GTPase that cycles between an active GTP-bound and inactive GDP-bound state. In the active state, RAC2 binds to a variety of effector proteins to regulate cellular responses, such as secretory processes, phagocytosis of apoptotic cells, and epithelial cell polarisation. RAC2 expression is increased in macrophages during *Mtb* infection. The differential expression patterns of RAC1 and RAC2 in *Mtb*-infected macrophages and dendritic cells may affect intracellular trafficking of *Mtb*, as well as various signaling cascades (Tailleux et al., 2003). RAC2 was predicted to interact with *Mtb* proteins ClpB, ClpC1, IlvD, Kdc, Pnp, and Tuf. These are the same proteins predicted to interact with CDC42. RAC2 has been shown to be activated by HIV-1 protein Tat, which may play a role in cell adhesion, spreading, and motility (Toschi et al., 2006).

Endoplasmin (HSP90B1) is a molecular chaperone that functions in the transport of secreted proteins. According to GO annotation, HSP90B1 is involved in biological processes such as innate immune response, and has molecular functions such as virion binding and RNA binding. The cellular component it is found in is the plasma membrane and perinuclear region of the cytoplasm – which is the same as HIV-1 protein Nef, one of its interacting proteins.

HSP90B has been shown to induce dendritic cell maturation and ICAM-1 expression, and with Nef the effect was enhanced (Mercier et al., 2013). Mercier et al. (2013) showed that these maturing HIV-1 infected dendritic cells clustered more effectively with T-cells than immature dendritic cells, and thus enhance viral transfer to T-cells. In addition, while Nef mediates CD4

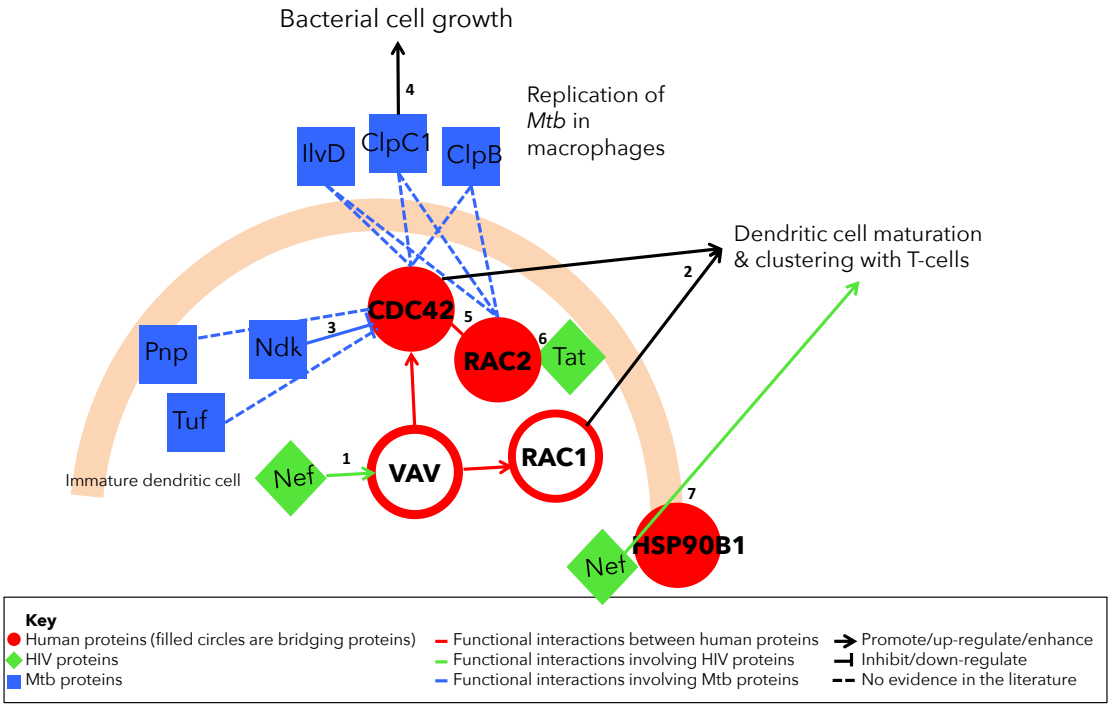


Figure 2.11 Host-pathogen interactions that regulate dendritic cell maturation during HIV-TB co-infection.
(**1**) Nef is able to target VAV and enhance its activation of CDC42 and RAC1. (**2**) Enhancing CDC42 and RAC1 changes the cell morphology to look more like mature dendritic cells, which enables clustering with T-cells. (**3**) CDC42 is predicted to interact with *Mtb* proteins PnP and Tuf, most likely via interactions with the *Mtb* protein Ndk, which is known to downregulate CDC42 in order to slow down dendritic cell maturation. (**4**) *Mtb* proteins IlvD, CLpC1, and ClpB are predicted to interact with both CDC42 and RAC2, and enhance bacterial cell growth. (**5**) RAC2 interacts with CDC42 and expression of RAC2 is increased in *Mtb* infected cells. (**6**) RAC2 is activated by Tat. (**7**) HSP90B1 interacts with Nef which enhances dendritic cell maturation and clustering with T-cells

downregulation, knockdown of HSP90B1 has been shown to increase CD4 levels in HIV-1 infected cells, but did not have a significant effect on HIV-1 replication (Landi et al., 2014). HSP90B1 was predicted to interact with *Mtb* protein ThiC; however, there is no evidence of this in the literature.

Interactions involved in regulating dendritic cell maturation are illustrated in Figure 2.11.

2.4.1.9 Response to pathogens

60 kDa heat shock protein, mitochondrial (HSDP1), also known as Hsp60, is a protein that is involved in mitochondrial protein import, and may promote correct folding of imported proteins. It is involved in biological processes such as viral process, T and B cell activation, regulation of macrophage activation, regulation of apoptotic process, B cell proliferation, and response to unfolded protein and protein refolding. HSPD1 interacts with HIV-1 proteins Nef and Gag. HSPD1 has been shown to be incorporated into the virion membrane, similarly to related protein hsp70, which interacts with Gag (Gurer et al., 2002). Gag, Nef, and hsp60 have been shown to be part of the Staufen1 RNA-binding-protein complexes in HIV-1 expressing

cells (Milev et al., 2012).

HSPD1 is predicted to interact with 10 *Mtb* proteins. clpC1, fba, groEL1, groEL2, groS, pknA, pknD, pknE, and pknF where predicted by Huo et al. (2015), while hemL was predicted by Rapanoel et al. (2013) to interact with HSPD1. hemL is a Glutamate-1-semialdehyde 2,1-aminomutase. groEL1, groEL2, and groS (also known as groES) are chaperonin proteins. groEL1 and groEL2 are hsp60 homologues in *Mtb*, which explains its prediction. groES is a *Mtb* homologue of hsp10 and is an important T-cell antigen (Qamra et al., 2005). hsp60 and hsp10 are important stimulators of the immune system (Qamra et al., 2005). As a pathogen infects the host, encounters in the host activate heat-shock proteins in response to the stress, and act as foreign antigens that elicit a strong immune response from the host (Qamra et al., 2005). groES and groEL bind to one another, which also explains the prediction of groES and HSPD1 interacting.

Like the heat-shock proteins, pkns (Serine/threonine-protein kinases) are homologous to eukaryotic proteins. This could explain the predicted interactions. PknA has been speculated to be important for growth and development. PknD and E are located in the same operon as the ATP-binding cassette (ABC) transporter gene and have been predicted to be regulators of phosphate transport (Av-Gay and Everett, 2000).

2.4.1.10 Viral transcription

HEXIM1 is a transcriptional regulator that is a RNA Polymerase II (RNA Pol II) transcription inhibitor. It binds to P-TEFb in order to prevent phosphorylation of RNA Pol II, and thereby prevents transcriptional elongation. It is predicted to interact with the *Mtb* protein hemL (Rapanoel et al., 2013), a Glutamate-1-semialdehyde 2,1-aminomutase.

Tat controls transcription of HIV-1 by RNA Pol II. Tat interacts with TAR and P-TEFb, recruiting P-TEFb to phosphorylate RNA Pol II, thus resulting in transcriptional elongation of Tat (Barboric et al., 2007). HEXIM1 binds P-TEFb by interacting with CycT1, and 7sK facilitates the binding. However, Tat disrupts 7sK snRNP and releases P-TEFb (Barboric et al., 2007). According to GO annotation, Tat is found in the host cell nucleus and host cell cytoplasm with HEXIM1.

Similar to HEXIM1, YBX1 (Y-box transcription factor) was also predicted by Rapanoel et al. (2013) to interact with *Mtb* protein hemL, and YBX1 was predicted to interact with HIV-1 protein Tat. YBX1 mediates pre-mRNA alternative splicing regulation and regulates the transcription of many genes. During HIV-1 infection, transcriptional regulation is mediated by viral and cellular factors. TAR is a RNA regulatory element found at the ends of viral transcripts and is necessary for HIV-1 transcription. Tat targets TAR. YBX1 has been shown to interact with TAR, as well as interact with Tat directly (Ansari et al., 1999). Over expressing YBX1 resulted in activation of the HIV-1 promoter by Tat, whereas Tat mediated activation was inhibited in YBX1 mutants.

2.4.1.11 Regulating T-cell activation

LCK is a non-receptor tyrosine-protein kinase that plays an important role in T-cell maturation. It was predicted to interact with *Mtb* proteins Gap, GroEL1, GroEL2, and HtpG (Huo et al., 2015). During *Mtb* infection, it has been shown that ManLAM inhibits T-cell activation by suppressing the phosphorylation of LCK at Tyr-394 (Mahon et al., 2012). Mahon et al. (2012) suggest that ManLAM might induce phosphatase activity to dephosphorylate LCK and thus inhibit T-cell activation. GroS, a *Mtb* hsp10, is an important T-cell antigen, and GroS and GroEL bind to one another (Qamra et al., 2005), which may explain the predicted interaction.

According to the HHPID, the literature suggests that LCK interacts with HIV proteins Nef, Env, Gag, and Tat. Nef, Env, and Gag are found in the host cell cytoplasm, and LCK is also located in this cellular component. Collette et al. (1996) found that Nef interacts with LCK *in vitro* and *in vivo*, and that Nef expression reduced LCK-mediated T-cell receptor signaling. In a more recent study, Pan et al. (2012) showed that Nef compartmentalises T-cell receptor signaling to adjust antigen responses. They showed that Nef does this by re-localising LCK from the plasma membrane to the trans-golgi network, thereby limiting T-cell receptor signaling at the plasma membrane, which, in turn, allows increased viral replication (Pan et al., 2012). In addition, Gag is important for viral assembly and release, which occurs at the plasma membrane (Strasner et al., 2008). Strasner et al. (2008) have shown that in addition to enhancing viral replication, LCK is important for targeting Gag to the plasma membrane in T-cells.

Nef interacts with HNRNP-K and nucleates LCK, which results in Tat-dependent HIV replication (Wolf et al., 2008). Env also inhibits LCK by phosphorylation of the auto-inhibitory tyrosine 505 residue (Morio et al., 1997). These studies and others suggest that LCK is inhibited by several HIV-1 genes, whereas during *Mtb* infection, ManLAM activates LCK, which, in turn, inhibits T-cell activation.

Protein Kinase C (PKC) was predicted to interact with two *Mtb* proteins, htpG and tuf (Huo et al., 2015). PKC regulates HIV-1 trans-activation by Tat (Jakobovits et al., 1990), and phosphorylates Tat (Holmes, 1996). According to the annotation by HHPID, PKC is a downstream effector of LCK and complexes with Nef and LCK to promote HIV transcription. This can be visualised in the generated human-pathogen PPIN in which PKC interacts with LCK, which both interact with Nef.

FYN is a Tyrosine-protein kinase that interacts with LCK and was predicted to interact with *Mtb* protein htpG by Huo et al. (2015), and with HIV-1 protein Nef. The biological processes FYN is involved in that relate to Nef function include viral process, regulation of defence response to virus by virus, innate immune response, and regulation of apoptotic process. In addition, FYN's cellular component is the plasma membrane and Nef is found in the host plasma membrane. There is evidence that the Nef-Fyn complex forms *in vivo*, and may affect T-cell receptor signaling (Arold et al., 1997).

The interactions involved in regulating T-cell activation are illustrated in Figure 2.12.

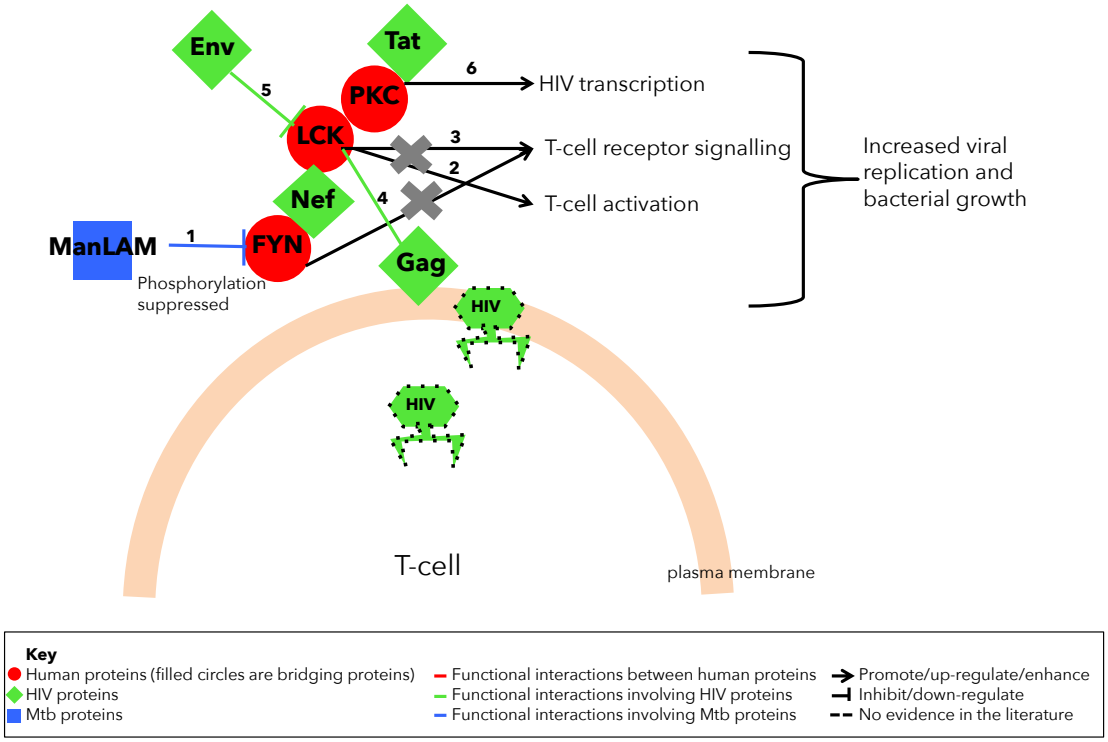


Figure 2.12 Host-pathogen interactions that regulate T-cell activation during HIV-TB co-infection. (1) During *Mtb* infection, ManLAM suppresses phosphorylation of LCK at the Tyr-394 residue, (2) which inhibits T-cell activation. (3) LCK interacts with Nef, which reduces LCK-mediated T-cell receptor signaling. (4) LCK is important for targeting Gag to the plasma membrane of T-cells, which assists in viral entry and release. (5) Env inhibits LCK by phosphorylation of Tyr 505. (6) PKC complexes with LCK and Nef to promote HIV transcription. (7) FYN interacts with Nef and LCK and the interaction with Nef may affect T-cell receptor signaling.

2.4.1.12 Protein Kinase R - a potential drug target

EIF2AK2, also known as protein kinase R (PKR), is the interferon-induced, double-stranded RNA-activated protein kinase, and functions in biological processes, including innate immune response, defence response to viral infection, and negative regulation of viral genome replication. EIF2AK2 was predicted to interact with htpG by Huo et al. (2015).

HIV-1 Tat is found in the nucleus and cytoplasm along with EIF2AK2. According to HHPID, there is evidence in the literature suggesting that Tat activates Li et al. (2005), binds, upregulates, downregulates, inhibits, and is phosphorylated by EIF2AK2. Li et al. (2005) showed that Tat induced the expression of cytokines IL-6, IL-10, and TNF-α, and that inhibiting PKR reduces Tat phosphorylation of PKR, as well as reducing TNF-α production by Tat. This is significant, as the increase of TNF-α expression in HIV-1 infected cells has been shown to lead to increased HIV-1 replication and increased *Mtb* growth, and, as such, the antimycobacterial activity of TNF-α may be reduced (Imperiali et al., 2001).

Furthermore, in *Mtb*, the absence of PKR has been shown to be beneficial to the host by reducing the viable *Mtb* in a mouse model (Wu et al., 2012b). Despite the many biological processes that PKR is involved in, mice knockouts were able to survive in its absence (Wu et al., 2012b). Wu et al. (2012b) found that PKR-deficient mice had lower *Mtb* counts, and less pulmonary pathology than wild type mice. In addition, PKR-deficient macrophages underwent more apoptosis than wild type macrophages, and Wu et al. (2012b) suggest that a PKR inhibitor could improve host immunity to TB.

Based on the role of PKR in Tat activity and the apparent reduction in *Mtb* pathogenicity in PKR deficient mice, it would be interesting to investigate how HIV infectivity is affected in PKR deficient mice, and whether co-infected knockouts also exhibit reduced *Mtb*. If reduced levels of PKR would be as viable for humans as it seemingly is in mice, and HIV is as affected by its absence as *Mtb*, this could be a very important host drug target for both pathogens. This is illustrated in Figure 2.13

2.4.2 Drug interactions

Proteins targeted by drugs had significantly higher betweenness and degree compared to non-drug targets. This is in line with the notion that higher network centrality implies higher biological importance and can be used to narrow down potential drug targets (Mulder et al., 2014). The only protein in the PPIN to have known interactions with both an anti-TB drug and an anti-HIV drug was Human Serum Albumin (HSA), which interacts with first-line anti-TB drug rifampicin and protease inhibitor HIV drug saquinavir, as well as four of the 28 bridge proteins, namely FN1, CTSD, NFκB1, and GAPDH. HSA is the most prominent protein in plasma and is known for its excellent drug binding capacity (Bocedi et al., 2004). Protease inhibitors, like saquinavir, bind to HSA, increasing serum levels. Studies have shown that rifampicin reduces the plasma concentration of saquinavir and ritonavir (which is taken with saquinavir) in HIV and TB co-infected patients, resulting in increased risk of virological failure (Ribera et al., 2007). Similar contraindications have been observed with other protease inhibitors and rifampicin (Ribera et al., 2007).

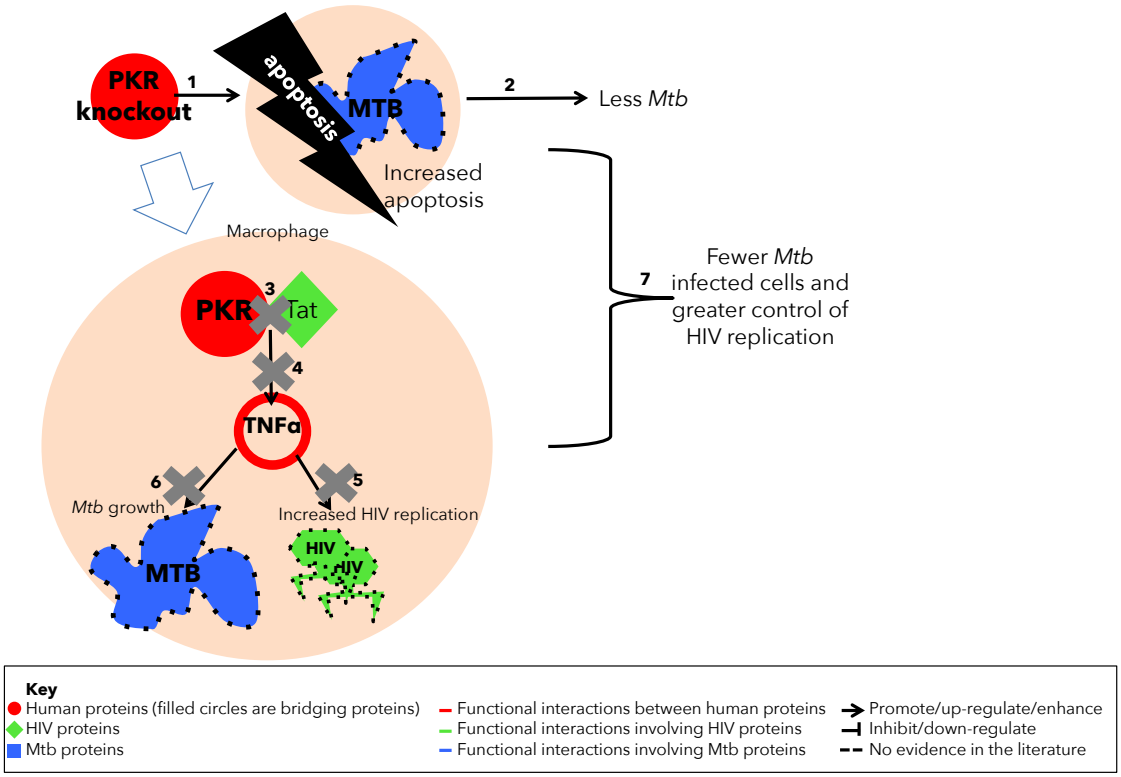


Figure 2.13 How inhibiting human Protein Kinase R may improve the host's ability to control HIV and TB during co-infection.
(1) The *Mtb* infected cells of PKR knockouts undergo more apoptosis than cells expressing PKR. (2) Increased apoptosis results in less *Mtb* in the host cells. (3) In HIV-infected macrophages of PKR knockouts, the HIV-1 protein Tat would not be able to bind to PKR. (4) Tat binding to PKR usually increases TNFα production, but this would be blocked. Blocking TNFα would, in turn, prevent the increased HIV replication, (5) as well as increased *Mtb* Growth (6) that are observed when TNFα is increased. (7) Altogether, PKR knockouts may have fewer *Mtb* infected cells and lower viral loads enabling greater control of the pathogens.

2.4.3 MHC proteins interact with proteins important for pathogen bridging

We hypothesised that the human MHC proteins would have network centrality measures that suggest they are more important than other human proteins for facilitating interactions between HIV-1 and *Mtb*. Although the MHC proteins were not found to have significantly higher or lower centrality measures in the PPIN compared to other proteins, they were found to have a significantly shorter distance to bridge proteins than other proteins in the network. This suggests that the MHC proteins may be important for regulating functional interactions between other human proteins and pathogen proteins. The apparent lack of functional interactions between MHC proteins and pathogen proteins, as well as the neutral network importance, could be explained by the variability in MHC genes. Because the MHC genes exhibit higher inter-population and intra-population variability than other human genes, they may not be sustainable targets for direct host-pathogen interactions, and are rather integral parts of pathways of interactions involving pathogen proteins. Of the 18 MHC genes that interacted with bridge proteins, 9 of them where HLA-DR and HLA-DQ proteins. Variants in the HLA-DQ and HLA-DR isotypes have been associated with TB in several populations (Lombard et al., 2006). Lombard et al. (2006) showed that polymorphisms in HLA-DR and HLA-DQ (HLA-DRB1*1302 and DQB1*0301-0304) were significantly associated with pulmonary TB in a cohort from the Venda population in South Africa. Furthermore, the HLA class II allele DRB1*1303 was shown to be associated with low plasma viral load in a South African population infected with HIV-1 clade C, as well as in a European population infected with HIV-1 clade B (Julg et al., 2011).

Another MHC gene that interacted with bridge proteins was TNF-α. TNF-α plays a role in many processes associated with immune response that are important for fighting HIV and TB. For example, TNF-α is involved in the formation and maintenance of granulomas, and can trigger cell lysis in infected macrophages (Gupta et al., 2012). Patel et al. (2007) showed that HIV and TB co-infected macrophages release less TNF-α and exhibit less TNF-dependent apoptosis than individuals only infected with *Mtb*. Furthermore, HIV and *Mtb* co-infection stimulates TNF-α production, which enhances HIV replication in macrophages (Kedzierska et al., 2003).

The observation that MHC proteins did not have significantly higher pathogenicity bridging centrality than non-MHC proteins, but did have significantly shorter distance to proteins with high pathogenicity bridging centrality than non-MHC proteins, further emphasises the usefulness of network analysis and highlights some of the subtleties. In terms of usefulness, it would have been possible and simpler to find the intersection between the human proteins that interact with TB and HIV without considering the human-human PPIs. By including human-human PPIs in the network, the importance of the MHC proteins could be identified. One of the subtleties of network analysis in this case is that the interaction edges are not weighted in terms of the biological significance of the interaction, but rather by the confidence in the interaction being real. The network centrality measures analyse the importance of the proteins based on how many other proteins they interact with; however, these interactions could be functional associations that exist but are not crucial for the biological process the proteins share. Thus, the MHC proteins play an important role in HIV-TB co-infection without any known functional associations with both pathogens in the

datasets included in this analysis, which may also be due to the limited knowledge of interactions between *Mtb* and human proteins.

2.5 Conclusion

This analysis has produced a comprehensive and high confidence functional protein-protein interaction network, that to our knowledge is the first to contain inter- and intraspecies interactions between human, HIV, *Mtb* and drug target interactions. Using the network, we were able to identify 28 human proteins that functionally interact with both pathogens, of which 27 have literature that supports the biological plausibility and, in some cases, importance of these interactions. The number of proteins identified and support for biological plausibility from the literature was largely affected by the limited datasets containing known or predicted *Mtb*-human protein interactions. In chapter 4, we will overlay gene expression data from HIV-TB infected cases and uninfected controls to determine whether or not the identified human proteins that may act as "pathogen bridges" are differentially expressed.

3. Identification of genetic variation in the MHC region using a sequence graph

3.1 Introduction

According to the Out of Africa model, humans originated in Africa and then dispersed across the rest of the world over the last 100 000 years (Campbell and Tishkoff, 2008). As such, human populations in Africa are the most genetically diverse (Campbell and Tishkoff, 2008). The pattern of genetic variation is affected by changes in population size, admixture, migration, exposure to pathogens in addition to natural selection, recombination, and mutation (Campbell and Tishkoff, 2008). As discussed in chapter 1, the current human reference genome (GRCh38) fails to capture inter-population variation in highly polymorphic regions of the genome such as the human major histocompatibility complex (MHC), particularly in African populations, which are highly divergent from the reference. The variability of the MHC region has made it a prominent genomic region of interest for identifying graph-based methods of genome representation to replace the current linear reference (Dilthey et al., 2015). The desire to unpack the MHC region is driven by the failure of the current linear reference genome representation to capture the variability between individuals and ancestrally diverse populations, coupled with the knowledge that variants in this region are strongly associated with several diseases, including HIV and TB (Dilthey et al., 2015).

Two categories of graph-based approaches have emerged: (1) augment the existing reference to incorporate variation using a graph structure (Paten et al., 2014), and (2) *de novo* assemble a graph using whole genome sequences from diverse populations with or without the reference (Turner et al., 2017). In both cases, the sequence graph can then be used for alignment by effectively simultaneously mapping reads to all references and selecting the path that best fits the data (Paten et al., 2014; Turner et al., 2017). Recent publications suggest that the first approach of augmenting the reference with known variation might achieve similar outcomes in a more sustainable way that is a natural progression from working with the linear reference genome (Rakocevic et al., 2019). However, when this thesis was started in 2015, a variety of graph-based methods for variant identification and genome assembly were being developed (Dilthey et al., 2015; Novak et al., 2017). The decision was taken to try and eliminate as much reference bias as possible by choosing a method that does not rely on the existing reference for construction and only uses it as a coordinate system for comparison post-construction (Turner et al., 2017). These reference-free methods for sequence graphs will be introduced, and later in this chapter the results will be discussed with comparison to the latest results that have been published.

3.1.1 de Bruijn graphs and sequence graphs

There are two widely used methods for assembling genomes, (1) overlap consensus layout (OLC) and (2) de Bruijn graphs. A visualisation of these two methods is available in Figure **??**.

OLC was first developed by Staden (1980). The method identifies overlaps between the sequencing reads, then scaffolds them to come up with a layout from which the consensus sequence can be determined (Li et al., 2011). The length of the sequencing reads used by OLC requires extensive computational resources, especially when sequences are high coverage.

The de Bruijn graph algorithm for DNA assembly was first introduced by Idury and Waterman (1995), and has become adopted as the primary algorithm used for genome assembly with the onset of next generation sequencing, which has resulted in multiple short reads (Li et al., 2011). A de Bruijn graph is constructed by splitting up the input reads into overlapping pieces of a fixed length (k), known as k-mers. The k-mers overlap, with one k-mer starting at every base, such that, for each read, the number of k-mers is equal to the read length $-\ k + 1$. The de Bruijn graph can then be constructed using the set of k-mers as the vertices, and the overlaps between them as the edges, where the edges cannot be parallel and the edges must share $k - 1$ bases. To avoid DNA sequence "palindromes" caused by double stranded DNA, k must always be an odd number. Splitting the reads into smaller k-mers increases the computational efficiency, making de Bruijn graphs a most widely used algorithm for *de novo* assembly (Li et al., 2011). The methods of assembly using OLC and de Bruijn graphs are illustrated in Figure 3.1. Sequence graphs are an extension of *de novo* assembly methods, which use the de Bruijn graph approach to assemble next generation sequencing reads (Turner et al., 2017). Most genome assemblers utilise de Bruijn graphs for assembly of reads and variant identification relative to a reference genome (Li et al., 2011). An assembler called Cortex that uses de Bruijn graphs, has been used for reference-free variant identification by building a map of genome variation within *Plasmodium falciparum* (Iqbal et al., 2012; Miles et al., 2015).

3.1.2 The linked de Bruijn graph

Traditionally, de Bruijn graphs are built up one read at a time at the cost of storing the graph in memory, as well as sacrificing long-range information in the read (Turner et al., 2017). Turner et al. (2017) proposed an advance to the traditional de Bruijn graph method, which they refer to as the "linked de Bruijn graph". In the linked de Bruijn graph, each vertex has a set of paths (referred to as links) through the graph that start at that vertex. Each link acts as a list of junction choices that when followed will recreate the given path. In addition, as the graph is traversed, the links record how many edges ago they were picked up (this is referred to as the "age" of the link). The next junction of the oldest link is then followed as this gives the most context about the current genome position. This allows the capturing of repeat sequences in the final assembly, which would have otherwise been missed in a de Bruijn assembly (Turner et al., 2017). An example of how a linked de Bruijn graph assembly could predict a sequence more accurately than an ordinary de Bruijn graph assembly is depicted in Figure 3.2.

3.1.3 Multi-coloured linked de Bruijn graphs

Turner et al. (2017) have extended the *Cortex* assembler developed by Iqbal et al. (2012) to use the linked de Bruijn graph method in their program *McCortex*, which can be used for

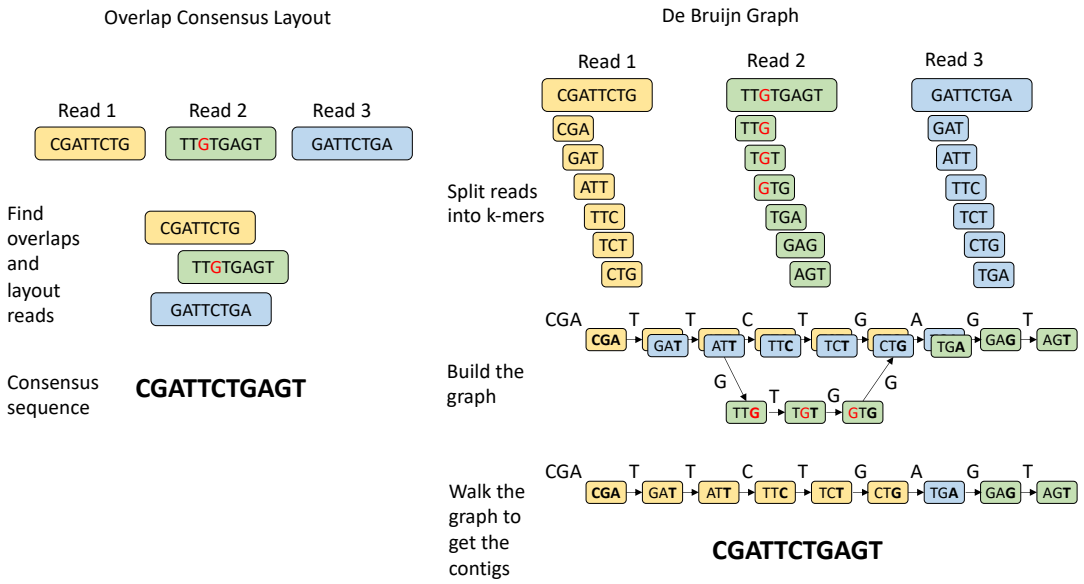


Figure 3.1 Overlap Consensus Layout vs. de Bruijn Graph sequence assembly.
This figure was adapted from Ayling et al. (2019) to illustrate the difference between Overlap Consensus Layout (OLC) and de Bruijn graph assembly. OLC involves finding overlaps between the reads, and laying them out to identify the most likely consensus sequence. Read coverage is used to determine the most likely sequence if there is variation as denoted in the figure. de Bruijn graph assembly involves splitting the reads into shorter equally odd lengthed fragments (called k-mers) that overlap by one less base than their length k. In the image the length of k is 3. Thereafter a graph is constructed by making each k-mer a node and connecting them with edges that denote the overlaps. Any variation (in this case a SNP) becomes another path in the graph. Traversing the graph can yield two paths, and the consensus sequence can be determined based on the k-mer coverage.

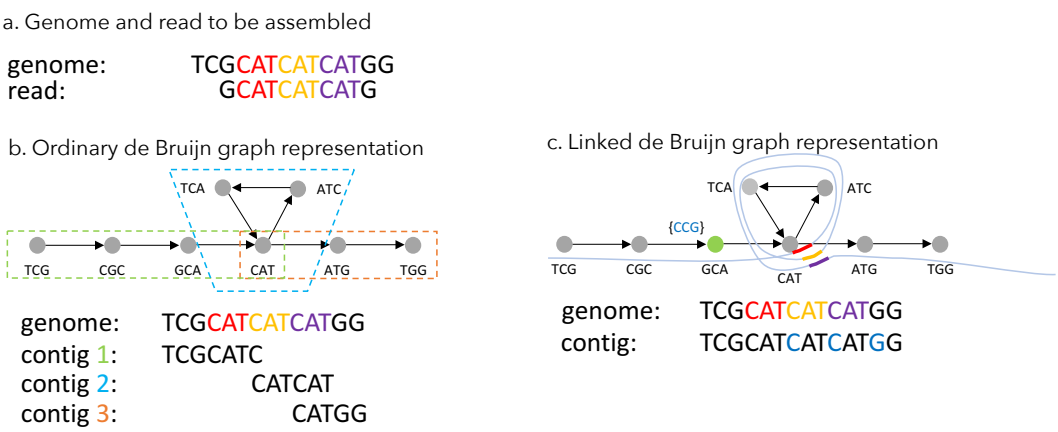


Figure 3.2 The usefulness of retaining link information when traversing de Bruijn graphs.
This figure was adapted from Turner et al. (2017) to explain the usefulness of retaining link information when traversing de Bruijn graphs in order to assemble a genome. (a) A 14bp genome that requires assembly, as well as one of the sequencing reads that spans a repeat in the genome. Each repeat is colour coded. (b) A de Bruijn graph constructed by splitting the genome reads up into k-mers of length=3. The k-mers are the grey nodes of the graph, and the arrows represent the edges of the graph. The dashed boxes correspond with 3 possible contigs listed below the graph, which could be scaffolded in various ways to assemble a possible genome, but not all assemblies will correctly represent the genome. (c) The same de Bruijn graph, except the traversed path uses link information from the read to inform the contig constructed. The link information for the repeat junction is stored in the k-mer before it (coloured in green), the letters in curly brackets indicate that the path needs to go through the repeat loop twice (depicted by the blue line).

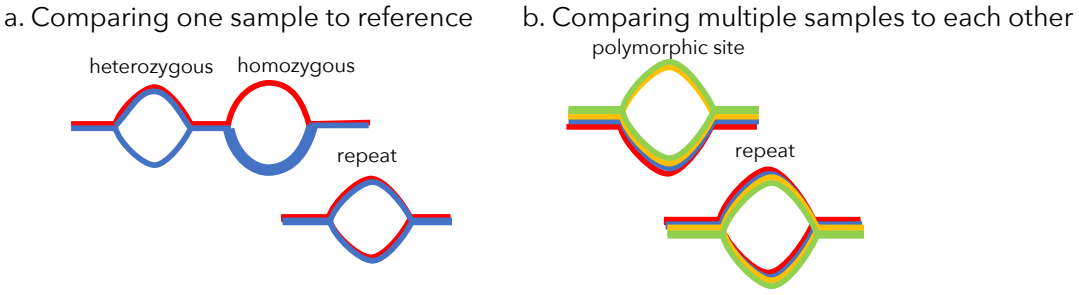


Figure 3.3 Analysing variants from bubbles in de Bruijn graphs.
This figure was adapted from Iqbal et al. (2012). (a) When comparing a diploid sample (blue lines) to a reference (red lines), you can detect heterozygous alleles where both sides of the bubble have blue lines. Similarly, you can detect homozygous variants where only one side of the bubble has a blue line. Repeats are detectable where both sides of the bubble contain both blue and red lines. (b) When comparing multiple samples, if only some colours are on one side of the bubble you can detect a polymorphic site. Repeats are detectable where both sides of the bubble contain colours from all samples.

population *de novo* assembly and variant calling. In this method, first each individuals' reads are *de novo* assembled into a linked de Bruijn graph. Then the individual graphs are combined into a single graph, with each k-mer carrying information regarding which individual it was observed in. Each individual is assigned a unique colour. To identify variation between individuals, "bubbles" in the graph are detected where both branches or sides of the bubble are present in the same colour (Iqbal et al., 2012; Turner et al., 2017) (see Figure 3.3). Once bubbles are detected, they can be converted into VCF format and mapped to a linear reference genome to identify positions and genotypes where possible. This enables comparison with variants identified through traditional variant calling mechanisms.

3.1.4 Aims and hypotheses

In this chapter, we describe how *McCortex* was used to construct a coloured de Bruijn graph of chromosome 6 for 33 sequences of African ancestry. We describe how the graph was *de novo* assembled using reads that previously mapped to chromosome 6, as well as unmapped reads, so as to not lose sequences that may have been unmapped due to divergence from the reference. The resulting graph was used to identify variation in the chromosome, including the MHC region. We hypothesise that (1) reads that were previously unmapped when aligned to the reference are incorporated into the reference graph; (2) new variants are identified that were not previously identified from calling the variants against the linear reference; and (3) less variants will be identified using the graph than from calling the variants against the reference as the individuals are ancestrally more similar to each other than to the reference.

3.2 Methods

Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team (http://hpc.uct.ac.za). Ethical approval was received from

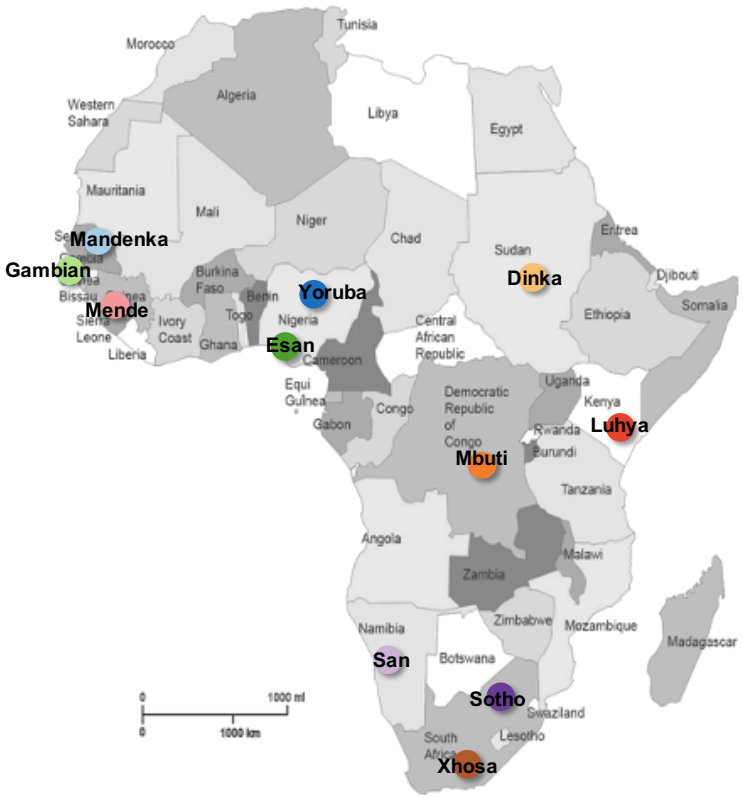


Figure 3.4 Geographical distribution of the individuals whose genomic data were included. The coloured circles each uniquely represent the population for which whole genome sequence data was available. The SAHGP granted access to whole genome sequences and variant data for eight Xhosa and eight Sotho individuals. The SGDP granted access to whole genome sequences for two individuals from each of the following populations: San, Mbuti, Mandenka, Dinka, and Yoruba. From 1000G, seven high coverage whole genome sequences were downloaded for the following populations: Yoruba (*n*=3), Luyha (*n*=1), Gambian (*n*=1), Mende (*n*=1), and Esan (*n*=1).

the University of Cape Town Human Research Ethics Committee (Reference number: HREC455/2017)

3.2.1 Whole genome sequence and variant file collection

A dataset of 33 high coverage whole genome sequences of individuals from African populations was compiled from three publicly available sources, namely the Simons Genome Diversity Project (SGDP), the Southern African Human Genome Programme (SAHGP), and the 1000 Genomes Project (1000G) (1000 Genomes Project Consortium, 2012; Choudhury et al., 2017; Simons Foundation, 2014). A detailed description of the data downloaded from each of the sources is provided in the paragraphs that follow, and the geographical distribution of the individuals whose genomic data were included is depicted in Figure 3.4 and detailed in Table 7.1.

3.2.1.1 SGDP sequences - sample origin and sequencing techniques

Ten whole genome sequences in BAM format and variant call files (VCFs) from five African populations were downloaded from the SGDP pilot project using *Globus online* (Foster, 2005). The pilot dataset is divided into two panels – Panel A and Panel B, each of which contains five African genomes including a San, Mbuti, Mandenka, Yoruba, and a Dinka individual (Prüfer et al., 2014). All of the samples were sequenced using the *Illumina HiSeq* next generation sequencing platform, and the sequenced reads were mapped to GRCh37/hg19 (extended with the Epstein-Barr virus) (Meyer et al., 2012; Prüfer et al., 2014). Average coverage was between 24 and 33-fold for the African genomes sequenced in Panel A (Meyer et al., 2012), and between 35 and 42-fold for Panel B (Prüfer et al., 2014). At the time, high coverage whole genome sequences for Panel C of the SGDP dataset, which included an additional 45 sequences from African individuals, were not available (Prüfer et al., 2014).

3.2.1.2 SAHGP sequences - sample origin and sequencing techniques

The SAHGP provided 16 high coverage (45-fold) whole genome sequences in BAM format and associated VCFs corresponding to two populations in South Africa: Sotho and Xhosa (Choudhury et al., 2017). These were sequenced on the *Illumina* next generation sequencing platform, using paired-end reads, and aligned to GRCh37/hg19.

3.2.1.3 1000G sequences - sample origin and sequencing techniques

The 1000G project provide thousands of low coverage genome sequences for 27 populations, including five African populations and two populations with admixed African ancestry (1000 Genomes Project Consortium, 2012). The seven high coverage sequences from African populations were included in this analysis, which provided sequences from the following five African populations: Yoruba in Ibadan, Nigeria (YRI); Luhya in Webuye, Kenya (LWK); Esan in Nigeria (ESN); Gambian in Western Divisions in the Gambia (GWD); and Mende in Sierra Leone (MSL). The VCF files and CRAM files for these sequences were downloaded and added to the dataset (see Table 7.1). YRI was the only population group with multiple high coverage whole genome sequences, with the trio of YRI sequences belonging to a related mother, father, and child.

3.2.2 Read extraction and preparation

Due to the large amount of read data, short read length, and the computational overhead of *de novo* assembly, sequencing reads are assembled by aligning to a reference sequence. After alignment, the reads are most often stored in BAM or CRAM format, which are binary versions of the human readable Sequence Alignment / Mapping (SAM) format. SAM files store the list of the reference information, followed by the list of alignments, usually indexed and sorted by the coordinates of the mapped reads to enable fast retrieval of alignments. Although this format is useful for downstream applications, for the purpose of this project, we wished to *de novo* assemble chromosome 6. To this end, we made use of the indexing of the BAM files to extract unmapped reads and reads mapping to chromosome 6, and convert

them into the unassembled Fastq format. The 1000G sequences were available as Fastq files, but, for consistency with the other sequences, the unmapped reads and reads that map to chromosome 6 were extracted from the aligned CRAM files.

3.2.2.1 Extraction and format conversion of reads

SAMtools is a set of tools for the post-processing of sequencing reads in SAM, BAM, and CRAM formats (Li et al., 2009). *SAMtools* was used to extract the reads mapping to the specific chromosome or unmapped reads directly from the original files. Thereafter, the files containing the read subsets were sorted by read name instead of by coordinate position before converting them to Fastq format. The sort was done to ensure that the reads are not in biased order before re-alignment. If the original BAM is not shuffled, the blocks of insert size will not be randomly distributed across the genome.

To assess the quality of the extracted reads, *FastQC* (Andrews, Simon, 2010) was run on each of the extracted read files, which all passed the overall statistics. The read length ranged between 94 and 101 base pairs, and in some cases varied slightly between the first-end and second-end reads. A detailed breakdown is available in Table 7.4, in which R1 and R2 denote the read length range for the first-end and second-end reads respectively. The percentage GC content ranged between 39% and 40% for chromosome 6, and between 39% and 45% for the unmapped reads.

3.2.2.2 Detection and removal of contamination in unmapped reads

Most unmapped sequence reads are from the target genome, but have not been aligned to the reference as they contain significant variation, sequencing errors or chimeric sequences (Tae et al., 2014). The remaining reads may be present due to contamination during sample preparation and sequencing. The contamination needs to be removed in order to preserve the integrity of downstream analysis, and to prevent errors in sequence assembly. Since our reads mapping to chromosome 6 had already been aligned to the human genome (GRCh37/hg19), it was necessary to inspect whether the unmapped reads aligned to human sequences or were non-human contamination. To this end, the standalone version of *DeconSeq*, a publically available tool that can detect and remove sequence contamination, was used (Schmieder and Edwards, 2011). *DeconSeq* was used over *BLAST* as it is based on the *BWA-SW* (*Burrows Wheeler Aligner*), which was shown to be over 10 times faster and is also able to handle corresponding regions such as gaps and variants (Li and Durbin, 2010). The *BWA-SW* alignment algorithm searches for the first significant match and then stops, whereas *BLAST* searches for all matches making it slower (Li and Durbin, 2010). Unlike in the original *BWA-SW* code, in which a N is replaced with a random nucleotide (A, G, C or T), *DeconSeq* uses modified code in which the base N will always mismatch the sequences in the reference databases, and sequences containing N's can therefore never be aligned with 100% identity unless the N's are overhanging the alignment start and end position (Schmieder and Edwards, 2011).

DeconSeq was used to extract the reads that aligned to databases of human DNA by treating

human DNA sequences as the "contaminants". A database of human reads was created using the sequence for alignment pipelines provided by Genbank GCA_000001405.15_GRCh38_full_plus_hs38d1_analysis_set.fna.gz (ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh38/seqs_for_alignment_pipelines/, Accessed 15 July 2015). This file contains all of the chromosomes from the GRCh38 primary assembly, the mitochondrial genomes, unlocalised scaffolds, unplaced scaffolds, the Epstein-Barr virus, alternate scaffolds from the GRCh38 ALT_REF_LOCI assembly units, as well as human decoy sequences from hs38d1 (GCA_000786075.2). To create the database, first, long stretches of N's were split from the sequences using command line perl. Then, BWA was used to create the database files (http://deconseq.sourceforge.net/faq.html).

DeconSeq compares the similarity of the sequences to specified databases, whereby the similarity is based on sequence coverage and alignment identity. Coverage is a measure of how much of the query sequence is similar to the database sequence and identity is a measure of how similar the query and the database sequence are to one another. The identity threshold was set to 98%, which specifies an average error rate of 0.01 [Threshold %=100% - (error rate +1% margin)]. Since the sequencing data was high coverage (>30-fold), 95% was chosen for the threshold coverage (Schmieder and Edwards, 2011). Thereafter, *DeconSeq* was run for every sequence, for both the first-end and second-end reads, by setting the created GRCh38 database as the remove database. As such our "contaminated" output file contained the reads that aligned to human DNA sequences, while the "clean" output file contained reads that did not align to human DNA sequences.

3.2.2.3 Merging overlapping paired-end reads

Paired-end reads often share overlapping sequences. These overlapping paired-end reads should ideally be merged prior to *de novo* assembly so that the long-distance connectivity information in the reads is not lost. *FLASH* (Fast Length Adjustment of SHort reads) was used to merge the overlapping read pairs, as the tool has been shown to reduce the number of misassembled contigs, as well as provide higher N50 values of contigs than other methods (Magoč and Salzberg, 2011). *FLASH* outputs a file containing the merged reads and two files containing the first-end and second-end reads that did not overlap. In some cases, there were different numbers of reads in the first-end and second-end read files, caused during reverse engineering from the BAM files which included unmapped reads. *FLASH* requires the first-end and second-end files to be perfectly aligned, and, as such, it was only run on reads that were present in both files. These were identified and extracted from the original files using *SEQTK* (Li, 2015). The reads that were only present in either the first-end or second-end read files were treated as single-end reads during graph assembly. On average, *FLASH* identified and merged 11.16% of reads with overlaps (standard deviation 15.86%).

3.2.3 PCA of chromosome 6 variants

In order to analyse the population structure of the samples included in this project, principal components analysis (PCA) was run on the SNPs in chromosome 6 that had been genotyped

by all of the contributing projects. Variants within the MHC region were excluded (chr6:28 477 797-33 448 354 in GRCh37/hg19), as highly variable regions and SNPs in linkage disequilibrium can skew the results of PCA. Initially all variants were included and a null value of 9 was assigned as the genotype for individuals where the variant had not been genotyped. This resulted in individuals being separated by project rather than by the population structure. As such, only variants that were genotyped in all individuals and in all of the projects were included. In addition, one individual (NA19240 from 1000G) was excluded from the PCA as this sequence is from the child of two of the other sequences (NA19239 and NA19238).

EIGENSTRAT's smartpca, a program within the *EIGENSOFT* package of tools, was used to conduct the PCA (Price et al., 2006). *EIGENSTRAT*'s smartpca can be used to apply PCA to genotype data to infer continuous axes of genetic variation (Price et al., 2006). Each axis of variation explains as much of the variability as possible, and often separates the data points graphically in terms of geographical location in samples with diverse ancestry (Price et al., 2006). *EIGENSTRAT*'s smartpca requires three input files: (1) a file of identifiers for the individual; (2) a SNP file, which contains the SNP identifier, the chromosome, the allele frequency, the chromosome position, the reference allele, and the alternate allele; and (3) a genotype file, which is essentially a genotype matrix that contains a column for each individual, and a row for each SNP with a genotype value of 0,1,2 or 9, corresponding to two copies, one copy, no copies or missing reference allele respectively. After extracting the relevant regions from the VCF files (Danecek et al., 2011), the VCF files were transformed into the *EIGENSTRAT* format for each individual. Thereafter, the files were filtered on postions that had been genotyped in each individual and the individuals' genotype files were combined into the genotype matrix. The final PCA was run for 32 individuals over 32 793 SNPs using smartpca version: 13050. *Genesis* was used to visualise the smartpca output (Buchmann and Hazelhurst, 2015).

After identifying variants from the graph, PCA was run on the newly identified SNPs (with the same region and sample excluded) so that the PCA plots could be compared. The final PCA was run for 32 individuals over 612 299 SNPs using smartpca version: 16000.

3.2.4 *De novo* assembly into a coloured multi-sample linked de Bruijn graph

McCortex was used to *de novo* assemble the sequences into a coloured multi-sample linked de Bruijn graph (Turner et al., 2017). Each sequence was assigned its own colour, enabling the identification of which samples have which k-mers and links. In this section, the *McCortex* workflow that was used will be described. A summary of this workflow is depicted in Figure 3.5.

3.2.4.1 Length of k-mers

The first step in constructing the graph is to split the reads into k-mers. The length of the parameter k needs to be carefully considered when constructing a de Bruijn Graph as the downstream analyses are highly sensitive to it. For example, increasing the length of k enables the graph to span short repeats; however, it also reduces the number of k-mers

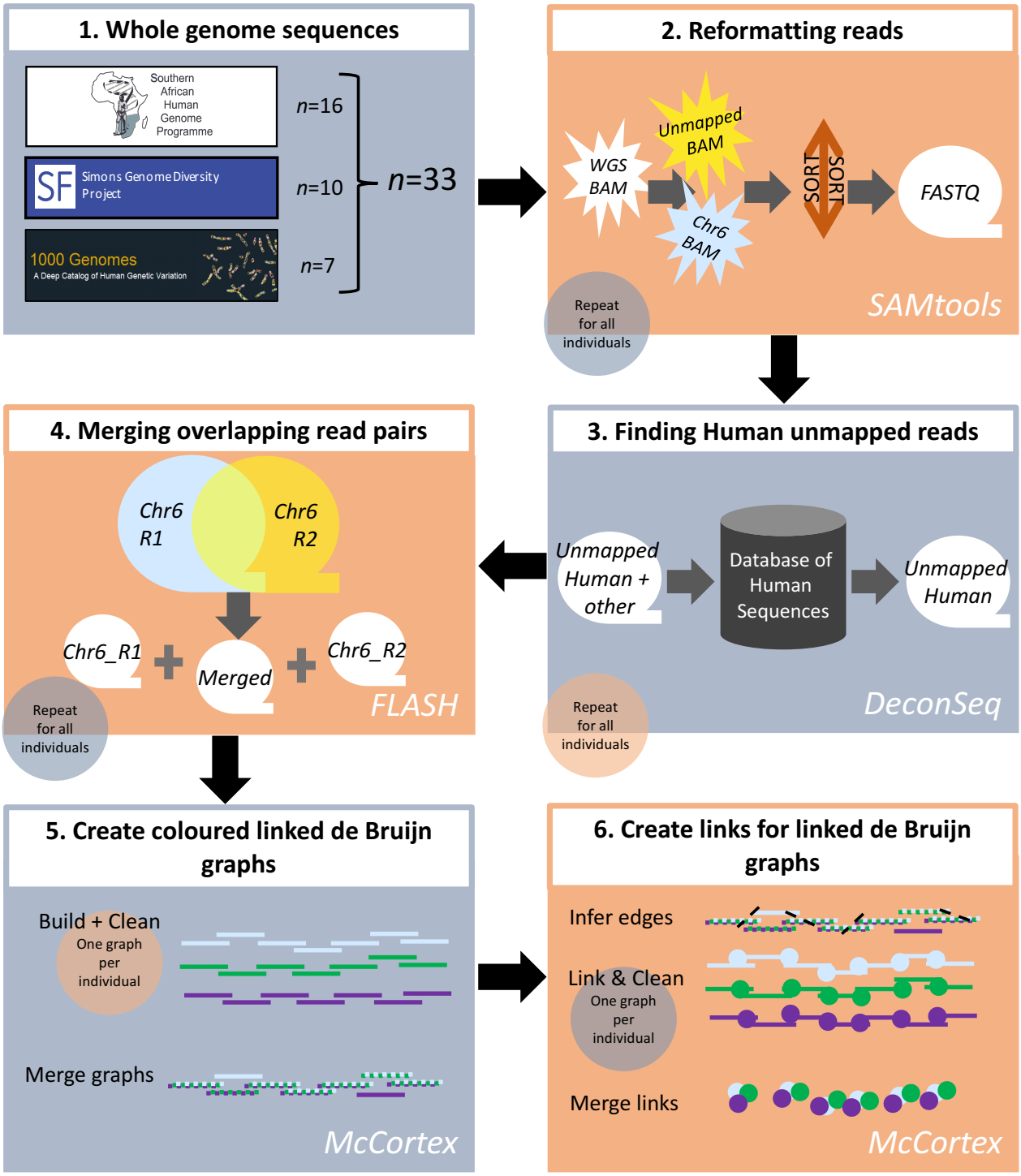


Figure 3.5 Constructing the multi-sample coloured linked de Bruijn graph. (1) 33 Whole genome sequences were acquired from three sources - the SAHGP, SGDP, and the 1000G project. (2) Reads mapping to chromosome 6, as well as unmapped reads, were extracted from the BAM files. The BAM files were then sorted by read name rather than position, and then converted to Fastq format using *SAMtools*. (3) *DeconSeq* was used to align the unmapped reads to a database of human sequences in order to isolate potentially human reads from other contamination. (4) Overlapping paired-end reads from chromosome 6 were merged using *FLASH* resulting in three Fastq outputs, including merged reads, first-end reads with no-overlaps, and second-end reads with no overlaps. (5) *McCortex* was used to build and clean a linked de Bruijn graph for each individual. These graphs were constructed using the unmapped human reads, the non-overlapping paired-end reads, and the merged reads. Then the individual linked de Bruijn graphs were merged into a 33 colour multi-sample graph. (6) Links between adjacent k-mers in the multi-sample graph where inferred. Then the reads were threaded for each individual against the merged graph and links between the k-mers were identified and cleaned. The links from all individuals were then merged into a combined links file.

available per read and increases the number of k-mers lost to sequencing error. In Iqbal et al. (2012), for 30-fold coverage, the optimum length of k was found to be 55. Because some of the sequences included were less than 30-fold coverage, but all were high coverage, k was set to 51. Publications after this work was started showed that the linked de Bruijn graph method is less sensitive to the length of k (Turner, 2019; Turner et al., 2017).

3.2.4.2 Constructing individual graphs

For each of the 33 individuals, graphs of chromosome 6 were constructed separately using the build command in *McCortex*. *McCortex* enables graphs to be built from a list of sequences, whereby the user indicates whether each sequence in the list was paired-end or single-end. The unmapped reads that were processed using *DeconSeq*, reads only present in the first-end or second-end file, as well as overlapping reads that were merged using *FLASH*, were treated as single-end reads. Unmerged paired-end reads outputted by *FLASH* were the only reads treated as paired-end reads during the build. On average, the build of chromosome 6 per individual took 5399 seconds (σ=5039 seconds), and required between 24GB and 96GB of RAM.

3.2.4.3 Filtering subgraphs of likely Chromosome 6 k-mers

To prevent stretches of graph comprised solely of unmapped reads that shared no edges with the rest of the graph, the graphs were filtered to remove k-mers that could not be aligned to chromosome 6 reads. The memory requirements ranged from 14GB to 100GB of RAM per graph, and on average 82.76% (σ=10.32%) of all k-mers could be aligned to chromosome 6 reads.

3.2.4.4 Error-cleaning graphs

Thereafter, individual graphs were cleaned to remove potential sequencing error. The program recommends a k-mer coverage cleaning threshold for the k-mers such that the false negative rate would be 0.001. For the samples included, the cleaning threshold ranged between a k-mer coverage of seven and 11-fold, which means that k-mers with a coverage less than the given threshold were removed. On average, 73.47% (σ=9.48%) of all k-mers were removed from the initial graphs resulting in between 129 103 975 and 173 180 697 k-mers per graph.

3.2.4.5 Merging graphs into a multi-sample graph

After individual graphs were constructed, *McCortex* enables the merging of graphs into a multi-coloured de Bruijn Graph. There is the option to load individual graphs as different or the same "colour" whereby a colour can be used to differentiate individuals or populations. Because we were interested in inter-individual variation and only usually had one or two sequences per population, each individual was treated as its own colour. This resulted in a 33 colour graph. After merging the graphs, a distance matrix was generated to identify how many k-mers were shared between each pair of individuals.

3.2.4.6 Threading the reads

After assembling k-mers, the edges between adjacent k-mers were inferred using *McCortex*, resulting in a connected graph. It should be noted that edges are not stored per individual graph, instead the edges are stored per k-mer and include a record of which individuals had those edges. *McCortex* was then used to generate "links" that connect k-mers to each other across collapsed repeats, which allows paths to be created. This is done using the thread command, which corrects reads for errors so that they match the graph and then identifies sequences matching the graph to generate the links. The thread command generates "link" files , which are plain text documents that contain a line per k-mer followed by lines per links that start at that k-mer. The link line contains the following information:

1. If the link starts with the k-mer in the forward or reverse strand
2. The number of junctions or forks in the graph that the link spans
3. The number of times the link is seen in the sample
4. The junction choices made by the link, listed as bases e.g. if the junctions are "AT", at the first junction take the edge starting with A, and at the second junction take the edge starting with T.

3.2.4.7 Cleaning graph links

After generating the links, they were cleaned to remove redundant and rare links. This reduces memory requirements for storing the graph connectivity information. The first 1000 k-mers with links were used to choose a cleaning threshold and then the links with a coverage less than that threshold are removed (average threshold was 10-fold, σ=1). Finally, the links for each individual could be merged into one links file that retains the original colour information for each individual, allowing the extraction of paths per individual, as well as the comparison of paths.

3.2.5 Identifying and annotating variants

3.2.5.1 Identifying variants

McCortex was used to call variants from the graph using the "bubble caller". The bubble caller identifies bubbles in the graph using the graph and links files. Thereafter, the putative 5' flanks of each variant were extracted and mapped to the reference (GRCh38) using BWA (Li and Durbin, 2010). The alleles were then aligned to the reference to generate a VCF file. During alignment, the minimum MAPQ threshold was set to 0, so that bubbles that aligned to more than one region would not be excluded. The rationale behind this was that the reference for chromosome 6 has alternate regions, and it is therefore plausible for a bubble to align to more than one location. When this parameter was left at the default threshold value of 30, few of the MHC variants could be identified, as they aligned to the reference, alternate contigs, and the HLA genes that were included in the reference file. After annotating with

McCortex, all variants in bubbles that had MAPQ scores $<$30 or mapped to more than one position on the reference chromosome 6 were excluded. If a variant mapped to a position on chromosome 6 and to a position on one of the alternate loci, then the variant was included.

3.2.5.2 Post-processing variants

The VCF file was post-processed using *BCFtools* to sort, remove duplicates, rename variants, compress, and index the VCF (Li et al., 2009). *McCortex* was used to annotate the VCF with k-mer coverage in the graph in order to indicate the average coverage of k-mers unique to the reference and the average coverage of k-mers unique to the alternative allele. Finally, the variants were genotyped using the expected k-mer coverage.

3.2.5.3 Combining variants called against the reference from the contributing projects

We obtained VCF files for each individual that contained the variants that were called against the reference by each project. The difference between these variants and variants identified in this project is that they were identified using sequences that were assembled by aligning to a reference, whereas here we have *de novo* assembled the sequences prior to variant calling. Each of these VCF files contained variants with genomic coordinates for GRCh37. These coordinates were mapped to GRCh38 using *BCFtools*. Thereafter, the unique combinations of genomic position and variant change were identified, so that they could be compared to the variants identified from the sequence graph.

3.2.5.4 Annotating variants

Once the variants were identified from the graph, *ANNOVAR* was used to annotate the variants and identify novel variants (Wang et al., 2010). *ANNOVAR* is an efficient tool that functionally annotates genetic variants. *ANNOVAR* was used for gene-based annotation, which maps variants to genes and identifies whether or not variants cause protein coding changes. RefSeq gene for GRCh38 was the database used for gene-based annotation. In addition, filter-based annotation was used to identify whether or not variants had previously been identified in any of the following databases:

1. dbSNP: A database of human genetic variation, including single nucleotide variations, microsatellites, insertions and deletions, annotated with information such as population frequency, molecular consequence, and genome coordinates (build 150) (Sherry et al., 2001).
2. 1000G: A database of genomic variation identified in over 1000 genomes (based on 201508 collection v5b) (1000 Genomes Project Consortium, 2012).
3. ClinVar: A database of genomic variation and its relationship with human health (build version 20180603) (Landrum et al., 2015).
4. ExAC: The Exome Aggregation Consortium dataset (version 0.3) of exome allele frequency data for over 60 000 individuals from various populations (Karczewski et al., 2016).

5. InterVar: A database of interpretations for missense variants (version 20180118) (Li and Wang, 2017).
6. dbNSFP: A database of non-synonymous human single nucleotide variants and their functional predictions (version 35c, 201810) (Liu et al., 2016).
7. gnomAD: The Genome Aggregation Database that contains aggregated exome and genome sequencing data (Karczewski et al., 2019).

In addition to these databases, the variants were compared to the variants originally identified in the previous projects using *Python* scripts.

3.2.5.5 Predicting the impact of the novel variants

After identifying potentially novel variants, research was conducted to investigate the probable impact that they would have on the gene, transcript or protein. Using the sequence viewer provided by NCBI (http://www.ncbi.nlm.nih.gov/projects/sviewer/), the 27 novel exonic variants were plotted and examined against previously annotated variants. The function of the genes the variants were found in was investigated and the functional host-pathogen PPIN was searched for any noteworthy protein associations. This in-depth investigation was restricted to the 27 novel exonic variants identified, as they are most likely to cause direct protein changes.

ANNOVAR was used to perform region-based annotation on the novel variants to determine whether or not they overlap with known variants or fall in regions of biological interest. The following region-based annotation was conducted on the novel variants:

1. The variants were compared against conserved genomic regions and annotated with regards to whether or not they fall within conserved genomic regions using the 100-way alignment from *phastCons* (Siepel et al., 2006).
2. Variants located within segmental duplications were annotated using *ANNOVAR's* database of segmental duplications. These variants may constitute sequencing alignment errors.
3. The novel variants were compared to previously published structural variants in the Database of Genomic Variants, which contains variants that involve segments of DNA that are larger than 50bp. Only the novel variants that involved changes longer than 20bp were compared for overlaps in the Database of Genomic Variants (MacDonald et al., 2013).
4. Lastly, the variants were compared to dbSNP based on the position of the variant (ignoring the nucleotide change).

The results of these investigations were summarised by the predicted variant type.

3.3 Results

In this chapter, a multi-coloured linked de Bruijn graph of chromosome 6 was assembled using sequences from 33 African individuals. During the graph assembly, reads that were previously unmapped in the reference-based assembly were incorporated into the graph. In addition, novel variants were identified from the graph assembly. In this section, we describe these results and other results leading up to, and post-construction of, the sequence graph.

3.3.1 Chromosome 6 sequence graph construction

A linked de Bruijn graph of chromosome 6 was constructed using the whole genome sequencing data from 33 African individuals. Each individual was assigned their own "colour" in the graph, enabling paths along the graph to be compared between individuals. The graph is represented across two files, one containing all the nodes in the graph, and one containing the links between them. The final graph contained 228 601 888 k-mers, as well as 22 191 474 paths between them. A log of the processing time, memory usage, as well as per sample k-mer and link output, is available as an Appendix in Table 7.5.

3.3.1.1 Incorporation of unmapped reads

To identify the proportion of unmapped reads that were incorporated into the graph, the unmapped reads were filtered based on whether or not they shared a k-mer with the graph. On average, 68.26% (σ=6.09%) of the unmapped reads per individual were incorporated into the individual's error-cleaned graph.

3.3.1.2 Sequence similarity between the individuals

A distance matrix illustrating the percentage of k-mers shared between any pair of samples is depicted in Figure 3.6. The percentage of k-mers shared between any two individuals ranged between 77.48% and 97.56%, with, on average, 93.58% (σ= 2.98%) of k-mers shared between any two individuals. In the related Yoruba trio, the parents shared 94.67% of k-mers with each other and the child shared over 96.57% of k-mers with each of their parents. However, the highest k-mer similarity (97.2%) was observed between a Mandenka individual (HGDP01284, SGDP) and a Yoruba individual (NA19239, 1000G). The lowest k-mer similarity (77.48%) was observed between a Gambian individual (HG02568, 1000G) and a Mandenka individual (HGDP01284, SGDP).

3.3.2 Identifying variation from the graph

Once the graph was built, it was used to identify variants between the individuals with respect to each other rather than against the existing reference genome. The variants were identified using the bubble caller built into *McCortex*. Thereafter, the variants were mapped to the current human reference genome (GRCh38). This enabled variant annotation with genomic coordinates so that the variants called from the graph could be compared to variants

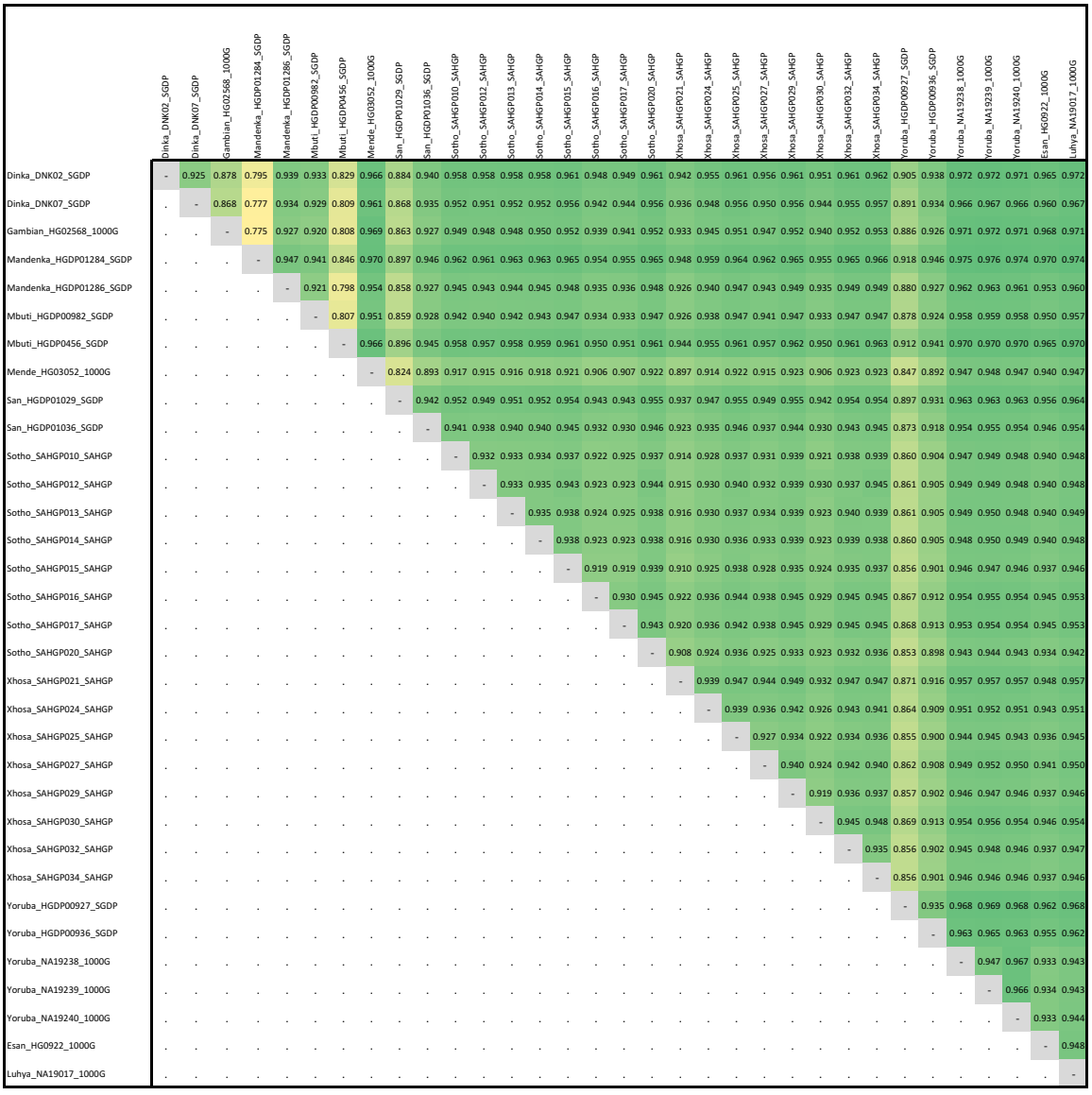


Figure 3.6 Distance matrix illustrating the proportion of k-mers shared between the 33 sequences
The individuals are labelled at the top and the left of the matrix, indicating the population, the individual's sequence identifier, and the sequencing project. The proportion of k-mers shared between the pair of individuals is written in each cell of the matrix. The cells are coloured on a scale ranging from yellow to green, whereby lighter yellow cells indicate lower proportions and darker green cells indicate higher proportions of k-mers shared respectively.

originally called from the reference genome for the sequences included in this study and from variation databases. Overall, 2 400 125 bubbles were identified in the graph, of which 2 354 405 (98.10%) could be mapped to the reference genome. When compared to the reference, 4 017 593 alternate alleles were identified within these bubbles. To annotate these alleles, the k-mer coverage per allele was calculated. However, this could only be calculated for 46.62% of the alleles, as the rest were too dense (over eight variants within one k-mer/51bp). Notably, variants were included regardless of the ability to annotate the alleles with k-mer coverage. 817 232 alleles were excluded due to their bubbles having a MAPQ score <30 or mapping to multiple positions on the chromosome. This resulted in a final number of 3 200 361 alternate alleles to the reference identified from the graph.

3.3.2.1 Annotating the variants

When filtered by chromosome location and nucleotide change, the bubbles represented 1 039 655 unique variants. Of these variants, 981 061 (94.36%) aligned to the reference chromosome 6, while the remainder aligned to the alternate reference locus sequences (alternate loci). The variants were predominantly SNPs (864 950, 83.20%).

In Table 3.1 the variants are described by location, with exonic variants being further described by function.

3.3.2.2 Comparing variants called from the graph to variants originally called against the reference

The variants called from the graph were annotated using *ANNOVAR* and compared to the variants that were originally called from the reference-based assembly. Only variants that aligned to the reference chromosome 6 could be compared, as the previously identified variants were not recorded as aligning to alternate loci. There were almost twice as many variants identified from the graph than there were called from the reference-based assembly. Even though more variants were identified from the graph, only 46.23% variants called from the reference-based assembly were re-identified in the graph. A detailed comparison of the number of variants identified, broken down by their gene location and sequence change, is available in Table 3.1.

Variants that could only be mapped to alternate loci for chromosome 6 are listed in the last column of Table 3.1. Only intergenic, upstream, and downstream variants that were identified as mapping to alternate loci could be cross-checked against the variants mapping to chromosome 6, because *ANNOVAR* annotates these variant types with distances relative to the gene, e.g. 500bp upstream of the gene. As such, it is possible that some of these variants listed in the "alternate loci" column are a subset of the variants that were mapped to chromosome 6.

Table 3.1 Comparison between variants called from the graph and variants called from the reference-based assembly

	Graph variants	Previous variants	Re-identified (%)	Novel	Alternate Loci
Total variants	981 061	466 334	46.23	44 864	58 594
SNPs (transitions)	552 975	288 726	46.65	7935	33 243
SNPs (transversions)	262 716	152 290	41.30	5529	16 016
Indels / substitutions	165 370	25 318	71.06	31 400	9335
Location/Function					
exonic	4368	3128	39.10	27	491
frameshift insertion	11	4	25.00	2	2
frameshift deletion	10	3	33.33	1	5
stopgain	23	25	28.00	0	1
stoploss	2	2	50.00	0	0
nonframeshift insertion	24	3	33.33	2	7
nonframeshift deletion	39	14	64.29	4	15
nonframeshift substitution	29	0	-	17	7
non-synonymous SNV	2176	1769	37.54	0	252
synonymous SNV	2049	1286	41.76	1	199
unknown	5	22	9.09	0	0
splicing	24	31	22.58	1	2
ncRNA	57 021	28 076	45.17	2281	6034
ncRNA-exonic	3100	1588	40.49	76	847
ncRNA-splicing	11	5	0.00	0	3
ncRNA-intronic	53 910	26 483	45.46	2205	5184
UTR5	1249	938	25.37	31	182
UTR3	7486	3614	49.75	268	406
UTR5/UTR3	2	1	0.00	0	0
intronic	335 818	162 842	47.60	13 488	12 437
upstream	4592	2879	31.96	146	840
downstream	5759	2983	39.56	344	772
upstream/downstream	162	120	25.83	0	42
intergenic	564 580	261 722	45.85	28 278	37 388

3.3.2.3 PCA of chromosome 6 variants

In order to ascertain the population structure and divergence of the individuals that would be included in the sequence graph, PCA was conducted on chromosome 6 variants prior to constructing the graph and after identifying variants from the graph. Both PCA plots excluded variants in the MHC region and excluded the child of the Yoruban trio. The first PCA used variants that were identified in all three projects in at least one of their samples, which left 32 793 SNPs from 32 individuals available for analysis. The second PCA used the variants identified between the individuals in the graph, for which many more variants (612 299 in total for the 32 individuals) could be included, as variants were called in the same way for all the projects.

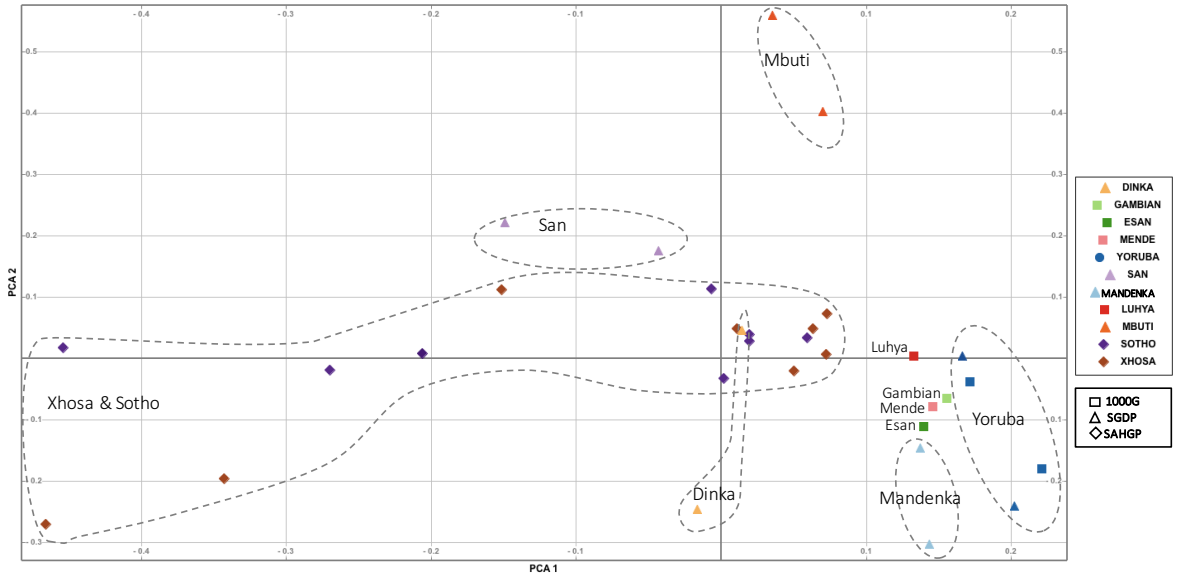
The annotated results of the two principal component analyses are displayed in Figure 3.7. In both plots, the individuals roughly fall into clusters by population, which loosely mirror the geographic location of the population (see Figure 3.8). In plot A, the first and second principal components explained 14.20% and 11.43% of the variance respectively. In plot B the first and second principal components explained 13.72% and 12.11% of the variance respectively. The variance explained by each of the 10 principal components is presented in Table 3.2. The highest divergence between individuals from the same population is observed in the Xhosa and Sotho populations, which are spread out along the first principal component, although this is less pronounced in the PCA of graph-identified variants. The Yoruba, Mandenka, and Dinka individuals display less variation along the second principal component in the PCA of graph-identified variants than in the PCA of variants identified by calling against the reference.

Table 3.2 Results of the PCA analysis

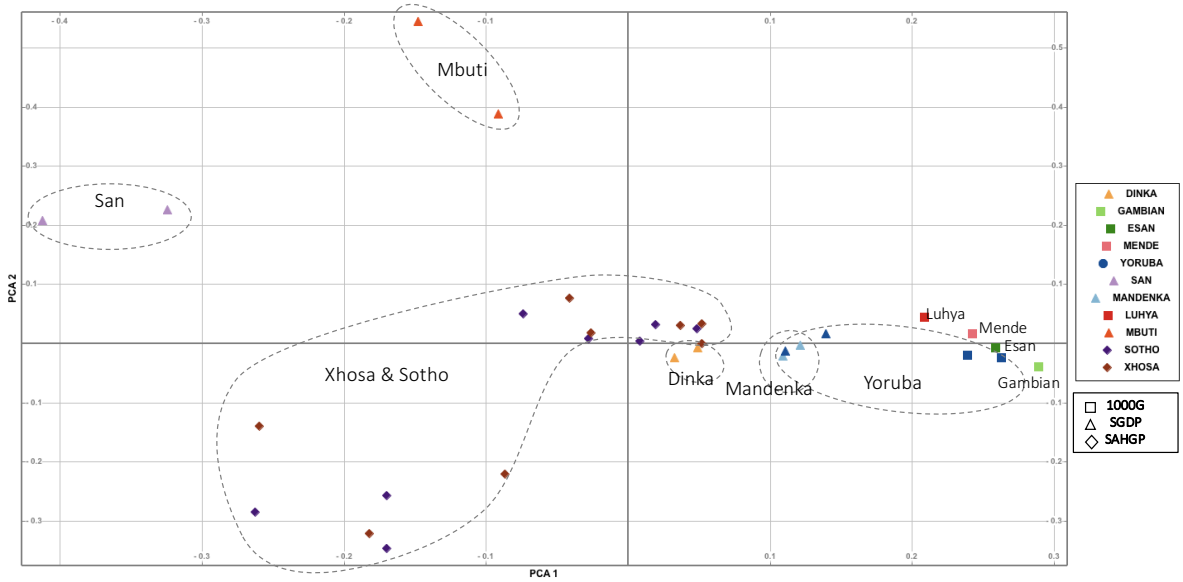
	A. Reference SNPs (*n*=32 238)		B. Graph SNPs (*n*=612 299)	
Principal component	Eigenvalue	Variance explained (%)	Eigenvalue	Variance explained (%)
1	1.909	14.20	1.681	13.72
2	1.536	11.43	1.484	12.11
3	1.464	10.89	1.351	11.03
4	1.375	10.23	1.236	10.09
5	1.301	9.68	1.125	9.18
6	1.263	9.40	1.120	9.14
7	1.212	9.02	1.108	9.04
8	1.154	8.58	1.071	8.74
9	1.135	8.44	1.043	8.51
10	1.094	8.14	1.032	8.42

3.3.2.4 Identifying novel variants

The variants identified from the graph were compared to publically available databases of genomic variation, including dbSNP, 1000 Genomes, ClinVar, ExAC, InterVar, gnomAD, and dbNSFP. Most of the variants identified between the individual sequences existed in at least one of these databases. Overall, 4.57% of the variants identified from the graph that could be



A. PCA of variants identified by the projects



B. PCA of variants identified through the graph

Figure 3.7 PCA of chromosome 6 variants genotyped in 32 African individuals.
This figure shows two biplots of PCA results analysing SNPs from Chromosome 6, with (A) analysing the variants identified by the contributing projects by calling against the reference, and (B) analysing the variants identified from the sequence graph. Each population is assigned a colour, and each project (SAHGP, SGDP, 1000G) is assigned a shape, so that each point on the plot can be identified with respect to the population it represents and the project that provided the sample. The outlined regions group together individuals from the same or co-located populations. SAHGP (n=16): Xhosa (n=8), Sotho (n=8); SGDP (n=10): San (n=2), Mbuti (n=2), Mandenka (n=2), Dinka (n=2), Yoruba (n=2); 1000G (n=7): Yoruba (n=3, child of trio excluded from plot), Luyha (n=1), Gambian (n=1), Mende (n=1), Esan (n=1)

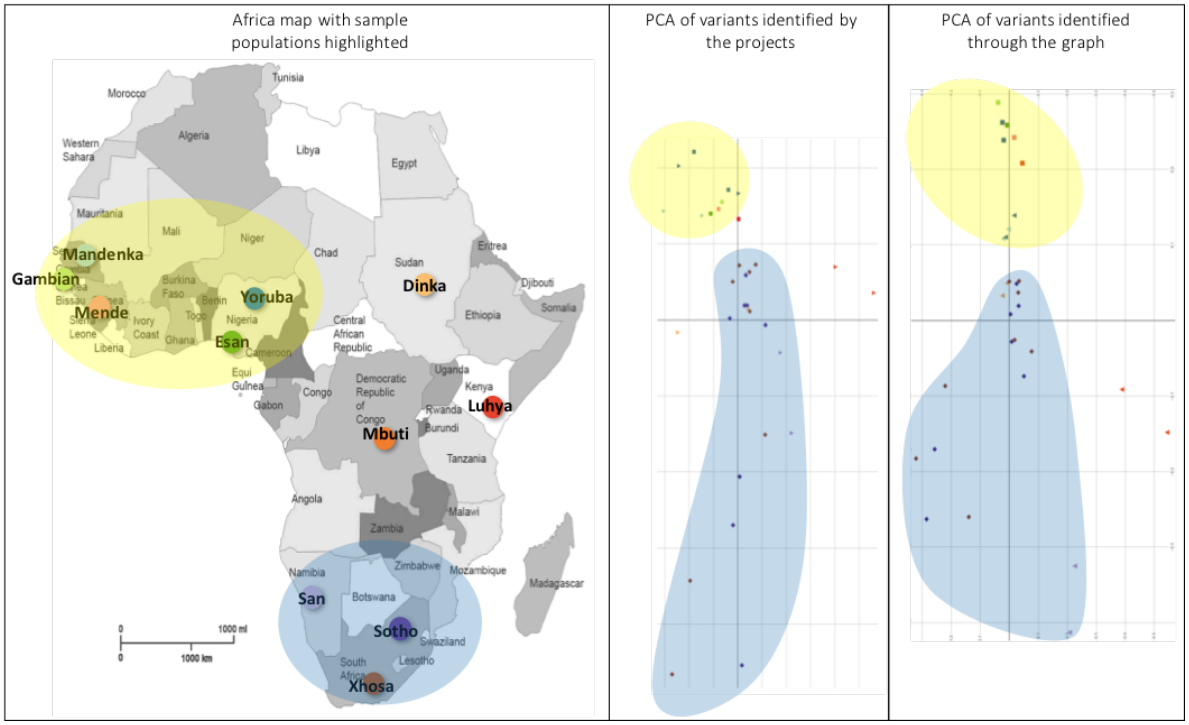


Figure 3.8 Comparison of the PCA distribution and geographical location.
The Africa map with individuals plotted alongside the PCA plot which has been rotated. The position of the points on the PCA loosely mirrors the geographic location.

mapped to reference chromosome 6 were not found in the databases (44 864 variants in total). A summary of the novel variant location is available in Table 3.3. The majority of the newly identified variants were indels and substitutions (69.99%). There were 27 new exonic variants identified, including 26 indels or substitutions and one synonymous SNP.

3.3.2.5 Identifying variants in the MHC region

The graph variants that aligned to the MHC region were compared to previously identified MHC variants for the sequences included in the graph. In the whole chromosome, the total number of variants identified from the graph was more than two times the number of previously identified variants. In contrast, in the MHC region 2.5 times more variants were identified from the graph; however, less than a quarter of the total MHC variants identified from the graph mapped to reference chromosome 6 (see Table 3.3). Of the previously identified variants, 7.51% were re-identified in the graph, and, in total, 10.89% of all variants that could be mapped to reference chromosome 6 were novel (1377 variants).

Most of the novel variants were identified in intergenic regions, with only two novel exonic variants identified, both of which were predicted to cause nonframeshift changes.

More than three times more variants mapped to alternate loci than to chromosome 6. These alternate loci are specifically for the HLA region, which explains why the vast majority (69.95%) of all the variants that mapped to alternate loci were located in the MHC region.

Table 3.3 Comparison between MHC variants called from the graph and MHC variants called from the reference-based assembly

	Graph variants	Previous variants	Re-identified (%)	Novel	Alternate Loci
Total variants	12 640	21 022	7.51	1377	40 992
SNPs (transitions)	7173	12 695	7.42	534	23 321
SNPs (transversions)	3453	7124	6.13	266	11 320
Indels / substitutions	2014	1203	16.63	577	6351
Location/Function					
exonic	162	736	3.94	2	468
frameshift insertion	3	1	0.00	0	2
frameshift deletion	2	1	0.00	0	4
stopgain	0	9	0.00	0	1
stoploss	0	0	-	0	0
nonframeshift insertion	2	1	0.00	0	7
nonframeshift deletion	4	4	0.00	1	15
nonframeshift substitution	2	0	-	1	7
non-synonymous SNV	94	457	4.16	0	239
synonymous SNV	51	240	3.33	0	178
unknown	4	22	9.09	0	0
splicing	1	11	0.00	1	1
ncRNA	1232	2830	6.78	38	187
ncRNA-exonic	253	401	7.23	7	54
ncRNA-splicing	0	3	0.00	0	0
ncRNA-intronic	979	2426	6.72	31	133
UTR5	24	169	2.96	0	176
UTR3	118	291	5.50	10	349
intronic	1024	3786	5.07	56	9006
upstream	264	521	6.72	11	525
downstream	340	566	6.36	17	537
upstream/downstream	40	45	2.22	0	0
intergenic	9435	12 068	8.89	1242	29 743

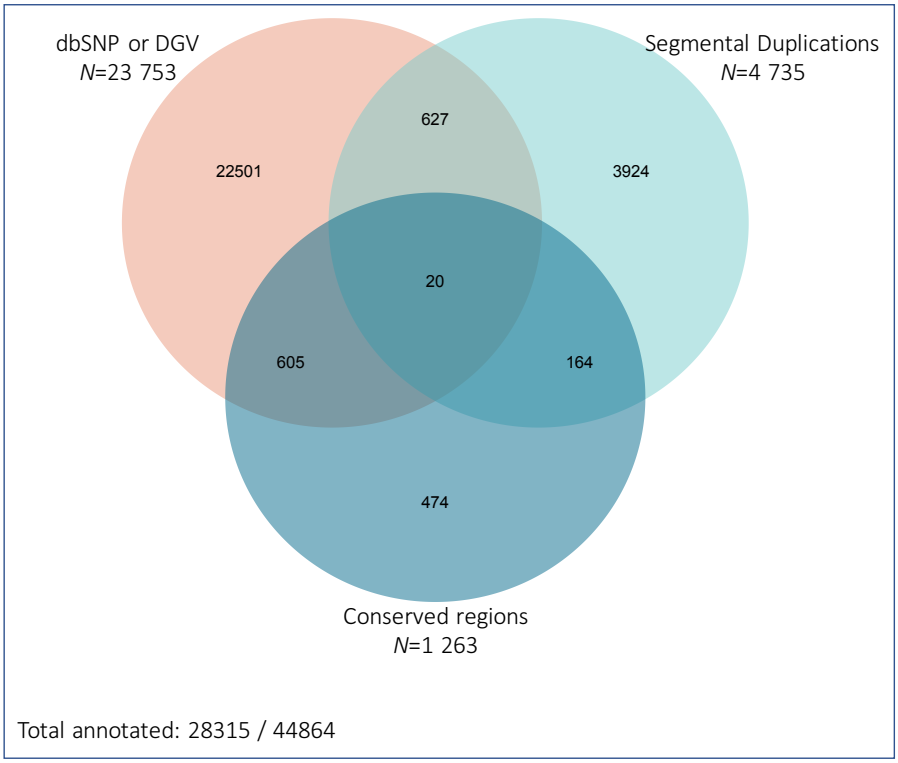


Figure 3.9 Venn diagram of how the novel variants were annotated
The Venn diagram shows the proportion of novel variants that were annotated based on either overlapping with a position in a database (dbSNP or DGV), overlapping with segmental duplications or overlapping with conserved regions.

3.3.3 Annotation of novel variants

Using *ANNOVAR*, the 44 864 novel variants were investigated further for overlaps with conserved regions, segmental duplications or positions of known SNPs, and structural variants. In total, 28 315 (63.11%) of the novel variants could be annotated, of which 23 753 (83.89%) overlapped with positions of known SNPs and structural variants, 4735 (16.72%) overlapped with segmental duplications, and 1263 (4.46%) were located in conserved regions (see Figure 3.9 for more detail). In other words, other variants had previously been identified in the same position, for example SNPs with a different allele, SNPs in a position where a substitution had been identified, or overlapping structural variants. The number of variants that could be annotated are broken down by variant type in Table 3.4.

Using the sequence viewer provided by NCBI, the 27 novel exonic variants were investigated further by reviewing the genes they were located in to identify the exact protein change and see if there were any other features of interest. The 27 variants were located in 19 different genes. Of the 27 exonic variants, 17 were 2bp nonframeshift substitutions and of those: ten variants were in the same position as two known adjacent SNPs, two variants overlapped with at least one known SNP, and one position had a SNP previously reported. Only five of the 27 "novel" exonic variants in total had no known or similar variants in the same position. The function of the genes, protein change of the variant, and any similar or variants are described for all 27 exonic variants in the paragraphs that follow.

1. ATXN1. This gene encodes the Spinocerebellar ataxia type 1 protein. Expansions of the glutamine repeat region elongate the polyglutamine tract and are known to be a cause of spinocerebellar ataxia type 1, an autosomal dominant neurodegenerative disorder. In our functional host-pathogen PPIN, ATXN1 functionally interacts with HIV protein Vif, whereby Vif upregulates the expression of ATXN1 in T-cells.
   Two nonframeshift insertions were identified in this gene:
   (1) chr6:16327684-16327684:->TGCTGATGC and
   (2) chr6:16327684-16327684:->TGCTGCTGC.
   The first variant inserts additional histidine and glutamine amino acids into the protein, while the second variant adds additional glutamine amino acids added to the polyglutamine tract of the protein. Similar insertions have been previously observed at this position (rs767434913, rs1554138052). In particular, rs1554138052 has been observed in African sequences at a higher frequency than other populations (0.002 vs global 0.0007 in ExAC).

2. BTN3A3.The butyrophilin (BTN) genes are located in the MHC class 1 region. BTN3A3 is involved T-cell mediated immunity, T-cell receptor signaling pathway, as well as regulation of the immune response. This gene is not in the functional host-pathogen PPIN.
   Two missense, nonframeshift substitutions were identified in this gene. Both of which are a combination of two adjacent known SNPs:
   (1) chr6:26451828-26451829:CA>TG: threonine (ACA) is replaced with methionine (ATG). The two adjacent known SNPs are rs755373159 C>T, and rs1226938598 A>G, the first of which has been observed in African samples at extremely low frequency 0.00001 (ExAC and gnomAD exomes).
   (2) chr6:26451762-26451763:GG>AT: arginine (CGG) is replaced with histidine (CAT). Two adjacent known SNPs are rs762891465 G>A, and rs1244731589 G>T, which have previously been observed at extremely low frequency 0.00001 (ExAC).

3. C6orf10. The testis expressed basic protein 1 is in the functional host-pathogen PPIN, but only functionally interacts with one other protein, the Amyloid protein-binding protein 2.
   The variant identified in this gene is a long nonframeshift deletion of 135 nucleotides.
   chr6:32293422-32293556:CTCACTCTTCTCTACCTGGGCTTCCTGTCCTTTCA
   GTACAACTGACTCCCTCTTCTTTACCTGGGCTTCCTGTCCCTTTGAGACACCAGA
   CTGACTCTTCTTTACTTGGGATTCCTGTCTTCTTGGCACACCCAT>-
   A similar deletion has been previously reported (rs1379518282), but no frequency information was available.

4. CDKAL1. The function of the CDK5 regulatory subunit associated protein 1-like-1 is not well characterised, but the protein encoded by this gene is a member of the methylthiotransferase family. CDKAL1 functionally interacts with human proteins in the functional host-pathogen PPIN. A nonframeshift substitution (chr6:21065218-21065219:CA>TG) was identified in this gene. The variant results in a missense mutation caused by replacing proline (CCA) with leucine (CTG).
   The substitution overlaps the position of two adjacent known SNPs: rs77152992 C>T

and rs56087852 A>G. Both SNPs have been observed in African populations (frequency of 0.06 and 0.2 for rs77152992 and rs56087852 respectively) in multiple studies (ExAC, gnomAD Genomes, and 1000G).

5. GPRC6A. G-protein coupled receptor family C group 6 member A is involved in G protein-coupled receptor activity and signaling receptor activity, as well as calcium mediated signaling, and functionally interacts with human proteins in the functional host-pathogen PPIN. A frameshift insertion (chr6:116792602-116792602:->TTCC) was identified in this gene. It is similar to a previously reported insertion rs111974433, which has been observed at higher frequencies in African populations in three studies gnomAD exomes (.247 vs .046), gnomAD genomes (.369 vs .154), and 1000G (.401 vs .159).

6. GSTA1. Glutathione S-transferase A1 is involved in several biological processes such as epithelial cell differentiation, and glutathione derivative biosynthetic process. It functionally interacts with human proteins in the functional host-pathogen PPIN. A nonframeshift substitution was identified in this gene (chr6:52794205-52794206:AT>TA) that causes two codon mutations, first a missense mutation where cysteine (TGT) is replaced by methionine (ATG), followed by a synonymous mutation to valine (GTA>GTT). One other SNP has been reported at this position (rs17414159).

7. GSTA2. Glutathione S-transferase A2 is involved in several biological processes such as epithelial cell differentiation, and glutathione derivative biosynthetic process. It functionally interacts with human proteins in the functional host-pathogen PPIN. A nonframeshift substitution (chr6:52752934-52752935:TA>AT) was identified in this gene, which causes two missense mutations replacing serine (AGT) and phenylanaline (TTT) with cysteine (TGT) and leucine (TTA). The first SNP (T>A) has been previously reported (rs1429852254) and observed at low frequency (0.001 by gnomAD).

8. GSTA5. Glutathione S-transferase A5 is involved in the glutathione metabolic process. It functionally interacts with human proteins in the functional host-pathogen PPIN. Five variants were identified in this gene:
(1) chr6:52834159-52834159:G>A is a synonymous mutation of a tyrosine amino acid (TAC > TAT). Another SNP has been reported at this position before (rs748204432).
(2) chr6:52834184-52834185:AC>TT leads to a missense mutation replacing valine (GTC) with asparagine (AAC). No other SNPs have been reported at this position.
(3) chr6:52834200-52834201:CA>TG leads to two missense mutations replacing aspartic acid (GAT) and alanine (GCC) with glutamic acid (GAA) and proline (CCC). The first SNP (C>T) rs759910242 is known, and has been previously observed at a low frequency 0.00001 (ExAC).
(4) chr6:52834247-52834248:GT>CC results in a missense mutation replacing threonine (ACT) with glycine (GGT). It is a combination of two adjacent known SNPs: rs1457106056 G>C, rs1178884319 T>C, which have been previously observed at extremely low frequency (gnomAD).
(5) chr6:52834193-52834194:GT>AG results in a missense mutation replacing threonine (ACT) with leucine (CTT).

9. KIAA0319. Dyslexia-associated protein KIAA0319 is involved in neuronal migration during development of the cerebral neocortex, and it functionally interacts with human proteins in the functional host-pathogen PPIN. A nonframeshift substitution (chr6:24578224-24578225:TC>GG) was identified, which results in a missense mutation replacing glutamic acid (GAA) with proline (CCA). It was identified in a Xhosa individual in our dataset.

10. MEP1A. Meprin A subunit alpha is involved in zinc ion binding and metalloendopeptidase activity, and it functionally interacts with human proteins in the functional host-pathogen PPIN. A nonframeshift substitution (chr6:46835364-46835365:CA>TG) was observed, which overlaps the position of two SNPs, first a synonymous SNP of alanine (GCC>GCT), followed by a missense SNP where methionine (ATG) is replaced by valine (GTG). Both of these SNPs are known (rs116465625 C>T, and rs2297019 A>G), and both have been shown to have higher frequency in African populations. rs116465625 has been observed at a frequency greater than 0.025 (ExAC, 1000G). rs2297019 has been observed at a frequency of at least 0.45 in African populations by three studies (ExAC, gnomAD exomes, gnomAD genomes, 1000G).

11. MICA. MHC class I polypeptide-related sequence A is involved in natural killer cell activation, immune response, T-cell mediated cytotoxicity, and defence response to virus and bacterium. In the functional host-pathogen PPIN it functionally interacts with HIV protein, Nef. A nonframeshift substitution (chr6:31415133-31415134:CA>TG) was identified, however, although *ANNOVAR* annotated this as exonic, on the genome viewer it is in the 3'UTR. It confers two changes, first a synonymous mutation followed by a missense one. It overlaps the position of two adjacent known SNPs: rs41545213 C>T and rs1882 A>G. Both have been observed in African populations in three studies. rs41545213 has frequency of 0.02-0.03 in both African and most non-African populations (gnomAD, ExAC, 1000G). rs1882 has frequency of 0.66-0.69 in African populations and 0.52-0.58 in European populations (gnomAD, ExAC, 1000G).

12. OR2A4. Olfactory receptor 2A4 is involved in G protein-coupled receptor activity. In the functional host-pathogen PPIN it is a high degree protein that functionally interacts with human proteins. A nonframeshift substitution (chr6:131701145-131701146:TG>AT) was identified that leads to a missense mutation replacing histidine (CAT) with isoleucine (ATT). The variant overlaps the position of two adjacent known SNPs: rs199675686 T>A, rs62423426 G>T. Both SNPs have been shown to have higher frequency in African populations than other populations. rs199675686 has a frequency of 0.2 (1000G). rs62423426 has a frequency of 0.2-0.33 (ExAC, gnomAD, 1000G).

13. PHACTR1. Phosphatase and actin regulator 1 is involved in actin cytoskeleton reorganisation, cell motility, and functionally interacts with human proteins in the network. A nonframeshift substitution (chr6:13014006-13014007:GC>AA) results in a missense mutation replacing arginine (CGA) with leucine (TTA). The variant overlaps the position of two adjacent known SNPs: rs962079940 G>A, rs994802085 C>A. Both SNPs have been shown to have higher frequency in African populations, >0.003 (gnomAD). It was observed in a Xhosa individual in our dataset.

14. PRIM2. DNA primase large subunit is the polymerase that synthesises small RNA primers for the Okazaki fragments made during discontinuous DNA replication and functionally interacts with human proteins in the network. The variant identified is a frameshift insertion (chr6:57537494-57537494:->A), in which the initial codon stays the same (glutamic acid), but the frameshift insertion would culminate in major changes to the protein thereafter. No similar variants have been previously reported.

15. PSMG4. Proteasome assembly chaperone 4 is a chaperone protein which promotes assembly of the 20S proteasome and functionally interacts with human proteins in the functional host-pathogen PPIN. A nonframeshift substitution (chr6:3264292-3264293:CT>TC) was identified that resulted in a missense mutation replacing leucine (CTT) with serine (TCT). The variant overlaps the position of two adjacent known SNPs: rs4959788 C>T, rs4959789 T>C, which have both been observed in African populations. rs4959788 has a frequency of 0.1-0.2 and rs4959789 has a frequency of 0.3-0.36 (ExAC, gnomAD, 1000G).

16. RAB44. Ras-related protein Rab-44 is involved in GTP binding, GTPase activity, calcium ion binding, and functionally interacts with human proteins in the functional host-pathogen PPIN. A nonframeshift substitution (chr6:36715503-36715504:GC>AA) was identified that resulted in a missense mutation replacing serine (AGC) with lysine (AAA). The variant is made up of two adjacent known SNPs: rs903821952 G>A, rs1000743433 C>A, which have previously been observed at extremely low frequency (gnomAD). It was observed in a Xhosa individual in our dataset.

17. TBP. TATA box-binding protein-like protein 1 is part of a transcription system that mediates the transcription of most ribosomal proteins. Deletions in the repeat instability region (expansion of the (CAG/CAA)n trinucleotide repeat is associated with spinocerebellar ataxia type 17) and functionally interacts with human proteins in the functional host-pathogen PPIN. Three variants were observed in TBP which were deletions of different numbers of glutamine (CAG) repeats (chr6:170561959-170561970:$(CAG)_4$/$(CAG)_5$/$(CAG)_6$>-). A similar deletion has been previously observed at extremely low frequency (gnomAD).

18. TDRD6. Tudor domain-containing protein 6 is involved in spermiogenesis, chromatoid body formation, and proper precursor and mature miRNA expression, and functionally interacts with human proteins in the functional host-pathogen PPIN. A frameshift deletion was identified in this gene (chr6:46693320-46693327:AATTAAGT>-). This was observed in a Sotho individual in our dataset. A similar indel (rs1562058168) has been previously reported.

19. TMEM14C. Transmembrane protein 14C is required for normal heme biosynthesis. A nonframeshift substitution (chr6:10728662-10728663:CA>TG) was identified, which results in two mutations, first a synonymous mutation of glycine (GGC>GGT) followed by a missense mutation replacing isoleucine (ATT) with valine (GTT). No variants have previously been observed in this region.

Table 3.4 Annotation of the novel variants

Variant type	Total "novel"	Any annotation	dbSNP	DGV	Conserved region	Segmental duplication
exonic	27	26 (96.3%)	20 (74.07%)	1 (3.7%)	12 (44.44%)	9 (33.33%)
splicing	1	1 (100%)	0 (0%)	0 (0%)	1 (100%)	1 (100%)
ncRNA exonic	76	42 (55.26%)	30 (39.47%)	5 (6.58%)	11 (14.47%)	10 (13.16%)
ncRNA intronic	2205	1476 (66.94%)	1237 (56.1%)	184 (8.34%)	50 (2.27%)	189 (8.57%)
UTR5	31	22 (70.97%)	16 (51.61%)	2 (6.45%)	4 (12.9%)	3 (9.68%)
UTR3	268	184 (68.66%)	127 (47.39%)	5 (1.87%)	35 (13.06%)	38 (14.18%)
intronic	13 488	8825 (65.43%)	7339 (54.41%)	1154 (8.56%)	370 (2.74%)	1103 (8.18%)
upstream	146	103 (70.55%)	86 (58.9%)	10 (6.85%)	10 (6.85%)	16 (10.96%)
downstream	344	226 (65.7%)	165 (47.97%)	19 (5.52%)	21 (6.1%)	60 (17.44%)
intergenic	28 278	17 410 (61.57%)	13 669 (48.34%)	2453 (8.67%)	749 (2.65%)	3306 (11.69%)
Total	44 864	28 315 (63.11%)	22 689 (50.57%)	3833 (8.54%)	1263 (2.82%)	4735 (10.55%)

3.4 Discussion

In this chapter, a sequence graph of human chromosome 6 was constructed using whole genome sequences of 33 individuals of African ancestry, without the use of the linear reference genome. From the graph, we were able to identify genetic variation between the individuals without bias to the linear reference genome. To our knowledge, this is one of the first attempts at constructing a sequence graph by *de novo* assembly of multiple high coverage human genome sequences that contains sequences of Southern African individuals. In this section, the caveats, usefulness, and potential replicability of the approach will be discussed in detail.

3.4.1 Variant identification from a *de novo* assembled graph

The graph presented in this thesis is the first known application of *McCortex* at this scale to human sequencing data. The method was chosen at the time this work was started, because the pre-cursor method to *McCortex*, *CortexVar*, had been shown to work well for human sequences and had been used to generate a population graph of the HLA region (Dilthey et al., 2015). We specifically chose this method as it could be used to generate a graph from which variation could be called between samples without the need for a reference genome, thus eliminating reference bias. Due to the ancestral similarity of the sequences included and their diversity from the reference, we hypothesised that (1) we would identify novel variation from the graph, and (2) that there would be fewer variants called from the graph compared to variants previously called against the reference.

After completing most of the analysis, simulations of the *McCortex* bubble caller were published and showed that the *McCortex* bubble calls may have a high false positive rate (1.93%) (Turner, 2019). To account for the suspected high false positive rate, we applied strict MAPQ thresholds to the variants that we identified prior to further analysis. Despite the reportedly high false positive rate, the proportion of novel variants in relation to the number of variants identified, as well as the proportion of novel variants according to genomic location, is reasonable and in line with expectations (O'Rawe et al., 2013).

Turner (2019) also showed that joint bubble calling in multiple samples using *McCortex* may reduce the amount of sequence available to map to the reference, thus reducing the power to identify the position of variants. However, only 1.90% of the bubbles called in the graph we constructed could not be mapped to the reference genome. Given that our graph also contained unmapped reads, it is expected that some bubbles would not be mapped to the reference genome. In addition, more than 95% of the variants identified from the graph were previously reported in one of the databases of known genetic variation. Furthermore, more than half (52.95%) of the variants that were not previously reported fell in the same position as other known variants. Thus, it seems that the power to identify the position of the variants was not drastically reduced by the joint bubble calling in *McCortex*.

While *McCortex* was able to identify the position of bubbles that were called, many of the variants that were previously called against the reference could not be re-identified using the graph-based approach, with only 46.23% of the previously identified variants being

re-identified. While this may have been affected by the joint bubble calling, a factor that may have further restricted the ability to re-identify the variants that were called by the original projects (SAHGP, SGDP, 1000G) is that the method of reference-based variant calling may have differed between the three projects. To mitigate this, we could have called the variants using the same method for all samples, rather than assuming that the approaches would have yielded the same results. Research has shown that often the concordance between standard reference-based variant calling methods is low. O'Rawe et al. (2013) showed that this can be as low as 60% for SNVs and even lower for indels. It is thus not surprising that there was a low intersection between the graph-based variant calls and reference-based variant calls. Had we done the reference-based variant calling ourselves, we could have also included the alternate loci. We found variants mapping to alternate loci, which could not be compared to what was previously called, especially for variants in the MHC region, of which 76% mapped to alternate loci. It is likely that some of the alternate sequences have an African source, explaining why a high percentage of variants mapped to the alternate loci, as well as the highlighting the usefulness of including the alternate loci sequences when assembling new genomes or calling variants (Sherman et al., 2019).

We expected less variants to be called from the graph than were called against the reference due to the ancestral similarity of the samples and their diversity from the reference. Instead, the graph-based method identified more than twice as many variants as the reference-based method. The largest difference was in the number of indels and substitutions identified, where the graph-based method identified 6.5 times more indels and substitutions than the reference-based method. Because most of the variants identified could be identified in one of the databases of genetic variation, it is possible that some of the increase could be explained by the differences in the original method of reference-based variant calling. Had we called the variants against the reference using the same methods for all samples, the differences may have been less pronounced. Another possible explanation is that inter-individual variation between the sequences included may be almost as high as their diversity from the reference, which is plausible given the huge diversity and geographical spread of the individuals.

3.4.2 The feasibility of reference-free sequence graphs

A recently published genome graph contains the linear reference as the spine on which structural variation and small-scale variation branch off in a graph structure (Rakocevic et al., 2019). This use of the reference as the spine enables the coordinate system of the linear reference genome to be used, enabling compatibility with all existing genomic methods. It also allows new sequences to be assembled by alignment to the graph, enabling simultaneous alignment to paths made up of variation and paths from the original linear reference. It already contains a lot of known variation, and as more known variation is added to it, it will allow more accurate assembly of more diverse sequences. However, this approach is is still biased to known variation, which is called with comparison to the reference genome, which itself is biased to specific populations. The pan-genome approach by Sherman et al. (2019) may more accurately eliminate this reference bias. Using 910 sequences of African descent, Sherman et al. (2019) took the sequencing reads that could not align to the linear reference and assembled them into contigs that represented approximately 10% additional

sequence to the linear reference. This observation is not unique to sequences of African descent, as Korean and Chinese genomes have also been shown to have additional sequence, some of which overlaps with these newly assembled contigs from African sequences (Sherman et al., 2019). This method of assembling contigs for sequences that diverge from the reference is akin to the additional HLA loci provided by the Genome Reference Consortium, and is likely the way forward in accurately representing all populations within the reference genome. Incorporating these contigs into a graph genome structure as proposed by Rakocevic et al. (2019) may give the best outcome.

Here we present a *de novo* assembled graph of a human chromosome containing sequences from 33 individuals from 11 African populations. The methods included unmapped reads as well as steps for removing contamination, similar to Sherman et al. (2019). A vast proportion of the decontaminated unmapped reads were incorporated into the graph (68%), showing that despite unmapped reads often being discarded, they may contain biologically important information. This sentiment is echoed by Sherman et al. (2019), who identified an additional 296.5Mb of novel DNA distributed across 125 715 sequences based on reads that were previously unmapped. In our case, the 68% of unmapped reads were filtered so that they included k-mers that aligned to chromosome 6 reads, however the overall percentage might still be an overestimate. It is possible that the unmapped reads map to multiple locations on the genome and were unmapped because of that. More work should be done to determine the locations of the unmapped reads that were incorporated into the graph. For example, instead of colouring the sequences by individual we could have used additional colours for unmapped reads so that they could be more easily distinguished.

While the *de novo* graph-based assembly approach we used may be useful for identifying variation between individuals, the extent to which it could act as a spine for assembly of new sequences has not been validated. Furthermore, the *de novo* graph-based assembly approaches have not yet evolved to share a coordinate system with the linear reference, which makes downstream analyses biased to regions that can be positioned on the linear reference and discards existing tools and methods for downstream analyses.

3.4.3 Recommendations for future reference graphs

While a reference-free *de novo* assembled sequence graph is a noble idea, it will likely be more sustainable to use a combination of the approaches discussed towards developing a comprehensive reference graph that represents diverse populations. In light of the recently published research and the findings presented in this thesis, we propose the following as components of a comprehensive reference graph:

1. Use the linear reference as the "spine" upon which to add additional paths made up of variation (Rakocevic et al., 2019). This will allow minimal deviations from the existing coordinate system, thus enabling downstream analyses to be largely unaffected by the new graph structure.
2. Assemble contigs from reads that were previously unmapped when assembling against the linear reference (Sherman et al., 2019). While a proportion of unmapped reads will

always be contamination, the remainder may be valid sequence of consequence to the individuals the reads come from. These contigs may be constructed from sequences from specific populations in a "pan-genome" approach. Incorporating these contigs as alternate paths on the reference will add to the diversity represented in the graph.

3. Incorporate known variation that has been called against the existing linear reference (Rakocevic et al., 2019). Millions of single nucleotide variants and structural variants have already been identified by comparing thousands of genomes to the existing linear reference. Incorporating all of this into the graph will enable the graph to capture a large proportion of the diversity between individuals.

4. Incorporate variation as identified between individuals through *de novo*-based methods as presented in this thesis. While known variation is likely to capture most of the diversity between individuals, the variation has been called against a reference genome that is biased to specific populations. Incorporating additional variation, such as the variation we identified using *McCortex*, will help to eliminate some of the bias. Additionally, individual genomes from the same populations could be collapsed into population-based graphs.

Using a combination of approaches will reduce the reference bias to specific populations as more variation for more diverse populations is added to the graph over time. Subsets of the final comprehensive graph may represent population specific reference genomes if needed, but the overall graph would be useful for almost unbiased assembly of all new sequences, including those of admixed ancestry.

3.5 Conclusion

This chapter aimed to identify variation in chromosome 6 using a reference-free graph-based method of variant identification between individuals, particularly for African sequences that are known to differ from the existing linear reference. More variants were identified within the graph than expected, and not all of the previously identified variants could be re-identified. Both differences are most likely explained by the discordance in the variant calls by different methods. Most of the previously unmapped reads were incorporated into the graph, highlighting that they carry biological information that should not be blindly disregarded due to not being in the reference. Finally, this chapter presents a pipeline for how a sequence graph could be constructed for multiple sequences in a reference-free manner. The usefulness of this graph structure is questionable given the vast deviation from the linear reference coordinate system and the downstream analyses that depend on it. However, we propose that variants identified through this approach could be a useful addition to reference graphs that use the linear reference as a spine, by introducing paths constructed from variants free of reference bias. We identified 44 864 novel variants, 27 of which were in protein-coding regions. Notably, at least 2 of the exonic variants were in proteins known to interact with HIV-1 proteins. In chapter 4, the variants identified in this chapter will be combined with all known variation and mapped onto the proteins in the functional host-pathogen PPIN constructed in

chapter 2 to investigate the potential impact that host genetic variation may have on HIV and TB co-infection.

4. Mapping human genetic variation and gene expression onto a host-pathogen protein interaction network

4.1 Introduction

Combining different types of high throughput biological data as well as experimentally verified observations when analysing a biological system may help to create a holistic view of the system. Multi-omic data integration may help to identify areas of interest that may have been overlooked by using one set of data, whilst additionally helping to triangulate findings across different data sets. Networks are a powerful tool to integrate different kinds of biological information, as they provide effective means to store, retrieve, and visualise complex data and relationships (Deffur et al., 2018; Pratt et al., 2015). In chapter 2, we created a functional human-pathogen PPIN between human, HIV-1 and *Mtb* proteins. Using this network, we were able to identify genes that functionally interact with both pathogens. In addition, we made use of network centrality measures to describe the possible importance of MHC proteins in facilitating interactions between the co-infecting pathogens and their host. In chapter 3, we used a graph-based assembly approach to identify additional variants in the MHC region that, due to reference mapping bias, may otherwise go undetected using traditional variant calling methods. It is well established that integrating gene expression data with PPINs is useful for validating the importance of proteins based on their network centrality, as well as for prioritising genes under certain conditions in which the gene expression was measured (Ma et al., 2007; Navlakha and Kingsford, 2010; Wu et al., 2012a). Furthermore, it has been shown to be beneficial to analyse variants with gene expression data, as highly differentially expressed genes have been shown to be more likely to have variants associated with disease (Chen et al., 2008). Using the functional human-pathogen PPIN generated in chapter 2 as the backbone, the aim of this chapter was to integrate gene expression data from HIV-TB co-infection with genetic variation data into the network, and to describe the possible impact of genetic variation on the interactions. In this section, we will introduce the utility of graph databases for integrating this information, methods for prioritising disease-associated genes using gene expression data, as well as the utility of using PPIs to predict the impact of genetic variants.

4.1.1 Graph databases for biological data integration

Graph databases, such as the freely available Neo4j, have been shown to be an efficient way to store and query biological network information (Deffur et al., 2018; Himmelstein et al., 2017). Deffur et al. (2018), created a complex association network called ANIMA that integrates multiple different types of transcriptomics datasets and clinical observations using

a Neo4j graph database. Using this network, they were able to reconstruct multiple features of disease states (Deffur et al., 2018). Himmelstein et al. (2017) used Neo4j to compile an integrative network of publicly available biological data from millions of studies called Hetionet, which can be mined for drug repurposing. Their network connects compounds, diseases, pathways, gene ontology annotations, and drug information, such as, side effects and symptoms (Himmelstein et al., 2017). In graph databases data points are stored as nodes, relationships are stored as edges, and properties can be assigned to both nodes and edges. The properties can be used to seed important nodes in the network in order to prioritise other nodes that may be important.

4.1.2 Prioritising disease-associated genes by integrating gene expression and protein interaction data

Analysing the relationship between gene expression during diseased states and PPIN centrality measures could assist to prioritise genes involved in co-infection. Annotation of proteins with differential expression during diseased states has been widely used to prioritise genes using network-based approaches (Guney and Oliva, 2012; Ma et al., 2007; Petrochilos et al., 2013; Wu et al., 2012a). Network-based approaches typically measure proximity to known disease genes, which act as seeds to prioritise candidate disease genes (Guney and Oliva, 2012). Several topology-based ranking algorithms have been proposed. It has been shown that while random walk-based methods typically perform better than other approaches, a combination of approaches typically yields the most complete results (Guney and Oliva, 2012; Navlakha and Kingsford, 2010). These approaches have been shown to work best with high quality interaction networks filtered on high confidence interactions, with edges weighted with the confidence of the interaction (Guney and Oliva, 2012), and nodes weighted with information that strengthens their association with a disease state, such as fold change in gene expression (Petrochilos et al., 2013).

Guney and Oliva (2012) present a network-based gene prioritisation framework called *GUILD* (Genes Underlying Inheritance Linked Disorders), in which they implement four widely used algorithms, namely: PageRank with priors, Functional Flow, Random walk with restart, and Network propagation; as well as four of their own algorithms, namely: NetShort, NetZcore, NetScore and NetCombo. They showed that their algorithms improved upon the existing algorithms. In particular, NetCombo, which combines the results of their other three algorithms, was the best performing method. The methods available in *GUILD* were used to rank genes in the PPIN in order to identify genes that may be associated with HIV-TB co-infection.

Gene expression data from a case-control study on a cohort of HIV-TB co-infected and latent TB infected African adults from Malawi and South Africa (Kaforou et al., 2013) were used to identify significantly differentially expressed genes that could be used as seeds for prioritisation of other relevant proteins in the network. In addition, this gene expression data were analysed alongside the protein interactions data to determine whether differentially expressed genes map to proteins with high network centrality measures.

4.1.3 Analysing the impact of variants by integrating genetic variation, gene expression, and protein interaction data

It has been found that genes that are highly differentially expressed are more likely to have disease-associated variants (Chen et al., 2008). Furthermore, using differential expression during a disease state to prioritise important genes will help with the prioritisation of variants (Guo et al., 2015; Laddach et al., 2018). Mapping variants onto PPINs can provide insight into the complex interplay of many variants on protein assemblies and may improve understanding of the consequences of variants on the PPIN (Laddach et al., 2018). In this chapter, known variants (with and without clinical significance) as well as variants identified in chapter 3 were mapped onto the network to identify variants that may be important. We will investigate the types of variants within prioritised genes, identify whether some have already been linked with HIV or TB or co-infection, identify others that may be novel, and analyse what impact they may have.

4.1.4 Aim and hypotheses

The aim of this chapter was to annotate the human-pathogen functional PPIN with genetic variation and gene expression data so that the relationships between the different data types could be interrogated. To this end, the different datasets were combined and a graph database containing the protein interactions, variants, and gene expression data was created. Using this integrated dataset, we investigated whether or not genes that were differentially expressed in HIV-TB co-infected individuals compared to uninfected individuals exhibit network properties that suggest they facilitate interactions between the host and pathogens. In addition, we investigate whether or not proteins with high network properties have variants that may be clinically significant. We hypothesise that (1) genes that are differentially expressed in HIV-TB co-infected individuals compared to uninfected individuals have high network importance in the functional host-pathogen PPIN; (2) variants with known clinical consequences for the progression of the two diseases are found in human proteins with high network importance; and (3) human proteins that functionally interact with both HIV-1 and *Mtb* proteins have additional variants that change gene structure, expression or function, which have not yet been shown to have clinical consequences.

4.2 Methods

4.2.1 Analysis of gene expression data

One of the aims of this study was to identify whether genes that are differentially expressed during HIV-TB co-infection have network properties that indicate importance in the network. To this end, the microarray data from the study conducted by Kaforou et al. (2013) was accessed through NCBI's Gene Expression Omnibus (accession number GSE37250). Kaforou et al. (2013) recruited patients with the following conditions: HIV-infected patients with recently diagnosed active TB (co-infected individuals), HIV-infected patients with latent TB infection (HIV-infected individuals), HIV-uninfected patients with recently diagnosed active TB (TB-infected individuals), HIV-uninfected patients with latent TB infection (controls), along with

HIV-infected and uninfected patients with other diseases. Whole blood was collected before or within 24 hours of TB treatment initiation in presumed TB cases, and microarray analysis was performed on HumanHT-12 v.4 expression Beadarrays (Illumina) (Kaforou et al., 2013).

Using the Geo2R functionality within Gene Expression Omnibus, gene expression was compared between LTB individuals (controls, *n*=83) and HIV-TB co-infected individuals (*n*=98), of which the demographics can be viewed in Table 4.1.

Table 4.1 Demographics of individuals enrolled in the gene expression analysis by Kaforou et al. (2013)

	HIV-TB co-infected		LTB	
Country	SA	Malawi	SA	Malawi
N	49	60	50	36
Age in years median (IQR)	33.7 (29.0-38.3)	34.5 (29.6-43.2)	20.6 (19.1-23.4)	38.9 (32.3-50.9)
Sex (male, %)	40	52	42	53

The Geo2R package was run using Benjamini Hochberg multiple testing correction. Any record with a corrected *p*-value <0.001 was considered to be significantly differentially expressed. There were 28 953 probes in the dataset. The proteins corresponding with these probes were identified by mapping the Illumina Probe IDs to UniProt identifiers, this resulted in 18 633 proteins. Of these, 15 122 (81.16%) were in the high confidence human-pathogen PPIN. Some proteins (3063) had multiple probes. In these cases, the differential expression reading with the lowest *p*-value was used. Of the 15 122 proteins, 3486 (23.05%) were differentially expressed between LTB and HIV-TB co-infected individuals. To test whether differentially expressed genes show higher importance in the network, Wilcoxon rank-sum tests were performed comparing the following network properties between HIV-TB co-infected and LTB individuals: Pathogen Bridging Centrality, Degree, Betweenness, Closeness, as well as the Minimum distance to MHC proteins, *Mtb* proteins, and HIV proteins.

4.2.2 Prioritisation of disease-associated genes by integrating gene expression and protein interaction data

The significantly differentially expressed genes were used to prioritise other genes potentially associated with HIV-TB co-infection by using a network-based gene prioritisation framework called *GUILD* (Guney and Oliva, 2012). The framework implements eight algorithms implemented in either C++, R or *Python*.

4.2.2.1 Scoring the nodes and edges

The algorithms require a node file that contains an identifer for the node and a phenotypic relevance score. The phenotypic relevance score was set to 0.001 for all genes that were not differentially expressed during HIV-TB co-infection, and for all pathogen proteins. For genes that were significantly differentially expressed, the score was set to the absolute value of the log fold change in expression as calculated by Geo2R. Petrochilos et al. (2013) showed that

fold change was a more reliable scoring method for seed genes than *p*-values. The scores ranged from 0.1 to 5.

The algorithms also required an edge file that contains two node identifiers as well as an interaction score. The STRING score was used for all human-human interactions, and *Mtb*-*Mtb* interactions. Host-pathogen interactions were scored according to the weighting described in subsections 2.2.4 and 2.2.3 for HIV-1-human PPIs and *Mtb*-human PPIs respectively.

4.2.2.2 Network-based algorithms for prioritisation

In this section, the algorithms that were used to prioritise genes in the PPIN based on gene expression is described. We chose to implement all four of the algorithms defined by Guney and Oliva (2012), specifically focusing on the results of their combined algorithm, and three of the other algorithms implemented in their package, *GUILD*.

4.2.2.2.1 NetShort NetShort is one of the algorithms created by Guney and Oliva (2012) as part of *GUILD*. The algorithm is based on the notion that in a given network a node that is important for a phenotype (seed node) would have shorter distances to other seed nodes. Instead of only using shortest path distance, the algorithm accounts for the number of edges that reach a seed node and the number of seed nodes that are in the path. The edge weight is adjusted so that edges connecting seed nodes are shorter. The algorithm defines the score of the node p as:

$$score(p) = \sum_{n \epsilon N, n \neq p} \frac{1}{d(p,n)}$$

where $d(p,n)$ is the length of the shortest path between nodes p and n with weighted edges of graph $G(N,E,f)$, defined by nodes N, edges E, and the edge weight mapping function, f. The edge weight mapping function f is defined as: $f : (a,b) \rightarrow \Re, \forall (a,b) \epsilon E$. The weight $f(a,b)$ is derived by multiplying the edge score and average of the initial scores of both nodes as per the following equation:

$$f(a,b) = score(a,b) * (\frac{score^0(a) + score^0(b)}{2})^{-1}$$

By this definition, the edge between two nodes will be short when the nodes forming the edge have high scores (seed nodes), otherwise the edge between the nodes will be long.

4.2.2.2.2 NetZcore NetZcore is another algorithm proposed by Guney and Oliva (2012). The algorithm assesses the relevance of a node for a given phenotype. The relevance of the node is highlighted by comparing it to the background distribution of the relevance of neighbouring nodes through the use of random networks. The score of a node is defined as

the average of the scores of its neighbouring nodes and is normalised using the z-score formula:

$$score_{k+1}(p) = \frac{score_k(p) - \mu_k^{random}}{\sigma_k^{random}}$$

where μ_k^{random} and σ_k^{random} are the mean and standard deviation respectively of the distribution of scores belonging to a set of random networks that have the same topology as the original graph. To generate networks with the same topology, a node p is swapped with another node q with the same degree. *GUILD* uses 100 random networks. Calculating the node scores based on the neighbour scores, is repeated a specified number of iterations k, varying from 1 to a parameter that can be specified as the maximum. The score of a node p at an iteration of k is calculated as:

$$score_k(p) = \frac{1}{\|Nb(p)\|} \sum_{q \epsilon Nb(p)} f(p,q) * score_k(q)$$

where in a graph $G(N, E, f)$ with nodes N, edges E,the set of neighbours of the node p is $Nb(p)$ and the weight of any edge between two nodes p and q is an edge weight mapping function $f(p, q)$.

4.2.2.2.3 NetScore NetScore is based on the propagation of information through nodes in a network (Guney and Oliva, 2012). In particular, the algorithm considers multiple shortest paths from the source node to the target node and ignores all other paths between the nodes. NetScore uses a message-passing scheme to calculate the information passed between nodes. In this scheme, each node sends its information as a message to its neighbours and then iteratively to their neighbours. Each message contains a node identifier and the weight of the path between the source and target node, where the weight is defined as the multiplication of the weights of all the edges along the path that the message has traveled. Only the messages arriving through all the shortest paths from a given node are considered. Each message is scored as the score of the node that sent the message plus the path weight. At each iteration, a node score is calculated as the average score for the messages received by that node. By iteration k, a node therefore has the score of the nodes reaching it from shortest paths of length k weighted by the edge weights in these paths. This algorithm was run with 3 repetitions and 2 iterations as these are the optimised values (Guney and Oliva, 2012).

4.2.2.2.4 NetCombo The fourth and final algorithm proposed by Guney and Oliva (2012), NetCombo, is a combination of the other three algorithms, NetScore, NetShort, and NetZcore. In this algorithm, the node score is the average of the normalised scores for each of the other three methods. In their analysis of prioritisation methods, Guney and Oliva (2012) showed that NetCombo produced significantly better predictions than Network Propagation,

the best of the other approaches. As such, instead of using all four methods proposed by Guney and Oliva (2012), NetCombo were used for this analysis, along with three of the other algorithms implemented as part of *GUILD*.

4.2.2.2.5 PageRank with priors

PageRank with priors uses a random walk-based model to score nodes based on phenotypic relevance, in which a random surfer has a higher chance of arriving at initially relevant nodes. Chen et al. (2009) proposed the use of this method for candidate gene prioritisation. Interactions in the network are treated as bidirectional edges and node association scores are allocated using the formula:

$$PR_{t+1}(u) = (1-d) * PR_0(u) + d * \sum_{(u,v)\epsilon E} \frac{weight(u,v) * PR_t(v)}{\|\{(u,v)\epsilon E\}\|}$$

where u is a node, v is a node that has a link with u, d is a damping factor (i.e. a probability that the process of following the links will continue), V the set of nodes, E the set of edges, and $PR_t(u)$ is the page-rank of node u at step t, which is updated iteratively. A node is assigned an initial page-rank value of 1 if it is a seed and a value of 0.01 if it is not a seed. The initial page-rank values are normalised so that $\sum_{u\epsilon V} PR_0(u) = 1$. In *GUILD* the damping factor is set to 0.15.

4.2.2.2.6 Random walk with restart

The Random walk with restart algorithm by Köhler et al. (2008) iteratively simulates random moves of a walker from a node to a random neighbouring node, such that the walk can be restarted at any time step depending on a preset probability. Random walk differs from PageRank with priors by the way it normalises the weights of links. The scores of nodes are calculated as follows:

$$p_{t+1} = (1-r) * Wp_t + r * p_0$$

where p_t is a vector where each element i in the vector holds the probability of arriving at node i of the network at a given time step t, W is a column normalised adjacency matrix and r is the restart probability (Guney and Oliva, 2012). p_0 is a vector that holds the initial probabilities for the nodes with the sum of these probabilities totalling 1 (similar to the initial page-rank values above). p_{t+1} holds the disease-association probabilities of nodes once the convergence is reached. The convergence is determined by p_{t+1} and p_t having a difference less than $10e^{-6}$ or by achieving the limit of the number of iterations (typically set at 50).

4.2.2.2.7 Network propagation

The network propagation algorithm modifies Random walk with restart by normalising the edge weight by the number of incoming edges as well as the number of outgoing edges. Guney and Oliva (2012) implemented the algorithm defined by Vanunu et al. (2010) and Erten et al. (2011) in *GUILD*. The network propagation is calculated as:

$$p_{t+1} = (1-r) * W'p_t + r * p_0$$

where p_t, p_0, and r are all defined as they were for Random walk with restart. While W' is the adjacency matrix whose elements are normalised by the square root of the multiplication of degrees of the nodes that define the edge at each cell (Guney and Oliva, 2012).

4.2.2.3 Analysis of prioritisation scores

The algorithms used for prioritisation were NetCombo, Page Rank with priors, Random walk with restart and Network propagation. After calculating the scores using these methods, the associations between the various scores assigned by the algorithms were tested with a Spearman's rank correlation (Zar, 2005). Spearman's rank correlation is a non-parametric test for measuring the strength and direction of association between two ranked, monotonic variables. The Spearman's rank correlation test returns a coefficient ρ, which is high when observations have a similar rank, and low when observations have a different rank. The test was executed using the built-in *Python* function, Scipy.stats.spearmanr, which takes two arrays as input and returns the coefficient ρ and a *p*-value. The percentile rank of the 28 proteins interacting with both pathogens was calculated for each method.

4.2.3 Variant data analysis

To assess if known variants with clinical consequences for HIV or TB infection are located in proteins that have high network importance, a database of clinically relevant variants was downloaded from ClinVar. ClinVar is a publicly available repository of relationships between human variations and phenotypes (Landrum et al., 2020). The data as of 13 February 2021 were downloaded. The ClinVar data were filtered based on the listed phenotype to only include variants associated with HIV or TB. The following phenotype descriptions were used to filter the variants: "Mycobacterium tuberculosis, protection against"; "Mycobacterium tuberculosis, susceptibility to"; "Mycobacterium tuberculosis, susceptibility to infection by"; "Human immunodeficiency virus type 1, susceptibility to"; "Human immunodeficiency virus type 1, increased perinatal transmission of"; "Human immunodeficiency virus dementia, susceptibility to"; "Human immunodeficiency virus type 1, rapid disease progression with infection by"; "HIV-1 viremia, susceptibility to"; and "Human immunodeficiency virus type 1, rapid progression to AIDS". In total, 37 variants met this criteria.

To test whether or not proteins that have clinically relevant variants have higher network importance, Wilcoxon rank-sum tests were performed comparing the following network properties between proteins with and without clinically relevant variants: Pathogen Bridging Centrality, Degree, Betweenness, Closeness, as well as the Minimum distance to MHC proteins, *Mtb* proteins, and HIV-1 proteins. In addition, the network prioritisation scores (Page rank with priors, Random walk with restart, Network Propagation, and NetCombo) were compared for these proteins.

To assess if human proteins that may be important in the functional host-pathogen PPIN, have variants that may be relevant to co-infection, common variants (minor allele frequency of at least 0.01 in at least one of the 1000 Genomes populations) in dbSNP build 151 were downloaded (the file downloaded from the NCBI FTP site was called

common_all_20180423.vcf). Common SNPs were chosen for this analysis, as if the variants are common and deleterious, they are probably more important to account for when studying drug response than if they are extremely rare. To limit the scope of the analysis to variants that will likely be consequential, only variants tagged in the VCF file as non-synonymous variants were included. The following tags were included: (1) NSF: non-synonymous frameshift, a coding region variation where one allele in the set changes all downstream amino acids; (2) NSM: non-synonymous missense, a coding region variation where one allele in the set changes protein peptide; and (3) NSN: non-synonymous nonsense, a coding region variation where one allele in the set changes to STOP codon. The variants were further filtered so that only genes that mapped to proteins in the PPIN were included. In total, 120 521 variants in 12 881 proteins were included, six of which were already included as clinically-associated variants. Of the variants, 1662 were frameshift, 116 747 were missense, 1532 were nonsense and 581 were both missense and nonsense variants.

SNPNexus, a web-based tool for variant annotation, was used to extract the predicted effect on protein function, amino acid changes, and allele frequencies in African, East Asian, and European populations for each of the variants (Oscanoa et al., 2020). Using SNPNexus, SIFT (Kumar et al., 2009) scores and PolyPhen (Adzhubei et al., 2010) scores were calculated to predict the effect of the variants on protein function. If a SNP affected multiple transcripts, then the most deleterious score was chosen per allele change. Allele frequencies in the African, East Asian, and European samples in 1000 Genomes, gnomAD exomes, and gnomAD genomes projects were extracted. If any of the minor allele frequencies from the three projects was ≥ 0.01, the variant was annotated as common in African populations. Fisher's exact tests were performed on each pair of populations allele frequencies for each project, using the *Python* module, Fisher. The resulting *p*-values were corrected using Benjamini Hochberg multiple testing correction using the *Python* function statsmodels.stats.multitest.fdrcorrection. Any SNPs that had significantly higher frequency in an African population than another population were flagged, and the projects in which the significant difference was observed were listed.

In addition, to investigate the potential impact of the 1377 novel variants identified in the MHC region in chapter 3, all of these variants were mapped to the PPIN, including non-coding variants and variants in regulatory regions.

4.2.4 Integrating the data using a graph database

Neo4j version 4.2.1 was used to construct the graph database. The graph database is constructed by defining nodes and edges (relationships between the nodes). Properties are assigned to nodes and edges to annotate them with useful information. Labels are used to broadly categorise nodes and edges so that information can be more efficiently searched for within the graph. Neo4j has an intuitive query language called *Cypher* that enables easy extraction of subsets of nodes and edges. Neo4j also has a visualisation component implemented within the Neo4j desktop application browser. In addition to visualisation within Neo4j, Cytoscape version 3.9.0 (Shannon et al., 2003) was used as visualisation and annotation tool for subnetworks so that they could be exposed in the interactive viewer on the Network

Data Exchange (NDEx) (Pillich et al., 2017).

All human, HIV-1, and *Mtb* proteins and drugs in the interaction network generated in chapter 2 were loaded as nodes, and the interactions between them as edges. Properties were created for the nodes to cover all available annotations, such as gene ontology terms, protein length, protein names, calculated centrality measures, flags for MHC proteins, and bridge proteins (human proteins that interacted with both pathogens). Nodes were additionally annotated with whether or not they were differentially expressed during HIV-TB co-infection, as well as the *p*-values, logFC, and mean expression values as outputted by Geo2R. In addition, nodes were annotated with the prioritisation scores from each of the algorithms used. Edges were annotated with the interaction scores as well as any details of the source of the interaction. Variants were also imported as nodes with edges linking them to the protein products of the genes they are located in.

4.3 Results

The human-pathogen PPIN, gene expression, and genetic variation data were integrated into a graph database using Neo4j. The final database contained 19 155 proteins (15 774 human, 3370 *Mtb*, 11 HIV-1), 36 drugs, and 158 651 human variant alleles (37 clinically-associated with HIV or TB, 157 324 listed as a common non-synonymous SNP in dbSNP, and 1290 novel MHC variants). The database contained 583 489 relationships including 407 996 functional human-human PPIs, 15 172 *Mtb* PPIs, 890 human-HIV-1 PPIs, 339 human-*Mtb* PPIs, 239 drug target interactions, and 158 823 links to variants. There were 3486 human proteins labelled as differentially expressed during co-infection. The nodes and relationships were annotated with the properties listed an Appendix in Table 7.6. The combined dataset enabled easier analyses and visualisations. In this section, we describe the network importance of differentially expressed genes and how gene expression was used to prioritise proteins. In addition, we describe the proteins with known clinically-associated variants, and identify other variants that may be clinically relevant. All figures are browsable as interactive networks and downloadable from NDEx at the following link https://www.ndexbio.org/#/networkset/8d0f15ef-d936-11ec-b397-0ac135e8bacf?accesskey=d53b38beb93e7be80f374d3d8cebc626dcd5680a6767521ac9bac80e92c20e7a. The networks can be opened directly in Cytoscape from the NDEx interface.

4.3.1 Network importance of differentially expressed genes

The results showed that differentially expressed genes had significantly higher Pathogen Bridging Centrality, Degree, Betweenness, and Closeness than non-differentially expressed genes (see Table 4.2). In addition, the differentially expressed genes had shorter minimum distance to the MHC proteins, as well as shorter minimum distance to the *Mtb* and HIV-1 proteins. Altogether, these results indicate that the proteins corresponding with differentially expressed genes functionally interact more closely with pathogen proteins, as well as human proteins that are involved in immune response than non-differentially expressed genes. Of the 28 proteins interacting with both pathogens, 11 mapped to differentially expressed genes, and an additional five would have mapped to differentially expressed genes if a

p-value $\leq$0.05 was used as the cut-off for significance. The log fold change in gene expression during co-infection, HIV infection, and TB infection compared to latent TB infection is displayed for the 28 bridge proteins in Figure 4.1. Three of the bridge proteins, namely NAP1L1, CDC42, and PRKCH, were differentially expressed in all three disease states compared to the latent TB infected controls (p<0.05).

Table 4.2 Wilcoxon rank-sum test comparing network properties of proteins mapped to genes differentially expressed during HIV-TB co-infection compared with latent TB infection.

Network property	DE genes (n=3486) median (IQR)	Non-DE genes (n=11 636) median (IQR)	S	p-value
Pathogenicity Bridging	0.0003 (0, 0.011)	0.0 (0, 0.004)	12.37	<0.001
Degree	21 (6, 66)	15 (4, 53)	8.45	<0.001
Betweenness	3577.86 (210.23, 20397.67)	1526.72 (43.92, 13073.65)	11.83	<0.001
Closeness	0.29 (0.27, 0.32)	0.28 (0.26, 0.31)	15.82	<0.001
Minimum distance to MHC proteins	2(1, 2)	2 (2, 2)	-10.73	<0.001
Minimum distance to *Mtb* proteins	3 (2, 3)	3 (2, 3)	-9.99	<0.001
Minimum distance to HIV-1 proteins	2 (2, 3)	2, (2, 3)	-9.53	<0.001

4.3.2 Gene prioritisation based on differential expression

Four algorithms were used to prioritise genes in the network based on genes that were differentially expressed, namely: Page rank with priors, Random walk with restart, Network propagation, and NetCombo. The prioritisation took into account both the strength of the differential expression (absolute log fold change for significantly differentially expressed genes), and the confidence score of the interaction. Spearman's rank correlation was used to assess the correlation between each pair of measures. The measures were all strongly positively correlated with each other (see Table 4.3). Network propagation and Random walk with restart were the highest correlated measures, while NetCombo and Page rank with priors were the least correlated measures.

Table 4.3 Spearman's rank correlation between scores of gene prioritisation algorithms.

Algorithm scores compared	Spearman's ρ	p-value
NetCombo vs. Random walk with restart	0.911	<0.001
NetCombo vs. Network Propagation	0.932	<0.001
NetCombo vs. Page rank with priors	0.904	<0.001
Network propagation vs. Random walk with restart	0.984	<0.001
Network propagation vs. Page rank with priors	0.930	<0.001
Random walk with restart vs. Page rank with priors	0.952	<0.001

The percentile rank of the 28 human proteins that functionally interacted with both an HIV-1

Protein	LTB vs. HIV-TB	LTB vs. PTB	LTB vs. HIV
EIF2AK2	-1.11	0.21	-0.79
NAP1L1	0.59	-0.23	0.33
CDC42	-0.76	0.85	-0.34
PRKCQ	0.60	-0.67	-0.01
LCK	0.56	-0.77	-1.05
GAPDH	-0.33	0.28	-0.05
CALM1	0.27	-0.38	0.04
FYN	0.41	-0.50	-0.18
CAV1	-0.75	0.14	-0.82
PRKCH	0.46	-0.67	-0.21
CTSD	-0.41	0.37	0.06
EPRS	0.22	-0.22	-0.05
HNRNPR	0.25	-0.26	0.12
MSN	0.17	-0.12	0.03
YBX1	0.29	0.03	-0.03
RAC2	-0.11	0.02	-0.12
MAP3K5	-0.11	0.13	0.06
HSPD1	0.20	0.24	0.09
STAT3	-0.16	0.30	0.09
CBFB	0.16	-0.23	0.07
RAB7A	-0.08	0.14	0.05
NFKB1	0.04	0.02	-0.01
HEXIM1	-0.06	0.04	-0.03
CD209	-0.47	0.95	-0.58
AKT2	-0.36	0.57	-0.68
HSP90B1	-0.05	0.07	-0.24
FN1	0.06	-0.13	0.25
ARF6	-0.01	0.03	-0.05

Significance level key

<0.001

<0.01

<0.05

Figure 4.1 Log fold change in gene expression levels of the bridge proteins during HIV-TB co-infection, HIV infection, and TB infection compared to latent TB infection.
This tabular diagram shows the log fold change in expression of the 28 bridge proteins during three disease states compare to latent TB infection (HIV-TB co-infection, HIV infection, and TB infection). For each column, latent TB infection (LTB) is the numerator, and the log is calculated as base 2. Cells are coloured by the significance level of the differential expression as indicated in the key. Proteins that were significantly differentially expressed ($p<0.001$) are in bold and proteins that were marginally significantly differentially expressed ($p<0.05$) are italicised. Proteins that were significantly or marginally significantly differentially expressed during all three disease states are indicated by an orange rectangle surrounding the cells.

and a *Mtb* protein (bridge proteins) was assessed for each algorithm. On average, the bridge proteins scored well above the upper quartile (83 percentile rank) across all four measures. Only three bridge proteins did not score in the upper quartile for any measure (HEXIM1, CBFB, and MSN). CDC42, EIF2AK2, and CAV1 were the three top scoring bridge proteins, scoring above the 95^{th} percentile for all measures. The prioritisation scores for each of the 28 bridge proteins as well as their percentile rank for each algorithm are displayed in Table 4.4

Table 4.4 Prioritisation scores and ranking of the bridge proteins

Protein	Page Rank Rank (Score)	Random Walk Rank (Score)	Network propagation Rank (Score)	NetCombo Rank (Score)
CDC42	99.9 (0.078)	98.0 (4.16E-04)	95.5 (2.78E-04)	96.8 (0.357)
EIF2AK2	95.0 (0.021)	97.6 (3.82E-04)	97.6 (3.72E-04)	98.1 (0.403)
CAV1	99.3 (0.044)	97.1 (3.54E-04)	95.3 (2.72E-04)	96.0 (0.337)
LCK	99.6 (0.051)	96.1 (3.13E-04)	93.0 (2.19E-04)	94.6 (0.315)
FYN	99.9 (0.076)	97.3 (3.63E-04)	91.1 (1.87E-04)	93.0 (0.293)
PRKCQ	97.7 (0.029)	92.5 (2.21E-04)	92.2 (2.03E-04)	93.7 (0.302)
NAP1L1	95.1 (0.022)	92.5 (2.20E-04)	92.6 (2.11E-04)	94.4 (0.312)
CALM1	99.7 (0.060)	94.3 (2.59E-04)	86.2 (1.31E-04)	89.3 (0.262)
CTSD	94.8 (0.021)	93.0 (2.30E-04)	88.3 (1.51E-04)	89.8 (0.266)
GAPDH	98.3 (0.032)	90.3 (1.85E-04)	86.9 (1.38E-04)	89.8 (0.266)
STAT3	99.8 (0.067)	94.3 (2.59E-04)	81.6 (5.81E-05)	82.4 (0.219)
NFKB1	99.6 (0.049)	87.6 (1.52E-04)	81.4 (5.06E-05)	84.9 (0.230)
FN1	97.7 (0.029)	82.6 (1.00E-04)	81.1 (4.35E-05)	81.6 (0.217)
PRKCH	75.6 (0.008)	87.4 (1.50E-04)	88.3 (1.51E-04)	87.9 (0.252)
RAC2	97.8 (0.029)	80.3 (7.07E-05)	79.0 (2.94E-05)	73.3 (0.200)
EPRS	90.2 (0.015)	78.6 (5.09E-05)	79.8 (3.22E-05)	79.4 (0.213)
YBX1	96.5 (0.025)	78.7 (5.19E-05)	78.7 (2.88E-05)	73.7 (0.201)
ARF6	97.0 (0.026)	80.6 (7.41E-05)	79.3 (3.02E-05)	69.9 (0.193)
HSPD1	94.2 (0.020)	79.0 (5.41E-05)	79.0 (2.95E-05)	73.4 (0.200)
AKT2	97.7 (0.029)	81.5 (8.83E-05)	79.2 (2.99E-05)	64.5 (0.185)
HSP90B1	96.8 (0.026)	80.6 (7.43E-05)	78.7 (2.87E-05)	66.1 (0.188)
MAP3K5	91.5 (0.017)	77.3 (4.20E-05)	76.7 (2.54E-05)	73.3 (0.200)
HNRNPR	90.5 (0.016)	75.7 (3.51E-05)	72.0 (2.01E-05)	73.6 (0.201)
RAB7A	85.7 (0.013)	76.6 (3.82E-05)	65.9 (1.62E-05)	66.2 (0.188)
CD209	73.0 (0.008)	74.4 (3.10E-05)	76.0 (2.43E-05)	63.8 (0.184)
MSN	72.9 (0.008)	65.7 (1.80E-05)	61.8 (1.44E-05)	58.9 (0.176)
CBFB	51.8 (0.003)	51.6 (7.46E-06)	50.1 (9.26E-06)	55.2 (0.170)
HEXIM1	48.2 (0.002)	46.0 (4.59E-06)	41.3 (5.86E-06)	45.4 (0.153)

The top ranking protein for Network propagation and NetCombo was IFI27 (Interferon alpha-inducible protein 27), and the top ranking protein for Page rank with priors and Random walk with restart was UBC (Polyubiquitin-C). Neither of these proteins functionally interacted with an HIV-1 or *Mtb* protein in this network. Both were significantly differentially expressed during co-infection compared with latent TB infection. UBC seems to have been prioritised due to it being a hub protein in the network, as it is involved in over 4000

interactions in the PPIN. IFI27 seems to have been prioritised due to it having the largest change in gene expression during co-infection (10 times higher during co-infection vs. latent TB infection). The top ten ranked proteins for each of the four prioritisation algorithms along with their priority score is displayed in Table 4.5.

Table 4.5 Top ten ranking proteins across each of the four prioritisation algorithms

Page Rank Protein (Score)	Random Walk Protein (Score)	Network Propagation Protein (Score)	NetCombo Protein (Score)
UBC (1.000)	UBC (6.06E-03)	IFI27 (1.48E-03)	IFI27 (1.000)
TP53 (0.158)	IFI27 (1.51E-03)	OTOF (1.32E-03)	C1QB (0.822)
SRC (0.125)	OTOF (1.32E-03)	C1QB (1.20E-03)	SERPING1 (0.822)
AKT1 (0.112)	C1QB (1.21E-03)	SERPING1 (1.19E-03)	OTOF (0.806)
HSP90AA1 (0.109)	SERPING1 (1.19E-03)	CD177 (1.15E-03)	CD177 (0.790)
CDK1 (0.108)	BATF2 (1.15E-03)	BATF2 (1.15E-03)	BATF2 (0.776)
RPS27A (0.099)	CD177 (1.14E-03)	SEPT4 (1.06E-03)	DEFA4 (0.775)
UBA52 (0.088)	ISG15 (1.13E-03)	CEACAM8 (1.05E-03)	RSAD2 (0.752)
MYC (0.085)	FCGR1A (1.06E-03)	C1QC (1.05E-03)	SEPT4 (0.749)

Of the 145 MHC proteins in the network, 125 were assessed for differential expression, of which 46 were shown to be significantly differentially expressed during co-infection. Analysis of the prioritisation scores of the MHC proteins showed that they had consistently higher prioritisation scores across the four measures; however, only NetCombo was marginally significantly higher (p=0.04) for MHC proteins than non-MHC proteins (see Table 4.6). The expression values and prioritisation scores for all the MHC proteins is available as an appendix in Table 7.7.

Table 4.6 Wilcoxon rank-sum test comparing prioritisation measures for MHC and non-MHC proteins

Prioritisation measure	MHC protein (n=145) median (IQR)	Non-MHC protein (n=15 629) median (IQR)	S	p-value
Page Rank with priors	0.006 (0.0007, 0.01)	0.003 (0.0008, 0.01)	1.88	0.06
Random Walk with restart	2.44e-05 (2.5e-06, 1.27e-04)	1.2e-05 (2.0e-06, 5.61e-05)	1.76	0.08
Network propagation	1.8e-05 (5.19e-06, 1.19e-04)	1.29e-05 (4.52e-06, 3.00e-05)	1.35	0.17
NetCombo	0.19 (0.15, 0.24)	0.18 (0.14, 0.21)	2.09	0.04

4.3.3 Known variants associated with HIV and *Mtb*

The database of clinical variants (ClinVar) was filtered to identify any variants related to HIV-1 or *Mtb* infection. 37 variants were identified, including 1 insertion, 2 deletions, and 34 single nucleotide variants. The 37 variants were located within 21 genes, and all 21 genes were in the functional human-pathogen PPIN. One "bridge" protein in the network, CD209, had a variant associated with both HIV-1 and *Mtb* susceptibility. One MHC protein, TNF, had variants associated with HIV-1. Seven of the 21 proteins were significantly differentially

expressed during HIV and TB co-infection, namely: CCR2, CCL2, TLR2, IFNG, IFNGR1, SP110, BST2. In addition, one of the proteins, CCR5, was listed as a drug target.

The 21 proteins that had clinically-associated variants had significantly stronger network properties than the 15 753 human proteins without clinically-associated variants. Similarly, the prioritisation scores that were calculated based on gene expression were significantly higher for the 21 proteins than the other human proteins. The results of this analysis are displayed in Table 4.7.

Table 4.7 Wilcoxon rank-sum test comparing network properties of proteins mapped to genes with and without variants associated with HIV-1 or *Mtb* infection.

Network property	With clinical variants (*n*=21) median (IQR)	Without clinical variants (*n*=15 753) median (IQR)	S	*p*-value
Pathogenicity Bridging	0.03 (0.02, 0.11)	0.0 (0, 0.005)	5.18	<0.001
Degree	103 (48, 212)	16 (4, 58)	4.33	<0.001
Betweenness	25931.99 (6144.68, 79033.22)	1754.71 (50.28, 14272.72)	4.02	<0.001
Closeness	0.31 (0.29, 0.32)	0.28 (0.26, 0.31)	3.84	<0.001
Minimum distance to MHC proteins	1(1, 2)	2 (2, 2)	-3.52	<0.001
Minimum distance to *Mtb* proteins	2 (2, 2)	3 (2, 3)	-3.86	<0.001
Minimum distance to HIV-1 proteins	2, (1, 2)	2, (2, 3)	-4.59	<0.001
Page Rank with priors	0.015 (0.008, 0.02)	0.003 (0.0008, 0.01)	4.96	<0.001
Random Walk with restart	5.97e-05 (2.24e-05, 1.95e-04)	1.2e-05 (2.0e-06, 5.64e-05)	3.80	<0.001
Network propagation	3.63e-05 (1.92e-05, 1.88e-04)	1.29e-05 (4.52e-06, 3.02e-05)	3.60	<0.001
NetCombo	0.22 (0.21, 0.29)	0.18 (0.14, 0.21)	4.47	<0.001

4.3.4 Variants in prioritised proteins

Non-synonymous common variants in dbSNP were mapped to the functional interaction network along with the clinically-associated variants for HIV and TB. Variants were categorised within prioritised proteins and bridge proteins based on their allele frequency and the predicted effect of the SNP on the protein. In total, 12 884 proteins in the network had at least one common non-synonymous SNP or clinically-associated variant. There were 120 552 SNPs when counting by the SNP RS identifier, which corresponded with 157 361 different unique variants when counting by allele change. Allele frequencies from 1000 Genomes, gnomAD genomes and gnomAD exomes for African, East Asian and European populations were extracted where available. Variants with an allele frequency of at least 0.01 in at least one of the aforementioned studies were considered common in African populations, which resulted in 39 071 of the 157 361 alleles (24.82%) being classified as common. Variants with a significantly higher allele frequency in the African population than one of the other populations in at least one of the three studies were identified. In total, 62 962 of the 157 361

alleles (40.01%) had a significantly higher frequency in an African population in one of the studies. Of the alleles that had a higher frequency in an African population, 32 115 were also common (51.01%). We used SIFT scores and PolyPhen scores to identify potentially damaging or deleterious SNPs. Of the 157 361 alleles, 61 240 (38.91%) had a SIFT score between 0 and 0.05 (deleterious) or a PolyPhen score between 0.85 and 1 (probably damaging). Of these potentially damaging or deleterious variants, 11 771 (19.22%) were common in African populations. In addition, 22 888 of the potentially damaging or deleterious variants (37.37%) had a significantly higher frequency in an African population.

Variants mapping to the 28 bridge proteins that functionally interact with HIV-1 and *Mtb* proteins were identified. In total, there were 139 non-synonymous or clinically-associated alleles that have been reported in 20 of the 28 bridge proteins. Of these variants, 69 (49.64%) were potentially damaging or deleterious SNPs according to SIFT or PolyPhen. Of the damaging or deleterious variants within bridge proteins, 9 (13.04%) were common in African populations, and 24 (34.78%) had significantly higher frequency in an African population. In addition to the bridge proteins, variants in the MHC proteins, drug-targets, and the proteins with average priority scores in the top 1% were categorised (see Table 4.8).

Table 4.8 Description of variants within prioritised proteins

Protein set	Total alleles	Damaging or Deleterious	Common in African populations	Higher frequency in African populations	Higher frequency in African populations of those Damaging or Deleterious
Total with variants (*n*=12 884)	157 361	61 240 (38.92%)	39 078 (24.83%)	62 972 (40.02%)	22 888 (37.37%)
Top 1% priority scores (*n*=160)	1780	805 (45.22%)	428 (24.04%)	730 (41.01%)	308(38.26%)
MHC (*n*=99)	1327	577 (43.48%)	465 (35.04%)	583 (43.93%)	258 (44.71%)
Drug target (*n*=64)	959	339 (35.35%)	220 (22.94%)	379 (39.52%)	120(35.40%)
Bridge (*n*=20)	139	69 (49.64%)	34 (24.46%)	57 (41.01%)	24(34.78%)
Damaging or deleterious variants in bridge proteins (*n*=20)	69	69 (100.00%)	9 (13.04%)	24 (34.78%)	24(34.78%)

Using the graph database and Neo4j Desktop, the variants can be visualised alongside the protein and drug interactions for further exploration as shown in Figure 4.2. This figure shows all interactions with human proteins that functionally interact with both *Mtb* and HIV-1 proteins and were differentially expressed during HIV-TB co-infection. Clicking on any node or relationship from the Neo4j browser will display all properties associated with that element.

A Neo4j database backup has been uploaded to a private folder in the Open Science Framework (OSF) and may be downloaded from the following link https://osf.io/expdb/?view_only=454f800e41e440e990d53ff59701ed54. All of the

subnetworks displayed in the figures that follow have been annotated in Cytoscape and are accessible in NDEx. Neo4j is recommended for querying the full network, however the network is also queryable and downloadable from NDEx at the following link: https://www.ndexbio.org/#/network/80686816-74a3-11ec-b3be-0ac135e8bacf?accesskey=3978a315c9936737927a3cfa9aedae463b1ec32281d33d69e42c7dc6ef3db2b1.

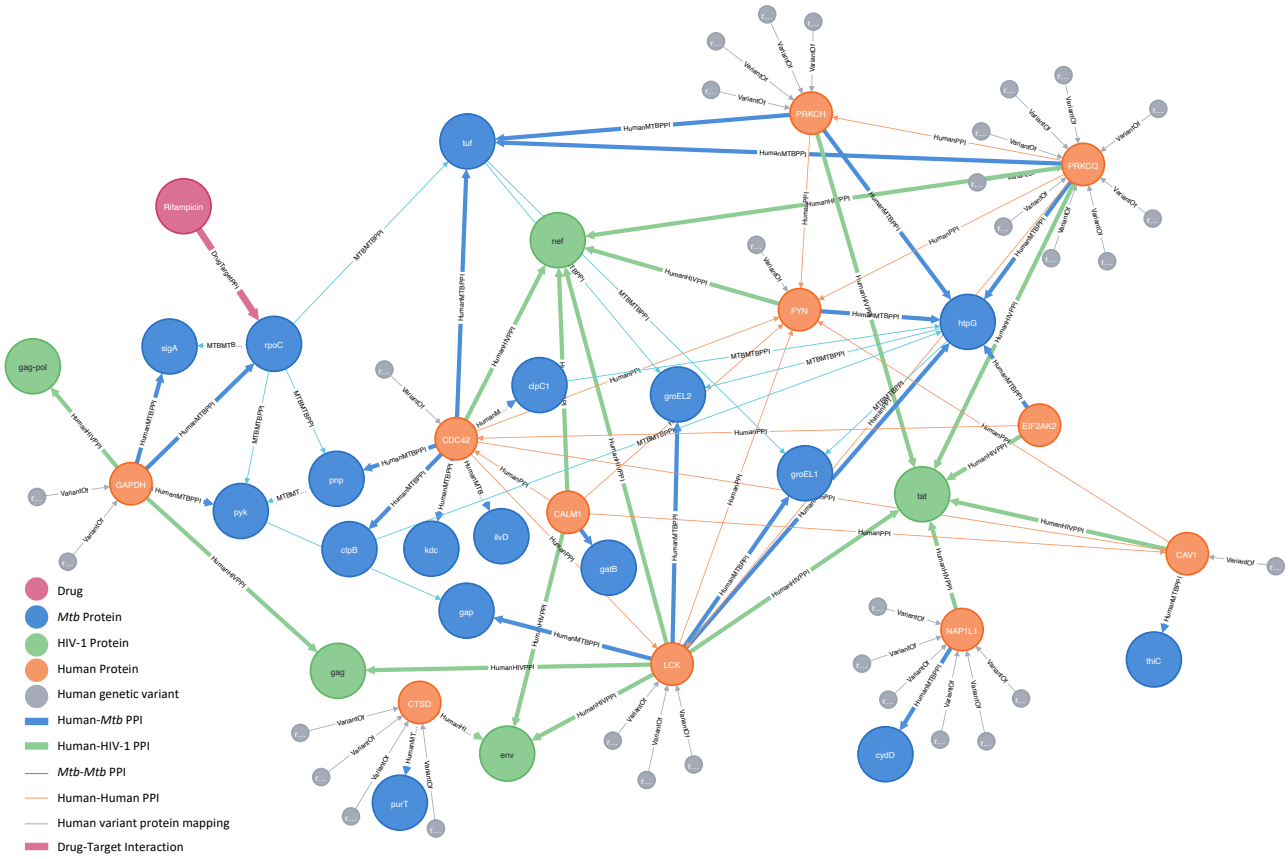


Figure 4.2 Visualisation of PPIs, Drug-target interactions and variants. The visualisation from Neo4j shows protein-protein interactions, drug-target interactions, mappings between proteins and variants as well as differential expression information. The human proteins (orange circles) have been filtered on proteins that are "bridges" and are differentially expressed during HIV and TB co-infection. The blue and green circles represent *Mtb* and HIV-1 proteins respectively, with thick blue and green lines representing human-*Mtb* and human-HIV-1 protein interactions. The smaller grey circles are variants of the human proteins they map to. The pink circles are drugs and the thick pink lines are drug target interactions. This subnetwork is available in an interactive viewer at the following link: https://www.ndexbio.org/#/network/903731d3-7645-11ec-b3be-0ac135e8bacf?accesskey=3f091c6e873cb49d789fefaa1bf4c1999f30db7b77c1d7b09db6028fd39399bf.

4.3.5 Properties of MHC proteins with novel variants identified from the sequence graph

In addition to common and clinical variants, the novel variants in the MHC region that were identified from the sequence graph created in chapter 3 were mapped to the host-pathogen PPIN. All variants identified were included regardless of the gene region, so that the possibilities for exploring the impact of non-coding variants on the PPIs could be illustrated.

Table 4.9 Novel MHC variants that were mapped to the PPIN

Variant type	Number of variants	Number of proteins
exonic	3	3
upstream	9	3
UTR3	10	1
downstream	16	5
intronic	53	12
intergenic	1199	27

Of the 1377 variants identified, 1290 (93.68%) mapped to proteins in the PPIN. There were only 3 exonic variants and 1199 (92.95%) intergenic variants of which 172 had both genes in the PPIN. The number of variants by location is displayed in Table 4.9.

The 1290 variants mapped to 36 proteins, of which 30 were classified as MHC proteins. The six non-MHC proteins were included, because intergenic variants could also map to genes outside of the MHC region. Of the MHC proteins, ten interacted with HIV-1 proteins, ten interacted with bridge proteins and two interacted with drug targets. In addition, seven were differentially expressed during HIV-TB co-infection. A visualisation of the variants and interactions is displayed in Figure 4.3. In addition, a list of the 30 MHC proteins is provided in Table 7.8 with a summary of the number of variants, pathogen interactions, bridge protein interactions, drug target interactions, differential expression, and priority ranking.

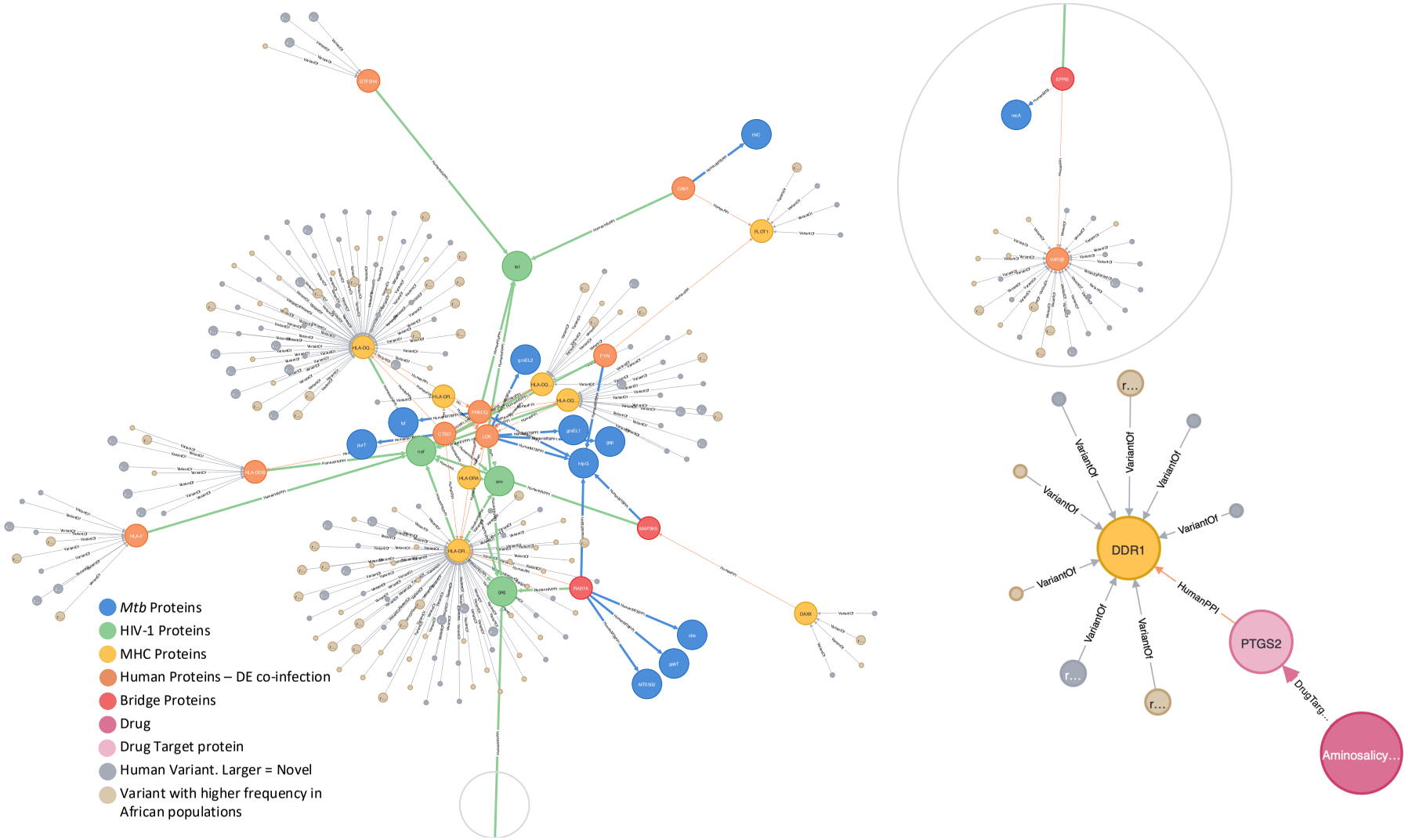


Figure 4.3 Visualisation of MHC proteins, with their variants and interactions. The visualisation shows MHC proteins (yellow), differentially expressed proteins (orange) as well as their interactions with bridge proteins (red), HIV-1 proteins (green), *Mtb* proteins (blue), drug target proteins (light pink), and drugs (dark pink). It also shows links to variants in the MHC proteins, with smaller grey circles representing known variants, larger grey circles representing novel variants and beige circles representing known variants with higher allele frequency in an African population. Intergenic novel variants were excluded from the visualisation for clarity. The light grey circle over the subnetwork on the top right is a continuation of the large subnetwork on the left where it gets cut-off at the bottom of the page. On the left the visualisation shows a complex subnetwork of interactions between many different MHC proteins, bridge proteins and pathogen proteins. On the right a simple subnetwork is shown illustrating the interaction between a MHC protein (DDR1) and a drug-target protein (PTGS2) as well as the drug (Aminosalicylic acid). Eight variants of DDR1 are displayed, three of which are novel and four had a higher frequency in an African population. This subnetwork is available in an interactive viewer at the following link: https://www.ndexbio.org/#/network/6ddf7917-7700-11ec-b3be-0ac135e8bacf?accesskey=af63b2847108576c058f7575e2f48b703877fe56c35d2f6336c601acc6ba5985.

4.4 Discussion

In this chapter, we integrated gene expression and genetic variation information with the functional human-pathogen protein-protein interaction network created in chapter 2. We showed that genes that were differentially expressed during HIV-TB co-infection also have significantly stronger network properties than genes that were not differentially expressed. Similarly, variants with known clinical significance in HIV or TB were shown to map to proteins with significantly stronger network properties than proteins without such variants. Finally, we identified common and consequential variants within prioritised proteins that may be clinically-associated with HIV and TB. In this section, we discuss these results as well as the potential impact of the common variants identified within proteins that interact with both pathogens.

4.4.1 Differentially expressed genes have higher network importance

To our knowledge this study is the first to analyse the network properties of genes within a host-pathogen protein interaction network alongside differential expression during HIV-TB co-infection. The finding that differentially expressed genes had significantly higher network importance than non-differentially expressed genes provides reassurance that network measures can be used as a proxy for biological relevance. The measure proposed to detect importance for involvement in functional interactions between the pathogens (Pathogenicity bridging centrality), was shown to be significantly higher for differentially expressed genes than non-differentially expressed genes. This shows that the measure may be useful for ranking genes based on their potential importance in HIV-TB co-infection.

Three of the bridge proteins were significantly differentially expressed during HIV-TB co-infection as well as HIV and TB mono-infection compared with latent TB infection, namely: NAP1L1, CDC42, and PRKCH. Both NAP1L1 and CDC42 functionally interact with HIV-1 protein Tat, while CDC42 functionally interacts with HIV-1 protein Nef. CDC42 and PRKCH both functionally interact with *Mtb* proteins tuf and htpG. The interactions between these three bridge proteins and the pathogen proteins, as well as the variants in the three bridge proteins are depicted in Figure 4.4. All three of the proteins had at least one variant with a higher frequency observed in an African population than an East Asian or European population. In addition, PRKCH and CDC42 had possibly deleterious variants.

Differentially expressed genes were used as seeds to assign priority ranking scores to other proteins in the network. The prioritised genes, as such, tended to be both differentially expressed and have high network centrality measures. The top scoring bridge proteins included EIF2AK2 (Protein Kinase R). In chapter 2, we proposed that EIF2AK2 could be a potential drug target as inhibition of this protein has been shown to lead to apoptosis of *Mtb* infected cells and inhibition of the protein may reduce viral infectivity. EIF2AK2 expression was significantly lower during latent TB infection compared with HIV-TB co-infection (the log fold change was -1.11 which equates to just over 2 times lower).

While few MHC genes were significantly differentially expressed during co-infection, and, probably as a result, MHC proteins did not rank significantly higher on any of the priority

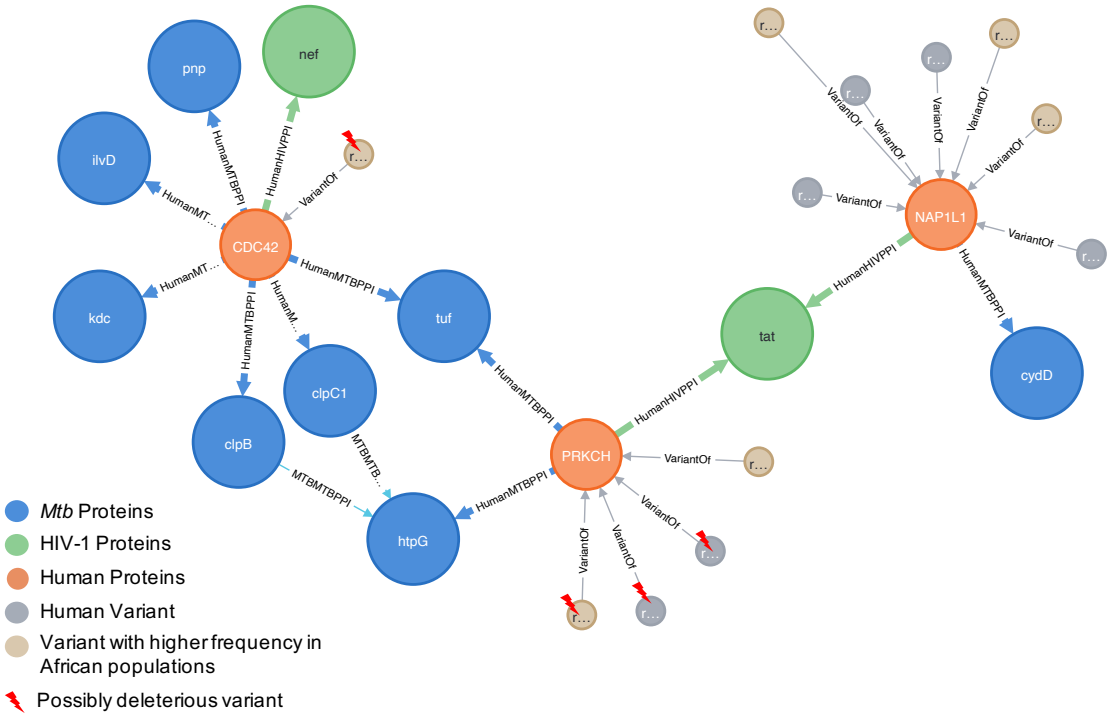


Figure 4.4 Host-pathogen PPIs for bridge proteins that are DE during mono- and co-infection. The human proteins (orange circles) are bridge proteins that are DE during HIV-TB co-infection and mono-infection of the two diseases (NAP1L1, CDC42 and PRKCH). The blue and green circles represent *Mtb* and HIV-1 proteins respectively. The small beige and grey circles represent human variants, with the beige circles highlighting variants with a higher frequency in African populations than East Asian or European populations. Variants that are possibly deleterious based on SIFT/Polyphen scores have been flagged with a red lightning bolt. This subnetwork is available in an interactive viewer at the following link: https://www.ndexbio.org/#/network/9ddc6f74-76ff-11ec-b3be-0ac135e8bacf?accesskey=387cb1d7a05b1bc0570be635c66b4821b433ff1737aa250e1166dedba4814ec4.

scoring measures, differentially expressed genes were shown to have significantly shorter minimum distance to MHC proteins. This corroborates the results in chapter 2, and elevates the role of MHC proteins in facilitating interactions between the pathogens.

4.4.2 Variants associated with HIV and TB are found in proteins with higher network importance

There were few variants with known clinical association with HIV or TB, but the few proteins that had clinically-associated variants ranked significantly higher across the four priority measures used and had significantly higher network centrality. In addition, one third of the proteins that contained clinically-associated variants were differentially expressed during HIV-TB co-infection. This aligns with the findings by Chen et al. (2008) that differentially expressed genes are more likely to have disease-associated DNA variants.

The only variant that was clinically-associated with both HIV-1 and *Mtb* susceptibility was located in the protein CD209. The interactions between HIV-1 and *Mtb* with CD209 were discussed at length in chapter 2, section 2.4.1.6. CD209 both enables the infection of dendritic cells with *Mtb* and enhances the infection of target cells with HIV-1. CD209 is the gene that encodes DC-SIGN, a C-type lectin known to be the major receptor of *Mtb* on human dendritic cells (Barreiro et al., 2006). Within a South African cohort, Barreiro et al. (2006) found that two variants (-871G and -336A) of CD209 were associated with a lower risk of developing tuberculosis and that they may have a higher frequency in non-African populations. Martin et al. (2004) showed that one of the same variants identified by Barreiro et al. (2006) (-336C) was associated with higher susceptibility for HIV-1 infection in participants at risk of infection.

4.4.3 Potentially impactful variants in high priority proteins

Common non-synonymous variants were mapped to proteins in the interaction network and analysed for various sets of prioritised proteins in terms of their allele frequencies in African populations as well as their predicted effect. Several SNPs were identified within drug-target proteins that may impact HIV-TB co-infection. Using the network it is possible to find and visualise potentially deleterious variants in drug-target proteins of interest (refer to Figure 4.5). The treatment of drug-sensitive TB in HIV-infected individuals in South Africa typically includes a multi-drug TB regimen including both isoniazid and rifampicin, as well as an ART regimen which would usually include efavirenz as the NNRTI (Boulle et al., 2008). Using the network we could visualise that all three drugs interact with the human enzyme Cytochrome P450 3A4 (CYP3A4). Efavirenz is metabolised by CYP3A4, and low plasma concentration of efavirenz is associated with virologic failure while high plasma concentration is associated with drug toxicity (Atwine et al., 2018). Rifampicin induces CYP3A4, which enhances efavirenz metabolism and results in a lower plasma concentration (Atwine et al., 2018). This is typically offset by the inibitory effect of isoniazid on CYP3A4, which reduces efavirenz metabolism (Atwine et al., 2018). Two SNPs in CYP3A4 that were predicted to have deleterious effects and were known to have higher alternative allele frequency in African populations were visualised in the network (rs4986908 and rs57409622). rs57409622 was first identified in South African populations and was predicted to have an impact on the enzymes function due to the amino

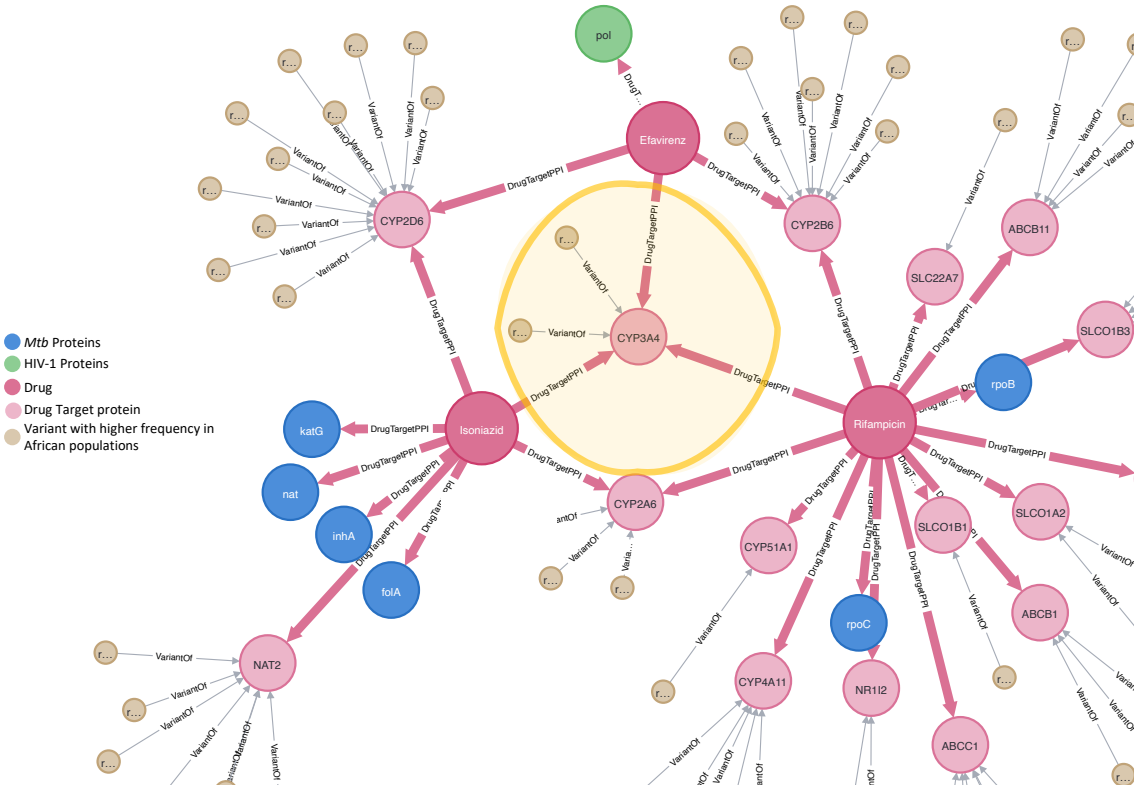


Figure 4.5 Potentially deleterious variants with higher frequencies in African populations in targets of the drugs rifampicin, isoniazid and efavirenz. The network includes three drugs (dark pink circles), human targets (light pink circles), *Mtb* targets (blue circles), HIV-1 targets (green circles) and human genetic variants (small beige circles). Only variants that have higher alternative allele frequency in African populations and that were potentially deleterious are included in the visualisation. A yellow circle highlights the protein CYP3A4 that interacts with all three drugs and has two potentially deleterious SNPs. This subnetwork is available in an interactive viewer at the following link: https://www.ndexbio.org/#/network/2d1bccf1-76fe-11ec-b3be-0ac135e8bacf?accesskey=cb64bc5bab6c318c41a7982ee7741039ee6cdf462a6cf5ca5e8d875c36627ed0.

acid change (Drögemöller et al., 2013). Both mutations may have functional effects on CYP3A4 and thus play a role in drug efficacy or toxicity during the treatment of TB in HIV-infected individuals.

4.4.4 Using the network to investigate the potential impact of MHC variants

Here we discuss an example of how the network can be used to investigate the potential impact of one of the novel variants within the MHC gene HLA-DOB. HLA-DOB was differentially expressed during HIV-TB co-infection and was in the 90th percentile for prioritisation. The novel variant identified was an intergenic five base pair insertion (-/ACAAC) between HLA-DOB and HLA-DQB2 at chromosome 6 position 32782641. An insertion has been found in this position before with a similar pattern (rs28987081, -/C -/CAAC -/CAACAAC) and the variant identified in this analysis may be another alternative allele. Using the network visualisation, both of these MHC proteins functionally interact with the HIV-1 protein Nef and the human bridge protein CTSD, which interacts with the *Mtb* protein PurT and the HIV-1 protein Env (refer to figure 4.6). In chapter 2, section 2.4.1.6 we discuss the

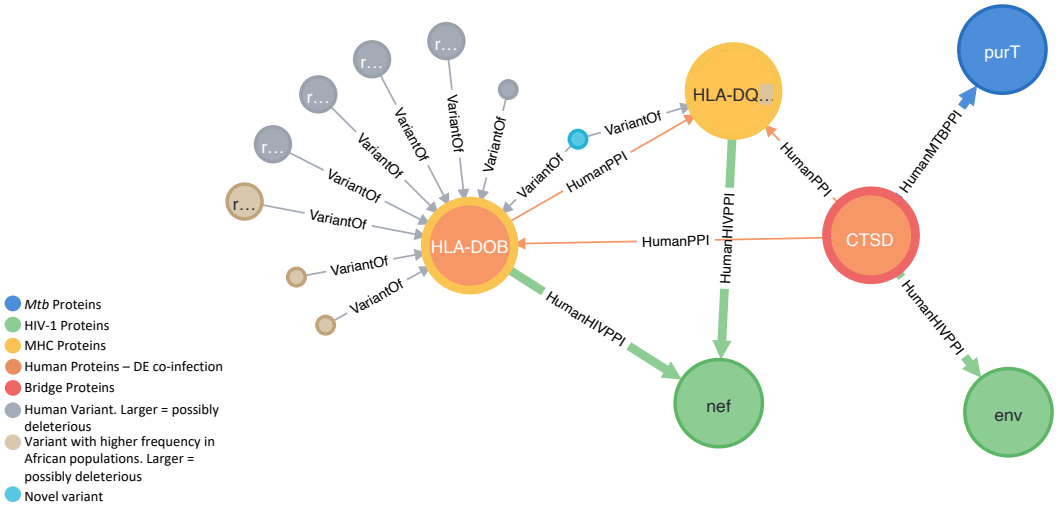


Figure 4.6 Visualisation of a variant in HLA-DOB in the context of the host-pathogen interaction network. The figure shows the novel variant (small light blue circle) is an intergenic variant of the two MHC proteins (large circles with yellow borders), HLA-DOB and HLA-DQB2. The MHC proteins both interact with the bridge protein (large circle with red border) CTSD which interacts with the HIV protein (green circle) Env and *Mtb* protein (large blue circle) PurT. HLA-DOB and CTSD are both differentially expressed during HIV and TB co-infection (illustrated by their orange fill). Additional variants of HLA-DOB are depicted by small grey and beige circles, with beige ones representing variants with higher alternative allele frequency in African populations and larger circles representing variants that are possibly deleterious. This subnetwork is available in an interactive viewer at the following link: https://www.ndexbio.org/#/network/f33d122d-76fa-11ec-b3be-0ac135e8bacf?accesskey=1bd036b286baebfb33cc3928a6649716e595e57b1a6d02b151b1b4c2f6749f66.

function of CTSD, which is involved in viral infectivity. Using the GWAS catalog, several intergenic variants between HLA-DOB and HLA-DQB2 have been significantly associated with various traits, although none specifically report HIV or tuberculosis (Buniello et al., 2019). However a variant in this intergenic region was associated with Kawasaki disease (Onouchi et al., 2012). Kawasaki disease is typically a childhood illness, however amongst adults with the disease there is a disproportionate number that are HIV-infected, which is thought to be due to the immuno-compromise brought about by HIV infection (Johnson et al., 2001). It is possible that this intergenic variant, and others in the region may also be associated with this susceptibility to Kawasaki disease amongst HIV-infected individuals.

4.4.5 Integrating different types of biological data into a graph database

In this chapter, we integrated the gene expression, genetic variation and drug and protein interactions into a graph database that enabled efficient identification of interesting

subnetworks and visualisations across different data types. The user friendly Neo4j browser and *Cypher* query language enable the data to be animated without the need for front-end development. To filter out as much noise as possible, only high confidence interactions and common variants that confer protein changes or variants with known clinical association with HIV or TB were included. While this has the advantage of reducing noise, it also raises a limitation due to study biases. Often interactions are high confidence, and variants well annotated with allele frequency by virtue of them being found in highly studied genes. Only including common missense coding variants is another potential limitation, as variants in non-coding regions can also be clinically relevant. Future work could include a wider range of variants. In addition, only one gene expression study was included in the analysis as it was the only one at the time that compared HIV-TB co-infection to uninfected controls. Triangulating gene expression data from multiple studies with larger sample sizes would be useful and may assist with more accurate prioritisation of proteins within the network.

However, the data included in the current database provide a comprehensive picture of various biological elements involved in HIV-TB co-infection. The annotations attached to each node and relationship allow rapid, in depth analyses and visualisations of subnetworks of interest.

4.5 Conclusion

This chapter aimed to integrate gene expression, genetic variation, and host-pathogen PPIs to facilitate the analysis of the potential impact of variation on HIV-TB co-infection. We identified that genes that are differentially expressed in HIV-TB co-infected individuals have higher network importance in the functional host-pathogen PPIN than genes that are not differentially expressed. Similarly, we showed that variants with known clinical association with HIV and TB map to proteins that have higher network importance. Finally, we identified missense variants within prioritised genes in the network that may warrant further investigation of their impact on HIV-TB co-infection. This analysis has produced a graph database that integrates different kinds of biological information, which provides three useful features, namely: (1) it enables rapid and high throughput prioritisation of sets of genes or variants, (2) it facilitates detailed investigations of smaller sets of genes or variants, and (3) it allows network-based visualisation of the relationships between various biological elements. . All subnetworks in the figures presented, as well as the complete network are browsable as interactive networks and downloadable from NDEx at the following link: https://www.ndexbio.org/#/networkset/8d0f15ef-d936-11ec-b397-0ac135e8bacf?accesskey=d53b38beb93e7be80f374d3d8cebc626dcd5680a6767521ac9bac80e92c20e7a. To extend this resource beyond HIV-TB co-infection, future work may go into providing a regularly updated database with this information alongside other host-pathogen interactions, and gene expression in different disease states.

5. Conclusion

This study focused on two often co-infecting pathogens, HIV and *Mtb*, which continue to evolve to evade host responses and treatment strategies. The interaction between the pathogens and their human host is further complicated by host genetic variation, the study of which is under-represented amongst African individuals. To contribute to the study of genetic variation in African populations, we used a reference free method to identify genetic variation in the highly diverse MHC region, which contains several genes involved in immune response. Additionally, to contribute to the study of HIV-TB co-infection, we aimed to generate a host-pathogen interaction network that integrated the genetic variation with functional protein interactions, gene expression data and drug interactions.

5.1 Contributions

The under-representation of African sequences in genome analysis is widely acknowledged. Despite containing the most genetic diversity, less than two percent of all human genomes analysed so far have originated in Africa (Wonkam, 2021). Analysing sequences from 33 African individuals spanning the continent, we identified novel variants in the MHC region using a reference-free graph-based method for variant identification, which to our knowledge has not been used at this scale before. The novel variants that we identified provide a contribution to the knowledge of variation in this region, particularly in African sequences, which are subject to reference mapping bias when using traditional methods of variant identification (Dilthey et al., 2015). Furthermore, in our analysis we developed a pipeline for sequence formatting, variant identification and variant annotation using existing tools that have not been used together in this manner previously. This pipeline could be used and improved by others.

We have generated, to our knowledge, the first host-pathogen interaction network that displays inter- and intraspecies interactions between HIV, *Mtb* and Human proteins, along with drug-target interactions. Using the network we identified 28 Human "bridge" proteins that functionally interact with both HIV and *Mtb* proteins. One of the bridge proteins, Protein Kinase R (EIF2AK2), was significantly differentially expressed during HIV-TB co-infection and had a high prioritisation score. The existing research regarding the function of this protein in HIV and TB mono-infection, coupled with it's prioritisation in the network made it stand out for further investigation. More apoptosis has been observed in *Mtb* infected cells in EIF2AK2 knockouts in mice (Wu et al., 2012b). Further more, during HIV-1 infection, HIV-1 protein Tat usually binds to EIF2AK2 increasing TNF-α production which in turn leads to increased viral replication and *Mtb* growth Imperiali et al. (2001); Li et al. (2005). We propose that if an EIF2AK2 inhibitor were a viable treatment method, it may reduce both *Mtb* presence and HIV-1 viral load.

In addition to identifying individual proteins of interest, the network enables the identification of pathways of interactions between the two pathogens via interactions with their human

host. The utility of the interaction network has been demonstrated by Koch et al. (2017). They used the interaction network to understand the biological significance of codons that they identified in three *Mtb* genes (celA2b, katG, and cyp138) that were evolving under directional selection influenced by co-infection with HIV-1. The network was used to find the shortest paths between the *Mtb* proteins and any HIV-1 proteins. They found that both CelA2b and KatG are in a pathway to HIV-1 protein that is bridged by the human protein, CD209 (DC-SIGN), an important receptor for HIV and *Mtb*.

The host-pathogen interaction network used by Koch et al. (2017) was the initial version presented in chapter 2. In chapter 4, this network was highly enriched with annotations and biological complexity, by integrating gene expression data and genetic variation data into the network. To our knowledge, this is the first time these kinds of data have been integrated with a host-pathogen interaction network. This resource may help explain observations from other studies by producing visualisations of various relationships between biological components. In addition, this methodology is reproducible and may be of benefit for analysing other pathogens, and co-morbid health conditions. The network is browsable as an interactive network and downloadable from NDEx at the following link: https://www.ndexbio.org/#/networkset/8d0f15ef-d936-11ec-b397-0ac135e8bacf?accesskey=d53b38beb93e7be80f374d3d8cebc626dcd5680a6767521ac9bac80e92c20e7a.

5.2 Limitations

While we have demonstrated that the host-pathogen PPIN developed in this work has been a useful tool for investigation, and that the integration of variation and gene expression data can extend this usefulness, we acknowledge that there are limitations tied to our study, as well as opportunities for further research. Since the host-pathogen PPIN was created, many more human-*Mtb* interactions have been identified that could be included in the network (Cao et al., 2019; Sun et al., 2018). Similarly, HIV-human interactions were filtered to reduce some of the potential noise from the vast database of associations, but some of the other associations may also be relevant and new associations have been added since this network was generated. If the integrated dataset created in this analysis were to continue to be a relevant resource, processes would have to be put in place to update the various datasets when new information is available. In addition, future work could be done to develop formatting instructions for input files so that researchers could submit their data for inclusion in the integrated dataset.

In addition, the variation dataset was restricted to human variants that were known to be clinically-associated with HIV or TB, common non-synonymous coding variants, and the novel variants from chromosome 6 that were identified in this analysis. The exclusion of other non-coding variants is a limitation as these variants may also be important. In addition, variants within the pathogen proteins could also be included, as these can drastically change the interactions with the host's immune system and drugs. Furthermore, information on drug target protein domains would be useful to include for predicting the impact of variants in these regions. Future work could include a more comprehensive variant data set in the integrated network.

While we were able to use the graph generated in chapter 3 successfully to identify variants, the lack of compatibility with existing reference genome implementations limits its utility as an alternative reference. For future work we suggest that variation identified using graph-based methods is incorporated as additional paths on the existing reference genome. Reference-free variant identification and graph-based reference implementations are an active area of development and many advances have been made since the work presented here was started and methods were chosen.

5.3 Directions for future research

The desire to understand the relationship between infectious diseases and their human host has motivated scientific research for centuries. In the backdrop of a global pandemic, as emerging pathogens continue to wreak havoc with public health care and our daily life, scientists are in a race to develop treatments and vaccinations. The network produced in this analysis could easily be extended to include host-pathogen interactions with other pathogens such as SARS-COV-2, the pathogen that causes COVID-19 disease. The risk of COVID-19 death has been shown to be higher in people infected with HIV regardless of treatment adherence. Similarly this risk was higher in people with current or previous tuberculosis infection and people with diabetes (Davies, 2020). Extending the network to include SARS-COV-2 host-pathogen interactions as well as gene expression information from studies of people with non-communicable diseases and infectious diseases may help improve the understanding of more complex interactions between various co-morbidities. This may be of particular importance in South Africa, which has a high prevalence of HIV, TB, non-communicable diseases such as diabetes, and, more recently, COVID-19.

The scope of this work extends beyond analysing HIV-TB co-infection. The vast availability of high throughput biological data can be leveraged to identify targets for therapeutic intervention. To benefit the most from the wide variety and vast availability of high throughput data, integrated datasets coupled with phenotypic observations could be leveraged for predictive analytics and to help explain findings related to specific genes or variants.

6. References

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249.

Aiken, C., Konner, J., Landau, N. R., Lenburg, M. E., and Trono, D. (1994). Nef induces CD4 endocytosis: requirement for a critical dileucine motif in the membrane-proximal CD4 cytoplasmic domain. *Cell*, 76(5):853–864.

Akhtar, L. N., Qin, H., Muldowney, M. T., Yanagisawa, L. L., Kutsch, O., Clements, J. E., and Benveniste, E. N. (2010). Suppressor of cytokine signaling 3 inhibits antiviral IFN-β signaling to enhance HIV-1 replication in macrophages. *The Journal of Immunology*, 185(4):2393–2404.

Ako-Adjei, D., Fu, W., Wallin, C., Katz, K. S., Song, G., Darji, D., Brister, J. R., Ptak, R. G., and Pruitt, K. D. (2014). HIV-1, human interaction database: current status and new features. *Nucleic Acids Research*, 43(D1):D566–D570.

Albini, A., Ferrini, S., Benelli, R., Sforzini, S., Giunciuglio, D., Aluigi, M. G., Proudfoot, A. E., Alouani, S., Wells, T. N., Mariani, G., et al. (1998). HIV-1 Tat protein mimicry of chemokines. *Proceedings of the National Academy of Sciences*, 95(22):13153–13158.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Andrews, Simon (2010). A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed: 2015-08-29.

Ansari, S. A., Safak, M., Gallia, G. L., Sawaya, B. E., Amini, S., and Khalili, K. (1999). Interaction of YB-1 with human immunodeficiency virus type 1 Tat and TAR RNA modulates viral promoter activity. *Journal of General Virology*, 80(10):2629–2638.

Arold, S., Franken, P., Strub, M.-P., Hoh, F., Benichou, S., Benarous, R., and Dumas, C. (1997). The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered T cell receptor signaling. *Structure*, 5(10):1361–1372.

Atwine, D., Bonnet, M., and Taburet, A.-M. (2018). Pharmacokinetics of efavirenz in patients on antituberculosis treatment in high human immunodeficiency virus and tuberculosis burden countries: a systematic review. *British Journal of Clinical Pharmacology*, 84(8):1641–1658.

Av-Gay, Y. and Everett, M. (2000). The eukaryotic-like Ser/Thr protein kinases of *Mycobacterium tuberculosis*. *Trends in Microbiology*, 8(5):238–244.

Ayling, M., Clark, M., and Leggett, R. (2019). New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics*, 21:1–11.

Badri, M., Ehrlich, R., Wood, R., Pulerwitz, T., and Maartens, G. (2001). Association between tuberculosis and HIV disease progression in a high tuberculosis prevalence area. *The International Journal of Tuberculosis and Lung Disease*, 5(3):225–232.

Baena, A. and Porcelli, S. A. (2009). Evasion and subversion of antigen presentation by *Mycobacterium tuberculosis*. *Tissue Antigens*, 74(3):189–204.

Barboric, M., Yik, J. H., Czudnochowski, N., Yang, Z., Chen, R., Contreras, X., Geyer, M., Peterlin, B. M., and Zhou, Q. (2007). Tat competes with HEXIM1 to increase the active pool of P-TEFb for HIV-1 transcription. *Nucleic Acids Research*, 35(6):2003–2012.

Barnes, P., Fong, S.-J., Brennan, P., Twomey, P., Mazumder, A., and Modlin, R. (1990). Local production of tumor necrosis factor and IFN-gamma in tuberculous pleuritis. *The Journal of Immunology*, 145(1):149–154.

Barreiro, L. B., Neyrolles, O., Babb, C. L., Tailleux, L., Quach, H., McElreavey, K., Van Helden, P. D., Hoal, E. G., Gicquel, B., and Quintana-Murci, L. (2006). Promoter variation in the DC-SIGN–encoding gene CD209 is associated with tuberculosis. *PLoS Medicine*, 3(2):e20.

Barreiro, L. B., Tailleux, L., Pai, A. A., Gicquel, B., Marioni, J. C., and Gilad, Y. (2012). Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection. *Proceedings of the National Academy of Sciences*, 109(4):1204–1209.

Barrero-Villar, M., Cabrero, J. R., Gordón-Alonso, M., Barroso-González, J., Álvarez-Losada, S., Muñoz-Fernández, M. Á., Sánchez-Madrid, F., and Valenzuela-Fernández, A. (2009). Moesin is required for HIV-1-induced CD4-CXCR4 interaction, F-actin redistribution, membrane fusion and viral infection in lymphocytes. *Journal of Cell Science*, 122(1):103–113.

Beste, D. J., Espasa, M., Bonde, B., Kierzek, A. M., Stewart, G. R., and McFadden, J. (2009). The genetic requirements for fast and slow growth in mycobacteria. *PLoS One*, 4(4):e5349.

Bigi, F., Gioffré, A., Klepp, L., de la Paz Santangelo, M., Alito, A., Caimi, K., Meikle, V., Zumárraga, M., Taboga, O., Romano, M. I., et al. (2004). The knockout of the lprG-Rv1410 operon produces strong attenuation of *Mycobacterium tuberculosis*. *Microbes and Infection*, 6(2):182–187.

Blagoveshchenskaya, A. D., Thomas, L., Feliciangeli, S. F., Hung, C.-H., and Thomas, G. (2002). HIV-1 Nef downregulates MHC-I by a PACS-1-and PI3K-regulated ARF6 endocytic pathway. *Cell*, 111(6):853–866.

Blanco-Melo, D., Venkatesh, S., and Bieniasz, P. D. (2012). Intrinsic cellular defenses against human immunodeficiency viruses. *Immunity*, 37(3):399–411.

Bocedi, A., Notaril, S., Narciso, P., Bolli, A., Fasano, M., and Ascenzi, P. (2004). Binding of Anti-HIV Drugs to Human Serum Albumin. *IUBMB life*, 56(10):609–614.

Boom, W. H., Canaday, D. H., Fulton, S. A., Gehring, A. J., Rojas, R. E., and Torres, M. (2003). Human immunity to M. tuberculosis: T cell subsets and antigen processing. *Tuberculosis*, 83(1):98–106.

Borgatti, S. P. and Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484.

Bossi, A. and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, 5(1):260.

Boulle, A., Van Cutsem, G., Cohen, K., Hilderbrand, K., Mathee, S., Abrahams, M., Goemaere, E., Coetzee, D., and Maartens, G. (2008). Outcomes of nevirapine-and efavirenz-based antiretroviral therapy when coadministered with rifampicin-based antitubercular therapy. *JAMA*, 300(5):530–539.

Brandt, D. Y., Aguiar, V. R., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data. *G3: Genes| Genomes| Genetics*, 5(5):931–941.

Brunet, A., Bonni, A., Zigmond, M. J., Lin, M. Z., Juo, P., Hu, L. S., Anderson, M. J., Arden, K. C., Blenis, J., and Greenberg, M. E. (1999). Akt promotes cell survival by phosphorylating and inhibiting a Forkhead transcription factor. *Cell*, 96(6):857–868.

Buchmann, R. and Hazelhurst, S. (2015). Genesis version 0.2.5.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012.

Campbell, M. C. and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*, 9:403–433.

Cao, T., Lyu, L., Jia, H., Wang, J., Du, F., Pan, L., Li, Z., Xing, A., Xiao, J., Ma, Y., et al. (2019). A Two-Way Proteome Microarray Strategy to Identify Novel *Mycobacterium tuberculosis*-Human Interactors. *Frontiers in Cellular and Infection Microbiology*, 9:65.

Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K., and O'brien, S. J. (1999). HLA and HIV-1: heterozygote advantage and B* 35-Cw* 04 disadvantage. *Science*, 283(5408):1748–1752.

Cavallaro, U., Mariotti, M., Wu, Z. H., Soria, M. R., and Maier, J. A. (1997). Fibronectin modulates endothelial response to HIV type 1 Tat. *AIDS Research and Human Retroviruses*, 13(15):1341–1348.

Chackerian, A. A., Alt, J. M., Perera, T. V., Dascher, C. C., and Behar, S. M. (2002). Dissemination of *Mycobacterium tuberculosis* is influenced by host factors and precedes the initiation of T-cell immunity. *Infection and Immunity*, 70(8):4501–4509.

Chaudhry, A., Verghese, D. A., Das, S. R., Jameel, S., George, A., Bal, V., Mayor, S., and Rath, S. (2009). HIV-1 Nef promotes endocytosis of cell surface MHC class II molecules via a constitutive pathway. *The Journal of Immunology*, 183(4):2415–2424.

Chen, J., Aronow, B. J., and Jegga, A. G. (2009). Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, 10(1):1–14.

Chen, L., Xie, Q.-w., and Nathan, C. (1998). Alkyl hydroperoxide reductase subunit C (AhpC) protects bacterial and human cells against reactive nitrogen intermediates. *Molecular Cell*, 1(6):795–805.

Chen, R., Morgan, A. A., Dudley, J., Deshpande, T., Li, L., Kodama, K., Chiang, A. P., and Butte, A. J. (2008). FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome Biology*, 9(12):1–15.

Chopra, P., Koduri, H., Singh, R., Koul, A., Ghildiyal, M., Sharma, K., Tyagi, A. K., and Singh, Y. (2004). Nucleoside diphosphate kinase of Mycobacterium tuberculosis acts as GTPase-activating protein for Rho-GTPases. *FEBS Letters*, 571(1):212–216.

Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Bardien, S., Botha, G., Chimusa, E. R., Christoffels, A., Gamieldien, J., Sefid-Dashti, M. J., et al. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nature Communications*, 8(1):2062.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R., et al. (2011). Modernizing reference genome assemblies. *PLoS Biology*, 9(7):e1001091.

Collette, Y., Dutartre, H., Benziane, A., Ramos-Morales, F., Benarous, R., Harris, M., and Olive, D. (1996). Physical and functional interaction of Nef with Lck HIV-1 Nef-induced T-cell signaling defects. *Journal of Biological Chemistry*, 271(11):6333–6341.

Consortium, . G. P. et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

Consortium, U. et al. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169.

Cooray, S. (2004). The pivotal role of phosphatidylinositol 3-kinase–Akt signal transduction in virus survival. *Journal of General Virology*, 85(5):1065–1076.

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1):D472–D477.

Dana, R. R., Eigsti, C., Holmes, K. L., and Leto, T. L. (2000). A regulatory role for ADP-ribosylation factor 6 (ARF6) in activation of the phagocyte NADPH oxidase. *Journal of Biological Chemistry*, 275(42):32566–32571.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., et al. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158.

Davies, M. (2020). HIV and risk of COVID-19 death: a population cohort study from the Western Cape Province, South Africa. *medRxiv : the preprint server for health sciences*.

Davis, F. P., Barkan, D. T., Eswar, N., McKerrow, J. H., and Sali, A. (2007). Host–pathogen protein interactions predicted by comparative modeling. *Protein Science*, 16(12):2585–2596.

De Marco, A., Dans, P. D., Knezevich, A., Maiuri, P., Pantano, S., and Marcello, A. (2010). Subcellular localization of the interaction between the human immunodeficiency virus transactivator Tat and the nucleosome assembly protein 1. *Amino Acids*, 38(5):1583–1593.

de Noronha, A. L., Bafica, A., Nogueira, L., Barral, A., and Barral-Netto, M. (2008). Lung granulomas from *Mycobacterium tuberculosis*/HIV-1 co-infected patients display decreased in situ TNF production. *Pathology-Research and Practice*, 204(3):155–161.

Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Smith, M. W., Allikmets, R., Goedert, J. J., Buchbinder, S. P., Vittinghoff, E., Gomperts, E., et al. (1996). Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. *Science*, 273(5283):1856–1862.

Deffur, A., Mulder, N. J., and Wilkinson, R. J. (2013). Co-infection with *Mycobacterium tuberculosis* and human immunodeficiency virus: an overview and motivation for systems approaches. *Pathogens and Disease*, 69(2):101–113.

Deffur, A., Wilkinson, R. J., Mayosi, B. M., and Mulder, N. M. (2018). ANIMA: Association network integration for multiscale analysis. *Wellcome Open Research*, 3:27.

Deregibus, M. C., Cantaluppi, V., Doublier, S., Brizzi, M. F., Deambrosis, I., Albini, A., and Camussi, G. (2002). HIV-1-Tat protein activates phosphatidylinositol 3-kinase/AKT-dependent survival pathways in Kaposi's sarcoma cells. *Journal of Biological Chemistry*, 277(28):25195–25202.

Diedrich, C. R. and Flynn, J. L. (2011). HIV-1/*Mycobacterium tuberculosis* coinfection immunology: how does HIV-1 exacerbate tuberculosis? *Infection and Immunity*, 79(4):1407–1417.

Diedrich, C. R., Mattila, J. T., Klein, E., Janssen, C., Phuah, J., Sturgeon, T. J., Montelaro, R. C., Lin, P. L., and Flynn, J. L. (2010). Reactivation of latent tuberculosis in cynomolgus macaques infected with SIV is associated with early peripheral T cell depletion and not virus load. *PLoS One*, 5(3):e9611.

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R., and McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. *Nature Genetics*, 47(6):682–688.

Doolittle, J. M. and Gomez, S. M. (2010). Structural similarity-based predictions of protein interactions between HIV-1 and *Homo sapiens*. *Virology Journal*, 7(1):82.

Drage, M. G., Tsai, H.-C., Pecora, N. D., Cheng, T.-Y., Arida, A. R., Shukla, S., Rojas, R. E., Seshadri, C., Moody, D. B., Boom, W. H., et al. (2010). *Mycobacterium tuberculosis* lipoprotein LprG (Rv1411c) binds triacylated glycolipid agonists of Toll-like receptor 2. *Nature Structural & Molecular Biology*, 17(99):1088–1095.

Drögemöller, B. I., Plummer, M., Korkie, L., Agenbag, G., Dunaiski, A., Niehaus, D., Koen, L., Gebhardt, S., Schneider, N., Olckers, A., et al. (2013). Characterization of the genetic variation present in CYP3A4 in three South African populations. *Frontiers in Genetics*, 4:17.

Du, Q., Wang, H., and Xie, J. (2011). Thiamin (vitamin B1) biosynthesis and regulation: a rich source of antimicrobial drug targets. *International Journal of Biological Sciences*, 7(1):41–52.

Dyer, M. D., Murali, T., and Sobral, B. W. (2007). Computational prediction of host-pathogen protein–protein interactions. *Bioinformatics*, 23(13):i159–i166.

Dziadek, J., Madiraju, M. V., Rutherford, S. A., Atkinson, M. A., and Rajagopalan, M. (2002). Physiological consequences associated with overproduction of *Mycobacterium tuberculosis* FtsZ in mycobacterial hosts. *Microbiology*, 148(4):961–971.

El Messaoudi, K., Thiry, L. F., Liesnard, C., Van Tieghem, N., Bollen, A., and Moguilevsky, N. (2000). A human milk factor susceptible to cathepsin D inhibitors enhances human immunodeficiency virus type 1 infectivity and allows virus entry into a mammary epithelial cell line. *Journal of Virology*, 74(2):1004–1007.

Erten, S., Bebek, G., Ewing, R. M., and Koyutürk, M. (2011). DADA: degree-aware algorithms for network-based disease gene prioritization. *BioData Mining*, 4(1):1–20.

Estorninho, M., Smith, H., Thole, J., Harders-Westerveen, J., Kierzek, A., Butler, R. E., Neyrolles, O., and Stewart, G. R. (2010). ClgR regulation of chaperone and protease systems is essential for *Mycobacterium tuberculosis* parasitism of the macrophage. *Microbiology*, 156(11):3445–3455.

Farrow, M. F. and Rubin, E. J. (2008). Function of a mycobacterial major facilitator superfamily pump requires a membrane-associated lipoprotein. *Journal of Bacteriology*, 190(5):1783–1791.

Foster, I. (2005). Globus toolkit version 4: Software for service-oriented systems. In *Network and Parallel Computing*, pages 2–13. Springer.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Fu, W., Sanders-Beer, B. E., Katz, K. S., Maglott, D. R., Pruitt, K. D., and Ptak, R. G. (2009). Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Research*, 37(suppl 1):D417–D422.

García-Expósito, L., Barroso-González, J., Puigdomènech, I., Machado, J.-D., Blanco, J., and Valenzuela-Fernández, A. (2011). HIV-1 requires Arf6-mediated membrane dynamics to efficiently enter and infect T lymphocytes. *Molecular Biology of the Cell*, 22(8):1148–1166.

Gehring, A. J., Dobos, K. M., Belisle, J. T., Harding, C. V., and Boom, W. H. (2004). *Mycobacterium tuberculosis* LprG (Rv1411c): a novel TLR-2 ligand that inhibits human macrophage class II MHC antigen processing. *The Journal of Immunology*, 173(4):2660–2668.

Geijtenbeek, T. B., Kwon, D. S., Torensma, R., van Vliet, S. J., van Duijnhoven, G. C., Middel, J., Cornelissen, I. L., Nottet, H. S., KewalRamani, V. N., Littman, D. R., et al. (2000). DC-SIGN, a dendritic cell–specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell*, 100(5):587–597.

Goldman, A. S. and Prabhakar, B. S. (1996). *Immunology overview*. University of Texas Medical Branch at Galveston, Galveston (TX).

Guha, D. and Ayyavoo, V. (2013). Innate immune evasion strategies by human immunodeficiency virus type 1. *ISRN AIDS*, 2013:Article ID 954806, 10 pages.

Guha, D., Nagilla, P., Redinger, C., Srinivasan, A., Schatten, G. P., and Ayyavoo, V. (2012). Neuronal apoptosis by HIV-1 Vpr: contribution of proinflammatory molecular networks from infected target cells. *Journal of Neuroinflammation*, 9(1):138.

Guney, E. and Oliva, B. (2012). Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization. *PLoS One*, 7(9):e43557.

Guo, H., Dong, J., Hu, S., Cai, X., Tang, G., Dou, J., Tian, M., He, F., Nie, Y., and Fan, D. (2015). Biased random walk model for the prioritization of drug resistance associated proteins. *Scientific Reports*, 5(1):1–14.

Gupta, A., Kaul, A., Tsolaki, A. G., Kishore, U., and Bhakta, S. (2012). *Mycobacterium tuberculosis*: immune evasion, latency and reactivation. *Immunobiology*, 217(3):363–374.

Gupta, P., Collins, K. B., Ratner, D., Watkins, S., Naus, G. J., Landers, D. V., and Patterson, B. K. (2002). Memory CD4+ T cells are the earliest detectable human immunodeficiency virus type 1 (HIV-1)-infected cells in the female genital mucosal tissue during HIV-1 transmission in an organ culture system. *Journal of Virology*, 76(19):9868–9876.

Gurer, C., Cimarelli, A., and Luban, J. (2002). Specific incorporation of heat shock protein 70 family members into primate lentiviral virions. *Journal of Virology*, 76(9):4666–4670.

Hadian, K., Vincendeau, M., Mäusbacher, N., Nagel, D., Hauck, S. M., Ueffing, M., Loyter, A., Werner, T., Wolff, H., and Brack-Werner, R. (2009). Identification of a heterogeneous nuclear ribonucleoprotein-recognition region in the HIV Rev protein. *Journal of Biological Chemistry*, 284(48):33384–33391.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA.

Hanekom, M., Van Der Spuy, G., Van Pittius, N. G., McEvoy, C., Ndabambi, S., Victor, T., Hoal, E., Van Helden, P. D., and Warren, R. M. (2007). Evidence that the spread of *Mycobacterium tuberculosis* strains with the Beijing genotype is human population dependent. *Journal of Clinical Microbiology*, 45(7):2263–2266.

Hazenberg, M. D., Hamann, D., Schuitemaker, H., and Miedema, F. (2000). T cell depletion in HIV-1 infection: how CD4+ T cells go out of stock. *Nature Immunology*, 1(4):285–289.

Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726.

Hingley-Wilson, S. M., Sambandamurthy, V. K., and Jacobs, W. R. (2003). Survival perspectives from the world's most successful pathogen, *Mycobacterium tuberculosis*. *Nature Immunology*, 4(10):949–955.

Hladik, F. and Doncel, G. F. (2010). Preventing mucosal HIV transmission with topical microbicides: challenges and opportunities. *Antiviral Research*, 88:S3–S9.

Hladik, F., Sakchalathorn, P., Ballweber, L., Lentz, G., Fialkow, M., Eschenbach, D., and McElrath, M. J. (2007). Initial events in establishing vaginal entry and infection by human immunodeficiency virus type-1. *Immunity*, 26(2):257–270.

Holmes, A. (1996). *In vitro* phosphorylation of human immunodeficiency virus type 1 Tat protein by protein kinase C: Evidence for the phosphorylation of amino acid residue serine-46. *Archives of Biochemistry and Biophysics*, 335(1):8–12.

Hoshino, Y., Nakata, K., Hoshino, S., Honda, Y., Doris, B. T., Shioda, T., Rom, W. N., and Weiden, M. (2002). Maximal HIV-1 replication in alveolar macrophages during tuberculosis requires both lymphocyte contact and cytokines. *The Journal of Experimental Medicine*, 195(4):495–505.

Hu, Y., Henderson, B., Lund, P. A., Tormay, P., Ahmed, M. T., Gurcha, S. S., Besra, G. S., and Coates, A. R. (2008). A *Mycobacterium tuberculosis* mutant lacking the groEL homologue cpn60. 1 is viable but fails to induce an inflammatory response in animal models of infection. *Infection and Immunity*, 76(4):1535–1546.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57.

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. A. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9):1–16.

Huo, T., Liu, W., Guo, Y., Yang, C., Lin, J., and Rao, Z. (2015). Prediction of host-pathogen protein interactions between *Mycobacterium tuberculosis* and *Homo sapiens* using sequence motifs. *BMC Bioinformatics*, 16(1):100.

Hwang, W., Cho, Y.-r., Zhang, A., and Ramanathan, M. (2006). Bridging centrality: identifying bridging nodes in scale-free networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 20–23.

Hwang, W., Kim, T., Ramanathan, M., and Zhang, A. (2008). Bridging centrality: graph mining from element level to group level. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 336–344. ACM.

Idury, R. M. and Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology*, 2(2):291–306.

Imperiali, F., Zaninoni, A., La Maestra, L., Tarsia, P., Blasi, F., and Barcellini, W. (2001). Increased *Mycobacterium tuberculosis* growth in HIV-1-infected human macrophages: role of tumour necrosis factor-α. *Clinical & Experimental Immunology*, 123(3):435–442.

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2):226–232.

Ishikawa, H., Sasaki, M., Noda, S., and Koga, Y. (1998). Apoptosis induction by the binding of the carboxyl terminus of human immunodeficiency virus type 1 gp160 to calmodulin. *Journal of Virology*, 72(8):6574–6580.

Iwasaki, A. (2012). Innate immune recognition of HIV-1. *Immunity*, 37(3):389–398.

Jäger, S., Kim, D. Y., Hultquist, J. F., Shindo, K., LaRue, R. S., Kwon, E., Li, M., Anderson, B. D., Yen, L., Stanley, D., et al. (2012). Vif hijacks CBF-[bgr] to degrade APOBEC3G and promote HIV-1 infection. *Nature*, 481(7381):371–375.

Jakobovits, A., Rosenthal, A., and Capon, D. (1990). Trans-activation of HIV-1 LTR-directed gene expression by tat requires protein kinase C. *The EMBO journal*, 9(4):1165.

Johnson, R. M., Little, J. R., and Storch, G. A. (2001). Kawasaki-like syndromes associated with human immunodeficiency virus infection. *Clinical Infectious Diseases*, 32(11):1628–1634.

Julg, B., Moodley, E. S., Qi, Y., Ramduth, D., Reddy, S., Mncube, Z., Gao, X., Goulder, P. J., Detels, R., Ndung'u, T., et al. (2011). Possession of HLA class II DRB1* 1303 associates with reduced viral loads in chronic HIV-1 clade C and B infection. *Journal of Infectious Diseases*, 203(6):803–809.

Kaforou, M., Wright, V. J., Oni, T., French, N., Anderson, S. T., Bangani, N., Banwell, C. M., Brent, A. J., Crampin, A. C., Dockrell, H. M., et al. (2013). Detection of tuberculosis in HIV-infected and-uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLoS Medicine*, 10(10):e1001538.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210.

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., et al. (2016). The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Research*, 45(D1):D840–D845.

Karim, Q. A., Karim, S. S. A., Frohlich, J. A., Grobler, A. C., Baxter, C., Mansoor, L. E., Kharsany, A. B., Sibeko, S., Mlisana, K. P., Omar, Z., et al. (2010). Effectiveness and safety of tenofovir gel, an antiretroviral microbicide, for the prevention of HIV infection in women. *Science*, 329(5996):1168–1174.

Kawai, T. and Akira, S. (2007). Signaling to NF-κB by Toll-like receptors. *Trends in Molecular Medicine*, 13(11):460–469.

Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., Koizumi, H., et al. (2009). Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, 458(7238):641–645.

Kedzierska, K., Crowe, S. M., Turville, S., and Cunningham, A. L. (2003). The influence of cytokines, chemokines and their receptors on HIV-1 replication in monocytes and macrophages. *Reviews in Medical Virology*, 13(1):39–56.

Kent, W. J. and Haussler, D. (2001). Assembly of the working draft of the human genome with GigAssembler. *Genome Research*, 11(9):1541–1548.

Khan, F. A., Minion, J., Pai, M., Royce, S., Burman, W., Harries, A. D., and Menzies, D. (2010). Treatment of active tuberculosis in HIV-coinfected patients: a systematic review and meta-analysis. *Clinical Infectious Diseases*, 50(9):1288–1299.

Kishimoto, N., Onitsuka, A., Kido, K., Takamune, N., Shoji, S., and Misumi, S. (2012). Glyceraldehyde 3-phosphate dehydrogenase negatively regulates human immunodeficiency virus type 1 infection. *Retrovirology*, 9(1):107.

Koch, A. S., Brites, D., Stucki, D., Evans, J. C., Seldon, R., Heekes, A., Mulder, N., Nicol, M., Oni, T., Mizrahi, V., et al. (2017). The influence of HIV on the evolution of Mycobacterium tuberculosis. *Molecular Biology and Evolution*, 34(7):1654–1668.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949–958.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073.

Kumawat, K., Pathak, S. K., Spetz, A.-L., Kundu, M., and Basu, J. (2010). Exogenous Nef is an inhibitor of Mycobacterium tuberculosis-induced tumor necrosis factor-α production and macrophage apoptosis. *Journal of Biological Chemistry*, 285(17):12629–12637.

Kwan, C. K. and Ernst, J. D. (2011). HIV and tuberculosis: a deadly human syndemic. *Clinical Microbiology Reviews*, 24(2):351–376.

Kwara, A., Lartey, M., Boamah, I., Rezk, N. L., Oliver-Commey, J., Kenu, E., Kashuba, A. D., and Court, M. H. (2009). Interindividual Variability in Pharmacokinetics of Generic Nucleoside Reverse Transcriptase Inhibitors in TB/HIV-Coinfected Ghanaian Patients: UGT2B7* 1c Is Associated With Faster Zidovudine Clearance and Glucuronidation. *The Journal of Clinical Pharmacology*, 49(9):1079–1090.

Laddach, A., Ng, J. C.-F., Chung, S. S., and Fraternali, F. (2018). Genetic variants and protein–protein interactions: a multidimensional network-centric view. *Current Opinion in Structural Biology*, 50:82–90.

Landi, A., Vermeire, J., Iannucci, V., Vanderstraeten, H., Naessens, E., Bentahir, M., and Verhasselt, B. (2014). Genome-wide shRNA screening identifies host factors involved in early endocytic events for HIV-1-induced CD4 down-regulation. *Retrovirology*, 11(1):1–12.

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., et al. (2020). ClinVar: improvements to accessing data. *Nucleic Acids Research*, 48(D1):D835–D844.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2015). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868.

Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., et al. (2013). DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Research*, 41(D1):D936–D941.

Larsen, J. E., Massol, R. H., Nieland, T. J., and Kirchhausen, T. (2004). HIV Nef-mediated major histocompatibility complex class I down-modulation is independent of Arf6 activity. *Molecular Biology of the Cell*, 15(1):323–331.

Lathe, W. C. and Bork, P. (2001). Evolution of tuf genes: ancient duplication, differential loss and gene conversion. *FEBS Letters*, 502(3):113–116.

Lawn, S. D., Myer, L., Edwards, D., Bekker, L.-G., and Wood, R. (2009). Short-term and long-term risk of tuberculosis associated with CD4 cell recovery during antiretroviral therapy in South Africa. *AIDS*, 23(13):1717.

Lee, M. J. and Park, J. H. (2009). Pathway analysis in HEK 293T cells overexpressing HIV-1 tat and nucleocapsid. *Journal of Microbiology and Biotechnology*, 19(10):1103–1108.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., et al. (2009). IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Research*, 37(suppl 1):D1006–D1012.

Lew, J. M., Kapopoulou, A., Jones, L. M., and Cole, S. T. (2011). TubercuList–10 years after. *Tuberculosis*, 91(1):1–7.

Li (2015). seqtk: Toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk. Accessed: 2015-08-29.

Li, B.-Q., Niu, B., Chen, L., Wei, Z.-J., Huang, T., Jiang, M., Lu, J., Zheng, M.-Y., Kong, X.-Y., and Cai, Y.-D. (2013). Identifying chemicals with potential therapy of HIV based on protein-protein and protein-chemical interaction network. *PLoS One*, 8(6):e65207.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, J. C., Lee, D. C., Cheung, B. K., and Lau, A. S. (2005). Mechanisms for HIV Tat upregulation of IL-10 and other cytokine expression: kinase signaling and PKR-mediated immune response. *FEBS Letters*, 579(14):3055–3062.

Li, Q. and Wang, K. (2017). InterVar: clinical interpretation of genetic variants by the 2015 ACMG-AMP guidelines. *The American Journal of Human Genetics*, 100(2):267–280.

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., and Fan, W. (2011). Comparison of the two major classes of assembly algorithms: overlap?layout?consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11(1):25–37.

Lin, P. L. and Flynn, J. L. (2010). Understanding latent tuberculosis: a moving target. *The Journal of Immunology*, 185(1):15–22.

Lin, S., Wang, X. M., Nadeau, P. E., and Mergia, A. (2010). HIV infection upregulates caveolin 1 expression to restrict virus production. *Journal of Virology*, 84(18):9487–9496.

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human mmtation*, 37(3):235–241.

Lodish, H. (2008). *Molecular Cell Biology*. W. H. Freeman.

Lombard, Z., Brune, A., Hoal, E., Babb, C., Van Helden, P., Epplen, J., and Bornman, L. (2006). HLA class II disease associations in southern Africa. *Tissue Antigens*, 67(2):97–110.

Ma, X., Lee, H., Wang, L., and Sun, F. (2007). CGI: a new approach for prioritizing genes by combining gene expression and protein–protein interaction data. *Bioinformatics*, 23(2):215–221.

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., and Scherer, S. W. (2013). The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(D1):D986–D992.

MacPherson, J. I., Dickerson, J. E., Pinney, J. W., and Robertson, D. L. (2010). Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems. *PLoS Computational Biology*, 6(7):e1000863.

Magoč, T. and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963.

Mahon, R. N., Sande, O. J., Rojas, R. E., Levine, A. D., Harding, C. V., and Boom, W. H. (2012). *Mycobacterium tuberculosis* ManLAM inhibits T-cell-receptor signaling by interference with ZAP-70, Lck and LAT phosphorylation. *Cellular Immunology*, 275(1):98–105.

Maiti, D., Bhattacharyya, A., and Basu, J. (2001). Lipoarabinomannan from *Mycobacterium tuberculosis* promotes macrophage survival by phosphorylating Bad through a phosphatidylinositol 3-kinase/Akt pathway. *Journal of Biological Chemistry*, 276(1):329–333.

Malik, Z. A., Thompson, C. R., Hashimi, S., Porter, B., Iyer, S. S., and Kusner, D. J. (2003). Cutting edge: *Mycobacterium tuberculosis* blocks Ca2+ signaling and phagosome maturation in human macrophages via specific inhibition of sphingosine kinase. *The Journal of Immunology*, 170(6):2811–2815.

Manosuthi, W., Sukasem, C., Lueangniyomkul, A., Mankatitham, W., Thongyen, S., Nilkamhang, S., Manosuthi, S., and Sungkanuparph, S. (2013). Impact of pharmacogenetic markers of CYP2B6, clinical factors, and drug-drug interaction on efavirenz concentrations in HIV/tuberculosis-coinfected patients. *Antimicrobial Agents and Chemotherapy*, 57(2):1019–1024.

Maroder, M., Scarpa, S., Screpanti, I., Stigliano, A., Meco, D., Vacca, A., Stuppia, L., Frati, L., Modesti, A., and Gulino, A. (1996). Human Immunodeficiency Virus Type 1tatProtein Modulates Fibronectin Expression in Thymic Epithelial Cells and Impairsin VitroThymocyte Development. *Cellular immunology*, 168(1):49–58.

Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B. E., Ghaffaari, A., Kersey, P., Kloosterman, W., Makinen, V., Novak, A., et al. (2016). Computational pan-genomics: status, promises and challenges. *BioRxiv*, page 043430.

Martin, M. P., Lederman, M. M., Hutcheson, H. B., Goedert, J. J., Nelson, G. W., Van Kooyk, Y., Detels, R., Buchbinder, S., Hoots, K., Vlahov, D., et al. (2004). Association of DC-SIGN promoter polymorphism with increased risk for parenteral, but not mucosal, acquisition of human immunodeficiency virus type 1 infection. *Journal of Virology*, 78(24):14053–14056.

Matsubara, M., Jing, T., Kawamura, K., Shimojo, N., Titani, K., Hashimoto, K., and Hayashi, N. (2005). Myristoyl moiety of HIV Nef is involved in regulation of the interaction with calmodulin in vivo. *Protein Science*, 14(2):494–503.

Mazandu, G. K. and Mulder, N. J. (2011). Generation and analysis of large-scale data-driven *Mycobacterium tuberculosis* functional networks for drug target identification. *Advances in Bioinformatics*, 2011:Article ID 801478, 14 pages.

McMichael, A. J., Borrow, P., Tomaras, G. D., Goonetilleke, N., and Haynes, B. F. (2010). The immune response during acute HIV-1 infection: clues for vaccine development. *Nature Reviews Immunology*, 10(1):11–23.

Mercier, S. K., Donaghy, H., Botting, R. A., Turville, S. G., Harman, A. N., Nasr, N., Ji, H., Kusebauch, U., Mendoza, L., Shteynberg, D., et al. (2013). The microvesicle component of HIV-1 inocula modulates dendritic cell infection and maturation and enhances adhesion to and activation of T lymphocytes. *PLoS Pathogens*, 9(10):e1003700.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338(6104):222–226.

Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., Gould, K., Mead, D., Drury, E., O'Brien, J., et al. (2015). Genome variation and meiotic recombination in Plasmodium falciparum: insights from deep sequencing of genetic crosses. *BioRxiv*, page 024182.

Milev, M. P., Ravichandran, M., Khan, M. F., Schriemer, D. C., and Mouland, A. J. (2012). Characterization of staufen1 ribonucleoproteins by mass spectrometry and biochemical analyses reveal the presence of diverse host proteins associated with human immunodeficiency virus type 1. *Frontiers in Microbiology*, 3(367):10–3389.

Möller, M., De Wit, E., and Hoal, E. G. (2010). Past, present and future directions in human genetic susceptibility to tuberculosis. *FEMS Immunology & Medical Microbiology*, 58(1):3–26.

Morio, T., Chatila, T., and Geha, R. S. (1997). HIV glycoprotein gp120 inhibits TCR-CD3-mediated activation of fyn and lck. *International Immunology*, 9(1):53–64.

Mulder, N., Mazandu, G., and Rapanoel, H. (2013). Using host-pathogen functional interactions for filtering potential drug targets in *Mycobacterium tuberculosis*. *Mycobacterial Diseases*, 3:2161–1068.

Mulder, N. J., Akinola, R. O., Mazandu, G. K., and Rapanoel, H. (2014). Using biological networks to improve our understanding of infectious diseases. *Computational and Structural Biotechnology Journal*, 11(18):1–10.

Nakata, K., Rom, W. N., Honda, Y., Condos, R., Kanegasaki, S., Cao, Y., and Weiden, M. (1997). *Mycobacterium tuberculosis* enhances human immunodeficiency virus-1 replication in the lung. *American Journal of Respiratory and Critical Care Medicine*, 155(3):996–1003.

National Department of Health (2015). National Consolidated Guidelines, For the prevention of mother-to-child transmission of HIV (PMTCT) and the management of HIV in children, adolescents and adults.

Navlakha, S. and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063.

Neil, S. J., Zang, T., and Bieniasz, P. D. (2008). Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. *Nature*, 451(7177):425–430.

Niederman, T., Garcia, J., Hastings, W. R., Luria, S., and Ratner, L. (1992). Human immunodeficiency virus type 1 Nef protein inhibits NF-kappa B induction in human T cells. *Journal of Virology*, 66(10):6213–6219.

Novak, A. M., Hickey, G., Garrison, E., Blum, S., Connelly, A., Dilthey, A., Eizenga, J., Elmohamed, M. S., Guthrie, S., Kahles, A., et al. (2017). Genome graphs. *BioRxiv*, page 101378.

O'Leary, S., O'Sullivan, M. P., and Keane, J. (2011). IL-10 blocks phagosome maturation in Mycobacterium tuberculosis–infected human macrophages. *American Journal of Respiratory Cell and Molecular Biology*, 45(1):172–180.

Ono, A., Ablan, S. D., Lockett, S. J., Nagashima, K., and Freed, E. O. (2004). Phosphatidylinositol (4, 5) bisphosphate regulates HIV-1 Gag targeting to the plasma membrane. *Proceedings of the National Academy of Sciences*, 101(41):14889–14894.

Onouchi, Y., Ozaki, K., Burns, J. C., Shimizu, C., Terai, M., Hamada, H., Honda, T., Suzuki, H., Suenaga, T., Takeuchi, T., et al. (2012). A genome-wide association study identifies three new risk loci for Kawasaki disease. *Nature Genetics*, 44(5):517–521.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E., et al. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine*, 5(3):28.

Oscanoa, J., Sivapalan, L., Gadaleta, E., Dayem Ullah, A. Z., Lemoine, N. R., and Chelala, C. (2020). SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Research*, 48(W1):W185–W192.

Ott, D. E., Coren, L. V., Kane, B. P., Busch, L. K., Johnson, D. G., Sowder, R., Chertova, E. N., Arthur, L. O., and Henderson, L. E. (1996). Cytoskeletal proteins inside human immunodeficiency virus type 1 virions. *Journal of Virology*, 70(11):7734–7743.

Pan, X., Rudolph, J. M., Abraham, L., Habermann, A., Haller, C., Krijnse-Locker, J., and Fackler, O. T. (2012). HIV-1 Nef compensates for disorganization of the immunological synapse by inducing trans-Golgi network–associated Lck signaling. *Blood*, 119(3):786–797.

Pasula, R., Wisniowski, P., and Martin II, W. J. (2002). Fibronectin facilitates *Mycobacterium tuberculosis* attachment to murine alveolar macrophages. *Infection and Immunity*, 70(3):1287–1292.

Patel, N. R., Zhu, J., Tachado, S. D., Zhang, J., Wan, Z., Saukkonen, J., and Koziel, H. (2007). HIV impairs TNF-α mediated macrophage apoptotic response to *Mycobacterium tuberculosis*. *The Journal of Immunology*, 179(10):6973–6980.

Paten, B., Novak, A., and Haussler, D. (2014). Mapping to a reference genome structure. *arXiv preprint arXiv:1404.5010*.

Percario, Z., Olivetta, E., Fiorucci, G., Mangino, G., Peretti, S., Romeo, G., Affabris, E., and Federico, M. (2003). Human immunodeficiency virus type 1 (HIV-1) Nef activates STAT3 in primary human monocyte/macrophages through the release of soluble factors: involvement of Nef domains interacting with the cell endocytotic machinery. *Journal of Leukocyte Biology*, 74(5):821–832.

Pérez-Núñez, D. and Martínez-Quiles, N. (2011). *Genetic factors that Influence HIV Infection: the role of the major histocompatibility complex system*. INTECH Open Access Publisher.

Petrochilos, D., Shojaie, A., Gennari, J., and Abernethy, N. (2013). Using random walks to identify cancer-associated modules in expression data. *BioData Mining*, 6(1):1–25.

Phillips, R. E., Rowland-Jones, S., Nixon, D. F., Gotch, F. M., Edwards, J. P., Ogunlesi, A. O., Elvin, J. G., Rothbard, J. A., Bangham, C. R., Rizza, C. R., et al. (1991). Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature*, 354(6353):453.

Pillich, R. T., Chen, J., Rynkov, V., Welker, D., and Pratt, D. (2017). NDEx: a community resource for sharing and publishing of biological networks. In *Protein Bioinformatics*, pages 271–301. Springer.

Pitarque, S., Herrmann, J.-L., Duteyrat, J.-L., Jackson, M., Stewart, G. R., Lecointe, F., Payre, B., Schwartz, O., Young, D. B., Marchal, G., et al. (2005). Deciphering the molecular bases of

Mycobacterium tuberculosis binding to the lectin DC-SIGN reveals an underestimated complexity. *Biochemical Journal*, 392(3):615–624.

Poyraz, Ö., Saxena, S., Schnell, R., Yogeeswari, P., Schneider, G., Sriram, D., et al. (2013). Discovery of novel inhibitors targeting the Mycobacterium tuberculosis O-acetylserine sulfhydrylase (CysK1) using virtual high-throughput screening. *Bioorganic & Medicinal Chemistry Letters*, 23(5):1182–1186.

Pratt, D., Chen, J., Welker, D., Rivas, R., Pillich, R., Rynkov, V., Ono, K., Miello, C., Hicks, L., Szalma, S., et al. (2015). NDEx, the network data exchange. *Cell Systems*, 1(4):302–305.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Pruess, M., Kersey, P., and Apweiler, R. (2005). The Integr8 project–a resource for genomic and proteomic data. *In Silico Biology*, 5(2):179–185.

Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49.

Puech, V., Guilhot, C., Perez, E., Tropis, M., Armitige, L. Y., Gicquel, B., and Daffé, M. (2002). Evidence for a partial redundancy of the fibronectin-binding proteins for the transfer of mycoloyl residues onto the cell wall arabinogalactan termini of *Mycobacterium tuberculosis*. *Molecular Microbiology*, 44(4):1109–1122.

Pugliese, A., Savarino, A., Cantamessa, C., and Torre, D. (1996). Influence of fibronectin on HIV-1 infection and capability of binding to platelets. *Cell Biochemistry and Function*, 14(4):291–296.

Qamra, R., Mande, S. C., Coates, A. R., and Henderson, B. (2005). The unusual chaperonins of *Mycobacterium tuberculosis*. *Tuberculosis*, 85(5):385–394.

Qidwai, T., Jamal, F., and Khan, M. (2012). DNA sequence variation and regulation of genes involved in pathogenesis of pulmonary tuberculosis. *Scandinavian Journal of Immunology*, 75(6):568–587.

Quaranta, M. G., Mattioli, B., Spadaro, F., Straface, E., Giordani, L., Ramoni, C., Malorni, W., and Viora, M. (2003). HIV-1 Nef triggers Vav-mediated signaling pathway leading to functional and morphological differentiation of dendritic cells. *The FASEB journal*, 17(14):2025–2036.

Quaranta, M. G., Tritarelli, E., Giordani, L., and Viora, M. (2002). HIV-1 Nef induces dendritic cell differentiation: a possible mechanism of uninfected CD4+ T cell activation. *Experimental Cell Research*, 275(2):243–254.

Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., Arsenijevic, V., Nadj, J., Ghose, K., Suciu, M. C., et al. (2019). Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, 51(2):354–362.

Ranjbar, S., Boshoff, H. I., Mulder, A., Siddiqi, N., Rubin, E. J., and Goldfeld, A. E. (2009). HIV-1 replication is differentially regulated by distinct clinical strains of *Mycobacterium tuberculosis*. *PLoS One*, 4(7):e6116.

Rao, V. S., Srinivas, K., Sujini, G., and Kumar, G. (2014). Protein-protein interaction detection: methods and analysis. *International Journal of Proteomics*, 2014:Article ID 147648, 12 pages.

Rapanoel, H. A., Mazandu, G. K., and Mulder, N. J. (2013). Predicting and analyzing interactions between *Mycobacterium tuberculosis* and its human host. *PLoS One*, 8(7):e67472.

Ribera, E., Azuaje, C., Lopez, R. M., Domingo, P., Curran, A., Feijoo, M., Pou, L., Sánchez, P., Sambeat, M. A., Colomer, J., et al. (2007). Pharmacokinetic interaction between rifampicin and the once-daily combination of saquinavir and low-dose ritonavir in HIV-infected patients with tuberculosis. *Journal of Antimicrobial Chemotherapy*, 59(4):690–697.

Rich, E., Torres, M., Sada, E., Finegan, C., Hamilton, B., and Toossi, Z. (1997). *Mycobacterium tuberculosis* (MTB)-stimulated production of nitric oxide by human alveolar macrophages and relationship of nitric oxide production to growth inhibition of MTB. *Tubercle and Lung Disease*, 78(5):247–255.

Roberts, J. D., Bebenek, K., and Kunkel, T. A. (1988). The accuracy of reverse transcriptase from HIV-1. *Science*, 242(4882):1171.

Rock, K. L. and Shen, L. (2005). Cross-presentation: underlying mechanisms and role in immune surveillance. *Immunological Reviews*, 207(1):166–183.

Roy, N. H., Lambelé, M., Chan, J., Symeonides, M., and Thali, M. (2014). Ezrin is a component of the HIV-1 virological presynapse and contributes to the inhibition of cell-cell fusion. *Journal of Virology*, 88(13):7645–7658.

Russell-Goldman, E., Xu, J., Wang, X., Chan, J., and Tufariello, J. M. (2008). A *Mycobacterium tuberculosis* Rpf double-knockout strain exhibits profound defects in reactivation from chronic tuberculosis and innate immunity phenotypes. *Infection and Immunity*, 76(9):4269–4281.

Rustagi, A. and Gale, M. (2014). Innate antiviral immune signaling, viral evasion and modulation by HIV-1. *Journal of Molecular Biology*, 426(6):1161–1177.

Saathoff, E., Pritsch, M., Geldmacher, C., Koehler, R. N., Maboko, L., Maganga, L., Geis, S., McCutchan, F. E., Kijak, G. H., Kim, J. H., et al. (2010). Viral and host factors associated with the HIV-1 viral load setpoint in adults from Mbeya Region, Tanzania. *Journal of Acquired Immune Deficiency Syndromes*, 54(3):324.

Sánchez, A., Espinosa, P., García, T., and Mancilla, R. (2012). The 19 kDa *Mycobacterium tuberculosis* lipoprotein (LpqH) induces macrophage apoptosis through extrinsic and intrinsic pathways: a role for the mitochondrial apoptosis-inducing factor. *Clinical and Developmental Immunology*, 2012:Article ID 950503.

Sanghvi, V. R. and Steel, L. F. (2011). The cellular TAR RNA binding protein, TRBP, promotes HIV-1 replication primarily by inhibiting the activation of double-stranded RNA-dependent kinase PKR. *Journal of Virology*, 85(23):12614–12621.

Schmieder, R. and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One*, 6(3):e17288–e17288.

Schmitt, E. K., Riwanto, M., Sambandamurthy, V., Roggo, S., Miault, C., Zwingelstein, C., Krastel, P., Noble, C., Beer, D., Rao, S. P., et al. (2011). The natural product cyclomarin kills Mycobacterium tuberculosis by targeting the ClpC1 subunit of the caseinolytic protease. *Angewandte Chemie International Edition*, 50(26):5889–5891.

Seto, S., Matsumoto, S., Ohta, I., Tsujimura, K., and Koide, Y. (2009). Dissection of Rab7 localization on Mycobacterium tuberculosis phagosome. *Biochemical and Biophysical Research Communications*, 387(2):272–277.

Shankar, E. M., Vignesh, R., EllegAard, R., Barathan, M., Chong, Y. K., Bador, M. K., Rukumani, D. V., Sabet, N. S., Kamarulzaman, A., Velu, V., et al. (2014). HIV–*Mycobacterium tuberculosis* co-infection: a 'danger-couple model' of disease pathogenesis. *Pathogens and Disease*, 70(2):110–118.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

Sheehy, A. M., Gaddis, N. C., and Malim, M. H. (2003). The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nature Medicine*, 9(11):1404–1407.

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics*, 51(1):30.

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

Shu, Q. and Nair, V. (2008). Inosine monophosphate dehydrogenase (IMPDH) as a target in drug discovery. *Medicinal Research Reviews*, 28(2):219–232.

Siepel, A., Pollard, K. S., and Haussler, D. (2006). New methods for detecting lineage-specific selection. In *Annual International Conference on Research in Computational Molecular Biology*, pages 190–205. Springer.

Silva Miranda, M., Breiman, A., Allain, S., Deknuydt, F., and Altare, F. (2012). The tuberculous granuloma: an unsuccessful host defence mechanism providing a safety shelter for the bacteria? *Clinical and Developmental Immunology*, 2012:Article ID 139127.

Simons Foundation (2014). Simons Genome Diversity Project Dataset. https://www.simonsfoundation.org/life-sciences/simons-genome-diversity-project-dataset/. Accessed: 2015-05-29.

Singh, V., Chandra, D., Srivastava, B. S., and Srivastava, R. (2011). Downregulation of Rv0189c, encoding a dihydroxyacid dehydratase, affects growth of *Mycobacterium tuberculosis* in vitro and in mice. *Microbiology*, 157(1):38–46.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M. H., Baldock, R., Barbiera, G., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, 43(W1):W589–W598.

Smith, M. W., Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Lomb, D. A., Goedert, J. J., O'Brien, T. R., Jacobson, L. P., Kaslow, R., et al. (1997). Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. *Science*, 277(5328):959–965.

Stacey, A. R., Norris, P. J., Qin, L., Haygreen, E. A., Taylor, E., Heitman, J., Lebedeva, M., DeCamp, A., Li, D., Grove, D., et al. (2009). 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*, 83(8):3719–3733.

Staden, R. (1980). A mew computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Research*, 8(16):3673–3694.

Statistics South Africa (2017). Mid-year population estimates. www.statssa.gov.za. Accessed: 2021-01-26.

Strasner, A. B., Natarajan, M., Doman, T., Key, D., August, A., and Henderson, A. J. (2008). The Src kinase Lck facilitates assembly of HIV-1 at the plasma membrane. *The Journal of Immunology*, 181(5):3706–3713.

Stremlau, M., Perron, M., Lee, M., Li, Y., Song, B., Javanbakht, H., Diaz-Griffero, F., Anderson, D. J., Sundquist, W. I., and Sodroski, J. (2006). Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5α restriction factor. *Proceedings of the National Academy of Sciences*, 103(14):5514–5519.

Stumptner-Cuvelette, P., Morchoisne, S., Dugast, M., Le Gall, S., Raposo, G., Schwartz, O., and Benaroch, P. (2001). HIV-1 Nef impairs MHC class II antigen presentation and surface expression. *Proceedings of the National Academy of Sciences*, 98(21):12144–12149.

Sun, J., Yang, L.-L., Chen, X., Kong, D.-X., and Liu, R. (2018). Integrating multifaceted information to predict *Mycobacterium tuberculosis*-human protein-protein interactions. *Journal of Proteome Research*, 17(11):3810–3823.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., et al. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, page gku1003.

Tachado, S. D., Li, X., Swan, K., Patel, N., and Koziel, H. (2008). Constitutive activation of phosphatidylinositol 3-kinase signaling pathway down-regulates TLR4-mediated tumor

necrosis factor-α release in alveolar macrophages from asymptomatic HIV-positive persons in vitro. *Journal of Biological Chemistry*, 283(48):33191–33198.

Tae, H., Karunasena, E., Bavarva, J. H., McIver, L. J., and Garner, H. R. (2014). Large scale comparison of non-human sequences in human sequencing data. *Genomics*, 104(6):453–458.

Tailleux, L., Schwartz, O., Herrmann, J.-L., Pivert, E., Jackson, M., Amara, A., Legres, L., Dreher, D., Nicod, L. P., Gluckman, J. C., et al. (2003). DC-SIGN is the major *Mycobacterium tuberculosis* receptor on human dendritic cells. *The Journal of Experimental Medicine*, 197(1):121–127.

Tastan, O., Qi, Y., Carbonell, J. G., and Klein-Seetharaman, J. (2009). Prediction of interactions between HIV-1 and human proteins by information integration. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 516. NIH Public Access.

TB Alliance (2020). Living with AIDS, dying of tuberculosis. https://www.tballiance.org/downloads/mediakit/Living%20with%20AIDS%20Dying%20of%20TB.pdf. Accessed: 2021-03-05.

Toor, J. S., Singh, S., Sharma, A., and Arora, S. K. (2014). *Mycobacterium tuberculosis* modulates the gene interactions to activate the HIV replication and faster disease progression in a co-infected host. *PLoS One*, 9(9):e106815.

Toossi, Z., Mayanja-Kizza, H., Hirsch, C., Edmonds, K., Spahlinger, T., Hom, D., Aung, H., Mugyenyi, P., Ellner, J., and Whalen, C. (2001). Impact of tuberculosis (TB) on HIV-1 activity in dually infected patients. *Clinical & Experimental Immunology*, 123(2):233–238.

Toossi, Z., Wu, M., Hirsch, C. S., Mayanja-Kizza, H., Baseke, J., Aung, H., Canaday, D. H., and Fujinaga, K. (2012). Activation of P-TEFb at sites of dual HIV/TB infection, and inhibition of MTB-induced HIV transcriptional activation by the inhibitor of CDK9, Indirubin-3-monoxime. *AIDS Research and Human Retroviruses*, 28(2):182–187.

Toschi, E., Bacigalupo, I., Strippoli, R., Chiozzini, C., Cereseto, A., Falchi, M., Nappi, F., Sgadari, C., Barillari, G., Mainiero, F., et al. (2006). HIV-1 Tat regulates endothelial cell cycle progression via activation of the Ras/ERK MAPK signaling pathway. *Molecular Biology of the Cell*, 17(4):1985–1994.

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36–46.

Turner, I. (2019). *Discovering Genetic Variation in Populations using Next Generation Sequencing and De Novo Assembly*. PhD thesis, University of Oxford.

Turner, I., Garimella, K. V., Iqbal, Z., and McVean, G. (2017). Integrating long-range connectivity information into de Bruijn graphs. *BioRxiv*, page 147777.

Urdahl, K. (2015). Understanding the Immune Response to M. tuberculosis. *Nature Education*, 8(3):6.

Usha, V., Gurcha, S. S., Lovering, A. L., Lloyd, A. J., Papaemmanouil, A., Reynolds, R. C., and Besra, G. S. (2011). Identification of novel diphenyl urea inhibitors of Mt-GuaB2 active against *Mycobacterium tuberculosis*. *Microbiology*, 157(1):290–299.

Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E. (2008). How correlated are network centrality measures? *Connections (Toronto, Ont.)*, 28(1):16.

Van Grol, J., Subauste, C., Andrade, R. M., Fujinaga, K., Nelson, J., and Subauste, C. S. (2010). HIV-1 inhibits autophagy in bystander macrophage/monocytic cells through Src-Akt and STAT3. *PLoS One*, 5(7):e11733.

Vannberg, F. O., Chapman, S. J., Khor, C. C., Tosh, K., Floyd, S., Jackson-Sillah, D., Crampin, A., Sichali, L., Bah, B., Gustafson, P., et al. (2008). CD209 genetic polymorphism and tuberculosis disease. *PLoS One*, 3(1):e1388.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6(1):e1000641.

Vardabasso, C., Manganaro, L., Lusic, M., Marcello, A., and Giacca, M. (2008). The histone chaperone protein Nucleosome Assembly Protein-1 (hNAP-1) binds HIV-1 Tat and promotes viral transcription. *Retrovirology*, 5(1):8.

Via, L. E., Deretic, D., Ulmer, R. J., Hibler, N. S., Huber, L. A., and Deretic, V. (1997). Arrest of mycobacterial phagosome maturation is caused by a block in vesicle fusion between stages controlled by rab5 and rab7. *Journal of Biological Chemistry*, 272(20):13326–13331.

Vidricaire, G. and Tremblay, M. J. (2005). Rab5 and Rab7, but not ARF6, govern the early events of HIV-1 infection in polarized human placental cells. *The Journal of Immunology*, 175(10):6517–6530.

Vieira, O. V., Botelho, R. J., and Grinstein, S. (2002). Phagosome maturation: aging gracefully. *Biochemical Journal*, 366(3):689–704.

Wang, J., Sönnerborg, A., Rane, A., Josephson, F., Lundgren, S., Ståhle, L., and Ingelman-Sundberg, M. (2006). Identification of a novel specific CYP2B6 allele in Africans causing impaired metabolism of the HIV drug efavirenz. *Pharmacogenetics and genomics*, 16(3):191–198.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164.

Wells, C. D., Cegielski, J. P., Nelson, L. J., Laserson, K. F., Holtz, T. H., Finlay, A., Castro, K. G., and Weyer, K. (2007). HIV infection and multidrug-resistant tuberculosis, the perfect storm. *Journal of Infectious Diseases*, 196(Supplement 1):S86–S107.

Whalen, C., Horsburgh, C. R., Hom, D., Lahart, C., Simberkoff, M., and Ellner, J. (1995). Accelerated course of human immunodeficiency virus infection after tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, 151(1):129–135.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, pages 80–83.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., and Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(suppl 1):D668–D672.

Wolf, A. J., Desvignes, L., Linas, B., Banaiee, N., Tamura, T., Takatsu, K., and Ernst, J. D. (2008). Initiation of the adaptive immune response to *Mycobacterium tuberculosis* depends on antigen production in the local lymph node, not the lungs. *The Journal of Experimental Medicine*, 205(1):105–115.

Wolf, A. J., Linas, B., Trevejo-Nunez, G. J., Kincaid, E., Tamura, T., Takatsu, K., and Ernst, J. D. (2007). *Mycobacterium tuberculosis* infects dendritic cells with high frequency and impairs their function in vivo. *The Journal of Immunology*, 179(4):2509–2519.

Wonderlich, E. R., Leonard, J. A., Kulpa, D. A., Leopold, K. E., Norman, J. M., and Collins, K. L. (2011). ADP ribosylation factor 1 activity is required to recruit AP-1 to the major histocompatibility complex class I (MHC-I) cytoplasmic tail and disrupt MHC-I trafficking in HIV-1-infected primary T cells. *Journal of Virology*, 85(23):12216–12226.

Wonkam, A. (2021). Sequence three million genomes across Africa. *Nature*, 590:209–211.

World Health Organisation (2015). Tuberculosis and HIV. http://www.who.int/hiv/topics/tb/en/. Accessed: 2021-03-05.

Wu, C., Zhu, J., and Zhang, X. (2012a). Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics*, 13(1):1–10.

Wu, K., Koo, J., Jiang, X., Chen, R., Cohen, S. N., and Nathan, C. (2012b). Improved control of tuberculosis and activation of macrophages in mice lacking protein kinase R. *PLoS One*, 7(2):e30512.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Research*, 28(1):289–291.

Yang, C.-S., Yuk, J.-M., and Jo, E.-K. (2009). The role of nitric oxide in mycobacterial infections. *Immune Network*, 9(2):46–52.

Yoneyama, M., Suhara, W., Fukuhara, Y., Fukuda, M., Nishida, E., and Fujita, T. (1998). Direct triggering of the type I interferon system by virus infection: activation of a transcription factor complex containing IRF-3 and CBP/p300. *The EMBO Journal*, 17(4):1087–1095.

Yu, B., Fonseca, D. P., O'Rourke, S. M., and Berman, P. W. (2010a). Protease cleavage sites in HIV-1 gp120 recognized by antigen processing enzymes are conserved and located at receptor binding sites. *Journal of Virology*, 84(3):1513–1526.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010b). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7):976–978.

Zar, J. H. (2005). *Spearman Rank Correlation*. John Wiley & Sons, Ltd.

Zhang, Y., Broser, M., and Rom, W. N. (1994). Activation of the interleukin 6 gene by Mycobacterium tuberculosis or lipopolysaccharide is mediated by nuclear factors NF-IL6 and NF-kappa B. *Proceedings of the National Academy of Sciences*, 91(6):2225–2229.

7. Appendices

Each of the tables in the appendices can be downloaded as excel files from the Open Science Framework at the following link:
https://osf.io/expdb/?view_only=dd996a99711d4deb85b78e04dc49af3b.

Table 7.1 Geographical distribution, sex, and sequence coverage of the individuals whose sequences where included

Data Source	Sample ID	Population	Country	Sex	Coverage
SGDP Panel B	HGDP00982	Mbuti	Congo	M	37
	HGDP01036	San	Namibia	M	38
	HGDP00936	Yoruba	Nigeria	M	39
	HGDP01286	Mandenka	Senegal	M	37
	DNK07	Dinka	Sudan	M	35
SGDP Panel A	HGDP0456	Mbuti	Congo	M	24
	HGDP01029	San	Namibia	M	33
	HGDP00927	Yoruba	Nigeria	M	32
	HGDP01284	Mandenka	Senegal	M	25
	DNK02	Dinka	Sudan	M	28
1000G	NA19238	Yoruba	Nigeria	F	19
	NA19239	Yoruba	Nigeria	M	25
	NA19240	Yoruba	Nigeria	F	33
	HG02922	Esan	Nigeria	F	NS*
	HG03052	Mende	Sierra Leone	F	NS
	NA19017	Luhya in Webuye	Kenya	F	NS
	HG02568	Gambian	The Gambia	M	NS
SAHGP	SAHGP010	Sotho	South Africa	M	45
	SAHGP012	Sotho	South Africa	M	49
	SAHGP013	Sotho	South Africa	M	48
	SAHGP014	Sotho	South Africa	M	46
	SAHGP015	Sotho	South Africa	M	50
	SAHGP016	Sotho	South Africa	M	43
	SAHGP017	Sotho	South Africa	M	45
	SAHGP020	Sotho	South Africa	M	51
	SAHGP021	Xhosa	South Africa	M	33
	SAHGP024	Xhosa	South Africa	M	48
	SAHGP025	Xhosa	South Africa	M	50
	SAHGP027	Xhosa	South Africa	M	47
	SAHGP029	Xhosa	South Africa	M	48
	SAHGP030	Xhosa	South Africa	M	44
	SAHGP032	Xhosa	South Africa	M	53
	SAHGP034	Xhosa	South Africa	M	48

Notes. 1000G VCFs are not per individual, but per population with pre-calculated allele frequencies, average coverage is below 10 for these sequences. *NS stands for coverage not specified; these samples were listed on the 1000G data portal as being high coverage whole genome sequences, but the coverage was not listed.

Table 7.2 Anti-TB drug-target interactions included in the analysis

Generic name	Target Name	Target UniProt ID	Target Organism	*Mtb* Orthologue	Target Action	Pharmacological Action	Target Type
Amikacin	30S ribosomal protein S12	P0A7S3	E-Coli	P9WH62	inhibitor	yes	target
Aminosalicylic Acid	Prostaglandin G/H synthase 2	P35354	Human		inhibitor	unknown	target
Aminosalicylic Acid	Inhibitor of nuclear factor κ-B kinase subunit α	O15111	Human		inhibitor	unknown	target
Aminosalicylic Acid	Arachidonate 5-lipoxygenase	P09917	Human		inhibitor	unknown	target
Aminosalicylic Acid	Group IIE secretory phospholipase A2	Q9NZK7	Human		inhibitor	unknown	target
Aminosalicylic Acid	Myeloperoxidase	P05164	Human		inhibitor	unknown	enzyme
Aminosalicylic Acid	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase	P64143	*Mtb*	P64143	unknown	unknown	target
Clofazimine	DNA		Human		intercalation	yes	target
Clofazimine	Multidrug resistance protein 1	P08183	Human		unknown	unknown	target
Clofazimine	Bile salt export pump	O95342	Human		unknown	unknown	target
Clofazimine	Cytochrome P450 3A4	P08684	Human		inhibitor/substrate	unknown	enzyme
Clofazimine	Bile salt export pump	O95342	Human		inhibitor	unknown	transporter
Clofazimine	Multidrug resistance protein 1	P08183	Human		inhibitor	unknown	transporter
Cycloserine	D-alanine–D-alanine ligase A	P0A6J8	E-Coli	P9WP30	inhibitor	yes	target
Cycloserine	Aromatic-L-amino-acid decarboxylase	P20711	Human		inhibitor	unknown	enzyme
Cycloserine	Proton-coupled amino acid transporter 2	Q495M3	Human		unknown	unknown	transporter
Cycloserine	Alanine racemase	Q9L888	M. avium	P9WQA8	inhibitor	yes	target
Ethambutol	Probable arabinosyltransferase C	P72059	*Mtb*	P72059	inhibitor	yes	target
Ethambutol	Probable arabinosyltransferase B	P72030	*Mtb*	P72030	inhibitor	yes	target
Ethambutol	Probable arabinosyltransferase A	P0A560	*Mtb*	P0A560	inhibitor	yes	target
Ethionamide	Catalase-peroxidase	Q08129	*Mtb*	Q08129	other/unknown	unknown	target
Ethionamide	Enoyl-[acyl-carrier-protein] reductase [NADH]	P0A5Y6	*Mtb*	P0A5Y6	adduct	yes	target
Imipenem	Penicillin-binding protein 2	P0AD65	E-Coli	Q8VKS5	inhibitor	yes	target
Imipenem	Penicillin-binding protein 1B	P02919	E-Coli	Q7D529	inhibitor	yes	target
Imipenem	Penicillin-binding protein 1A	P02918	E-Coli	Q7D529	inhibitor	yes	target
Imipenem	Beta-lactamase	P00807	Staph aureus	P9WKD2	other/unknown	yes	target
Isoniazid	Cytochrome P450 2C8	P10632	Human		unknown	unknown	target
Isoniazid	Cytochrome P450 1A2	P05177	Human		unknown	unknown	target
Isoniazid	Cytochrome P450 3A4	P08684	Human		unknown	unknown	target

Continued…

Generic name	Target Name	Target UniProt ID	Target Organism	*Mtb* Orthologue	Target Action	Pharmacological Action	Target Type
Isoniazid	Cytochrome P450 2C19	P33261	Human		unknown	unknown	target
Isoniazid	Cytochrome P450 2E1	P05181	Human		substrate/inhibitor/inducer	unknown	enzyme
Isoniazid	Cytochrome P450 1A2	P05177	Human		inhibitor	unknown	enzyme
Isoniazid	Cytochrome P450 2C9	P11712	Human		inhibitor	unknown	enzyme
Isoniazid	Cytochrome P450 2D6	P10635	Human		inhibitor	unknown	enzyme
Isoniazid	Cytochrome P450 2C19	P33261	Human		inhibitor	unknown	enzyme
Isoniazid	Cytochrome P450 3A4	P08684	Human		inhibitor	unknown	enzyme
Isoniazid	Arylamine N-acetyltransferase 2	P11245	Human		substrate	unknown	enzyme
Isoniazid	Cytochrome P450 2A6	P11509	Human		inhibitor	unknown	enzyme
Isoniazid	Cytochrome P450 2C8	P10632	Human		inhibitor	unknown	enzyme
Isoniazid	Catalase-peroxidase	Q08129	*Mtb*	Q08129	other/unknown	yes	target
Isoniazid	Enoyl-[acyl-carrier-protein] reductase [NADH]	P0A5Y6	*Mtb*	P0A5Y6	adduct	yes	target
Isoniazid	Dihydrofolate reductase	P0A546	*Mtb*	P0A546	unknown	unknown	target
Isoniazid	Arylamine N-acetyltransferase	P0A5L8	*Mtb*	P0A5L8	inhibitor	unknown	enzyme
Kanamycin	30S ribosomal protein S12	P0A7S3	E-Coli	P9WH62	inhibitor	yes	target
Kanamycin	Aminoglycoside 3′-phosphotransferase	P00551	E-Coli	L7N4U8	unknown	unknown	enzyme
Kanamycin	Aminoglycoside 2′-N-acetyltransferase	P0A5N0	*Mtb*	P0A5N0	unknown	unknown	enzyme
Levofloxacin	DNA gyrase subunit A	P43700	Haemophilus influenzae	P9WG46	inhibitor	yes	target
Levofloxacin	DNA topoisomerase 2-α	P11388	Human		inhibitor	unknown	target
Levofloxacin	Cytochrome P450 1A2	P05177	Human		inhibitor	unknown	enzyme
Levofloxacin	Cytochrome P450 3A4	P08684	Human		inhibitor	unknown	enzyme
Levofloxacin	Multidrug resistance protein 1	P08183	Human		substrate/inhibitor	unknown	transporter
Levofloxacin	Solute carrier family 22 member 6	Q4U2R8	Human		substrate	unknown	transporter
Levofloxacin	Solute carrier family 22 member 7	O15244	Human		inhibitor	unknown	transporter
Levofloxacin	Solute carrier family 22 member 4	Q9H015	Human		inhibitor	unknown	transporter
Linezolid	Amine oxidase [flavin-containing] A	P21397	Human		inhibitor	unknown	enzyme
Linezolid	Amine oxidase [flavin-containing] B	P27338	Human		inhibitor	unknown	enzyme
Moxifloxacin	DNA gyrase subunit A	P43700	Haemophilus influenzae	P9WG46	inhibitor	yes	target
Moxifloxacin	DNA topoisomerase 2-α	P11388	Human		inhibitor	unknown	target

Continued…

Generic name	Target Name	Target UniProt ID	Target Organism	*Mtb* Orthologue	Target Action	Pharmacological Action	Target Type
Ofloxacin	DNA gyrase subunit A	P43700	Haemophilus influenzae	P9WG46	inhibitor	yes	target
Ofloxacin	DNA topoisomerase 2-α	P11388	Human		inhibitor	unknown	target
Ofloxacin	Cytochrome P450 1A2	P05177	Human		inhibitor	unknown	enzyme
Ofloxacin	Solute carrier family 22 member 5	O76082	Human		inhibitor	unknown	transporter
Ofloxacin	Multidrug resistance-associated protein 1	P33527	Human		inhibitor	unknown	transporter
Ofloxacin	Solute carrier family 22 member 6	Q4U2R8	Human		inhibitor	unknown	transporter
Ofloxacin	Canalicular multispecific organic anion transporter 1	Q92887	Human		inhibitor	unknown	transporter
Ofloxacin	Solute carrier family 22 member 4	Q9H015	Human		inhibitor	unknown	transporter
Pyrazinamide	Xanthine dehydrogenase/oxidase	P47989	Human		substrate	unknown	enzyme
Pyrazinamide	Aldehyde oxidase	Q06278	Human		substrate	unknown	enzyme
Pyrazinamide	Cytochrome P450 1A2	P05177	Human		substrate	unknown	enzyme
Pyrazinamide	Cytochrome P450 3A4	P08684	Human		substrate	unknown	enzyme
Pyrazinamide	Fatty acid synthetase I (FASI)		*Mtb*		inhibitor	yes	target
Rifampicin	DNA-directed RNA polymerase subunit β	P0A8V2	E-Coli	P9WGY8	inhibitor	yes	target
Rifampicin	DNA-directed RNA polymerase subunit β'	P0A8T7	E-Coli	P9WGY6	inhibitor	yes	target
Rifampicin	Nuclear receptor subfamily 1 group I member 2	O75469	Human		agonist	unknown	target
Rifampicin	Solute carrier organic anion transporter family member 1B1	Q9Y6L6	Human		unkown	unknown	target
Rifampicin	Solute carrier organic anion transporter family member 1A2	P46721	Human		unkown	unknown	target
Rifampicin	Solute carrier organic anion transporter family member 1B3	Q9NPD5	Human		unkown	unknown	target
Rifampicin	Lanosterol 14-α demethylase	Q16850	Human		unkown	unknown	target
Rifampicin	Serum albumin	P02768	Human		unkown	unknown	target
Rifampicin	Solute carrier organic anion transporter family member 2B1	O94956	Human		unkown	unknown	target
Rifampicin	Cytochrome P450 3A4	P08684	Human		substrate/inducer	unknown	enzyme
Rifampicin	Cytochrome P450 1A2	P05177	Human		inducer	unknown	enzyme
Rifampicin	Cytochrome P450 2C8	P10632	Human		substrate/inhibitor/inducer	unknown	enzyme
Rifampicin	UDP-glucuronosyltransferase 1-1	P22309	Human		inducer	unknown	enzyme
Rifampicin	Cytochrome P450 2C9	P11712	Human		substrate/inducer	unknown	enzyme
Rifampicin	Cytochrome P450 2B6	P20813	Human		inducer	unknown	enzyme
Rifampicin	Cytochrome P450 2C19	P33261	Human		unknown	unknown	enzyme
Rifampicin	Cytochrome P450 2A6	P11509	Human		inhibitor	unknown	enzyme

Continued…

Generic name	Target Name	Target UniProt ID	Target Organism	*Mtb* Orthologue	Target Action	Pharmacological Action	Target Type
Rifampicin	Cytochrome P450 2E1	P05181	Human		substrate/inhibitor/inducer	unknown	enzyme
Rifampicin	Cytochrome P450 3A43	Q9HB55	Human		inducer	unknown	enzyme
Rifampicin	Cytochrome P450 3A5	P20815	Human		inducer	unknown	enzyme
Rifampicin	Cytochrome P450 3A7	P24462	Human		inducer	unknown	enzyme
Rifampicin	Cytochrome P450 4A11	Q02928	Human		inducer	unknown	enzyme
Rifampicin	Multidrug resistance protein 1	P08183	Human		substrate/inhibitor/inducer	unknown	transporter
Rifampicin	Solute carrier organic anion transporter family member 1B3	Q9NPD5	Human		subtrate/inhibitor	unknown	transporter
Rifampicin	Multidrug resistance-associated protein 1	P33527	Human		inhibitor	unknown	transporter
Rifampicin	Solute carrier organic anion transporter family member 2B1	O94956	Human		inhibitor	unknown	transporter
Rifampicin	Bile salt export pump	O95342	Human		inhibitor	unknown	transporter
Rifampicin	Solute carrier organic anion transporter family member 1A2	P46721	Human		inhibitor	unknown	transporter
Rifampicin	Solute carrier family 22 member 7	Q9Y694	Human		inhibitor	unknown	transporter
Rifampicin	Solute carrier organic anion transporter family member 1B1	Q9Y6L6	Human		inhibitor	unknown	transporter
Rifampicin	Multidrug resistance-associated protein 5	O15440	Human		inducer	unknown	transporter
Rifampicin	Canalicular multispecific organic anion transporter 1	Q92887	Human		inducer	unknown	transporter
Rifampicin	Canalicular multispecific organic anion transporter 2	O15438	Human		inducer	unknown	transporter
Streptomycin	30S ribosomal protein S12	P0A7S3	E-Coli	P9WH62	inhibitor	yes	target
Streptomycin	Protein-arginine deiminase type-4	Q9UM07	Human		unknown	unknown	target
Streptomycin	Solute carrier family 22 member 6	Q4U2R8	Human		inhibitor	unknown	transporter

Table 7.3 Anti-HIV drug-target interactions included in the analysis

Generic name	Target Name	Target UniProt ID	Target Organism	Target Action	Pharmacological Action	Target Type
Abacavir	Reverse transcriptase/RNaseH	Q72547	HIV-1	inhibitor	yes	target
Abacavir	Alcohol dehydrogenase 6	P28332	Human	substrate	unknown	enzyme
Abacavir	UDP-glucuronosyltransferase 1-1	P22309	Human	substrate	unknown	enzyme
Abacavir	Adenosine kinase	P55263	Human	substrate	unknown	enzyme
Atazanavir	HIV-1 Protease	Q72874	HIV-1	inhibitor	yes	target
Atazanavir	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor	unknown	enzyme
Atazanavir	Cytochrome P450 2C9	P11712	Human	substrate	unknown	enzyme
Atazanavir	Multidrug resistance protein 1	P08183	Human	substrate/inhibitor	unknown	transporter
Atazanavir	Multidrug resistance-associated protein 1	P33527	Human	substrate/inhibitor	unknown	transporter
Darunavir	HIV-1 Protease	Q72874	HIV-1	inhibitor	yes	target
Darunavir	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor	unknown	enzyme
Efavirenz	Reverse transcriptase/RNaseH	Q72547	HIV-1	inhibitor	yes	target
Efavirenz	Cytochrome P450 2B6	P20813	Human	substrate/inhibitor/inducer	unknown	enzyme
Efavirenz	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor/inducer	unknown	enzyme
Efavirenz	Cytochrome P450 2C19	P33261	Human	inhibitor	unknown	enzyme
Efavirenz	Cytochrome P450 2C9	P11712	Human	inhibitor	unknown	enzyme
Efavirenz	Cytochrome P450 1A2	P05177	Human	inhibitor	unknown	enzyme
Efavirenz	Cytochrome P450 2D6	P10635	Human	inhibitor	unknown	enzyme
Elvitegravir	Integrase	Q7ZJM1	HIV-1	inhibitor	yes	target
Emtricitabine	Reverse transcriptase/RNaseH	Q72547	HIV-1	inhibitor	yes	target
Emtricitabine	Deoxycytidine kinase	P27707	Human	substrate	unknown	enzyme
Enfuvirtide	Envelope glycoprotein	Q53I07	HIV-1	unknown	yes	target
Enfuvirtide	Cytochrome P450 2C19	P33261	Human	substrate	unknown	enzyme
Enfuvirtide	Cytochrome P450 2E1	P05181	Human	substrate	unknown	enzyme
Etravirine	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor	unknown	enzyme
Etravirine	Cytochrome P450 2C9	P11712	Human	substrate/inhibitor	unknown	enzyme
Etravirine	Cytochrome P450 2C19	P33261	Human	substrate/inhibitor	unknown	enzyme
Etravirine	Multidrug resistance protein 1	P08183	Human	inhibitor	unknown	transporter
Etravirine	Multidrug resistance protein 3	P21439	Human	inhibitor	unknown	transporter
Fosamprenavir	HIV-1 Protease	Q72874	HIV-1	inhibitor	yes	target

Continued...

Generic name	Target Name	Target UniProt ID	Target Organism	Target Action	Pharmacological Action	Target Type
Fosamprenavir	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor	unknown	enzyme
Lamivudine	Reverse transcriptase/RNaseH	Q72547	HIV-1	inhibitor	yes	target
Lamivudine	Deoxycytidine kinase	P27707	Human	substrate	unknown	target
Lamivudine	UMP-CMP kinase	P30085	Human	substrate	unknown	enzyme
Lamivudine	Phosphoglycerate kinase 1	P00558	Human	substrate	unknown	enzyme
Lamivudine	Nucleoside diphosphate kinase A	P15531	Human	substrate	unknown	enzyme
Lamivudine	Nucleoside diphosphate kinase B	P22392	Human	substrate	unknown	enzyme
Lamivudine	Choline-phosphate cytidylyltransferase A	P49585	Human	substrate	unknown	enzyme
Lamivudine	Ethanolamine-phosphate cytidylyltransferase	Q99447	Human	substrate	unknown	enzyme
Lamivudine	5′(3′)-deoxyribonucleotidase, cytosolic type	Q8TCD5	Human	substrate	unknown	enzyme
Lamivudine	Multidrug resistance-associated protein 1	P33527	Human	inhibitor	unknown	enzyme
Lamivudine	Solute carrier family 22 member 6	Q4U2R8	Human	substrate	unknown	transporter
Lamivudine	ATP-binding cassette sub-family G member 2	Q9UNQ0	Human	substrate	unknown	transporter
Lamivudine	Solute carrier family 22 member 1	O15245	Human	substrate	unknown	transporter
Lamivudine	Solute carrier family 22 member 2	O15244	Human	substrate	unknown	transporter
Lamivudine	Solute carrier family 22 member 3	O75751	Human	substrate	unknown	transporter
Lamivudine	Multidrug resistance protein 1	P08183	Human	substrate	unknown	transporter
Lamivudine	Multidrug resistance-associated protein 4	O15439	Human	substrate	unknown	transporter
Lamivudine	Canalicular multispecific organic anion transporter 2	O15438	Human	substrate	unknown	transporter
Lamivudine	Canalicular multispecific organic anion transporter 1	Q92887	Human	unknown	unknown	transporter
Lopinavir	HIV-1 Protease	Q72874	HIV-1	inhibitor	yes	target
Lopinavir	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor/inducer	unknown	enzyme
Lopinavir	Cytochrome P450 2D6	P10635	Human	inhibitor	unknown	enzyme
Lopinavir	Cytochrome P450 2C19	P33261	Human	inhibitor/inducer	unknown	enzyme
Lopinavir	Cytochrome P450 1A2	P05177	Human	inhibitor	unknown	enzyme
Lopinavir	Cytochrome P450 2B6	P20813	Human	inhibitor	unknown	enzyme
Lopinavir	Cytochrome P450 2C9	P11712	Human	inhibitor	unknown	enzyme
Lopinavir	Multidrug resistance protein 1	P08183	Human	inhibitor	unknown	transporter
Maraviroc	C-C chemokine receptor type 5	P51681	Human	antagonist	yes	target
Maraviroc	Cytochrome P450 3A4	P08684	Human	substrate	unknown	enzyme
Nevirapine	Reverse transcriptase/RNaseH	Q72547	HIV-1	inhibitor	yes	target
Nevirapine	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor/inducer	unknown	enzyme
Nevirapine	Cytochrome P450 2B6	P20813	Human	substrate/inhibitor	unknown	enzyme

Continued…

Generic name	Target Name	Target UniProt ID	Target Organism	Target Action	Pharmacological Action	Target Type
Nevirapine	Cytochrome P450 3A5	P20815	Human	substrate	unknown	enzyme
Nevirapine	Cytochrome P450 2C9	P11712	Human	substrate/inhibitor	unknown	enzyme
Nevirapine	Cytochrome P450 2A6	P11509	Human	substrate	unknown	enzyme
Nevirapine	Cytochrome P450 2D6	P10635	Human	substrate/inhibitor	unknown	enzyme
Nevirapine	Cytochrome P450 1A2	P05177	Human	inhibitor	unknown	enzyme
Raltegravir	Integrase	Q7ZJM1	HIV-1	inhibitor	yes	target
Raltegravir	UDP-glucuronosyltransferase 1-1	P22309	Human	substrate	unknown	enzyme
Rilpivirine	Reverse transcriptase/RNaseH	Q72547	HIV-1	inhibitor	yes	target
Rilpivirine	Nuclear receptor subfamily 1 group I member 2	O75469	Human	agonist	unknown	target
Rilpivirine	Cytochrome P450 2C19	P33261	Human	inhibitor	unknown	enzyme
Rilpivirine	Cytochrome P450 2B6	P20813	Human	inhibitor	unknown	enzyme
Rilpivirine	Cytochrome P450 1A2	P05177	Human	inducer	unknown	enzyme
Rilpivirine	Multidrug resistance protein 1	P08183	Human	inhibitor	unknown	transporter
Rilpivirine	ATP-binding cassette sub-family G member 2	Q9UNQ0	Human	inhibitor	unknown	transporter
Rilpivirine	Solute carrier organic anion transporter family member 1B1	Q9Y6L6	Human	inhibitor	unknown	transporter
Rilpivirine	Solute carrier organic anion transporter family member 1B3	Q9NPD5	Human	inhibitor	unknown	transporter
Ritonavir	HIV-1 Protease	Q72874	HIV-1	inhibitor	yes	target
Ritonavir	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor/inducer	unknown	enzyme
Ritonavir	Cytochrome P450 2D6	P10635	Human	inhibitor/substrate	unknown	enzyme
Ritonavir	Cytochrome P450 2C9	P11712	Human	inhibitor/inducer	unknown	enzyme
Ritonavir	Cytochrome P450 2C19	P33261	Human	inhibitor/inducer	unknown	enzyme
Ritonavir	Cytochrome P450 2B6	P20813	Human	substrate/inhibitor/inducer	unknown	enzyme
Ritonavir	Cytochrome P450 2C8	P10632	Human	inhibitor/inducer	unknown	enzyme
Ritonavir	Cytochrome P450 3A5	P20815	Human	substrate/inhibitor	unknown	enzyme
Ritonavir	Cytochrome P450 3A7	P20815	Human	substrate/inhibitor	unknown	enzyme
Ritonavir	Cytochrome P450 1A2	P05177	Human	substrate/inducer	unknown	enzyme
Ritonavir	Cytochrome P450 2E1	P05181	Human	inhibitor	unknown	enzyme
Ritonavir	Multidrug resistance protein 1	P08183	Human	substrate/inhibitor/inducer	unknown	transporter
Ritonavir	Multidrug resistance-associated protein 1	P33527	Human	inhibitor/inducer	unknown	transporter
Ritonavir	Canalicular multispecific organic anion transporter 1	Q92887	Human	substrate/inducer	unknown	transporter
Ritonavir	Solute carrier organic anion transporter family member 1A2	P46721	Human	inhibitor	unknown	transporter
Ritonavir	ATP-binding cassette sub-family G member 2	Q9UNQ0	Human	inhibitor	unknown	transporter
Ritonavir	Solute carrier organic anion transporter family member 1B1	Q9Y6L6	Human	inhibitor	unknown	transporter

Continued…

Generic name	Target Name	Target UniProt ID	Target Organism	Target Action	Pharmacological Action	Target Type
Saquinavir	HIV-1 Protease	Q72874	HIV-1	inhibitor	yes	target
Saquinavir	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor	unknown	enzyme
Saquinavir	Cytochrome P450 3A5	P20815	Human	substrate/inhibitor	unknown	enzyme
Saquinavir	Cytochrome P450 3A7	P20815	Human	substrate/inhibitor	unknown	enzyme
Saquinavir	Cholesterol side-chain cleavage enzyme, mitochondrial	P05108	Human	substrate	unknown	enzyme
Saquinavir	Cytochrome P450 2C19	P33261	Human	inhibitor	unknown	enzyme
Saquinavir	Cytochrome P450 2C8	P10632	Human	inhibitor	unknown	enzyme
Saquinavir	Cytochrome P450 2C9	P11712	Human	inhibitor	unknown	enzyme
Saquinavir	Cytochrome P450 2D6	P10635	Human	inhibitor/substrate	unknown	enzyme
Saquinavir	α-1-acid glycoprotein 1	P02763	Human	unknown	unknown	carrier
Saquinavir	Serum albumin	P02768	Human	unknown	unknown	carrier
Saquinavir	Multidrug resistance protein 1	P08183	Human	inhibitor	unknown	transporter
Saquinavir	Solute carrier family 22 member 1	O15245	Human	inhibitor	unknown	transporter
Saquinavir	Solute carrier organic anion transporter family member 1A2	P46721	Human	inhibitor	unknown	transporter
Saquinavir	ATP-binding cassette sub-family G member 2	Q9UNQ0	Human	inhibitor	unknown	transporter
Saquinavir	Solute carrier organic anion transporter family member 1B1	Q9Y6L6	Human	inhibitor	unknown	transporter
Saquinavir	Multidrug resistance-associated protein 1	P33527	Human	substrate/inhibitor	unknown	transporter
Saquinavir	Canalicular multispecific organic anion transporter 1	Q92887	Human	unknown	unknown	transporter
Tenofovir	Reverse transcriptase/RNaseH	Q72547	HIV-1	inhibitor	yes	target
Tenofovir	Cytochrome P450 1A2	P05177	Human	inhibitor	unknown	enzyme
Tenofovir	Adenylate kinase 2, mitochondrial	P54819	Human	substrate	unknown	enzyme
Tenofovir	Adenylate kinase 4, mitochondrial	P27144	Human	substrate	unknown	enzyme
Tenofovir	Nucleoside diphosphate kinase A	P15531	Human	substrate	unknown	enzyme
Tenofovir	Nucleoside diphosphate kinase B	P22392	Human	substrate	unknown	enzyme
Tenofovir	Solute carrier family 22 member 6	Q4U2R8	Human	substrate	unknown	transporter
Tenofovir	Solute carrier family 22 member 8	Q8TCC7	Human	substrate	unknown	transporter
Tenofovir	Multidrug resistance-associated protein 7	Q5T3U5	Human	substrate	unknown	transporter
Tenofovir	Multidrug resistance-associated protein 4	O15439	Human	substrate	unknown	transporter
Tenofovir	Canalicular multispecific organic anion transporter 1	Q92887	Human	substrate	unknown	transporter
Tipranavir	HIV-1 Protease	Q72874	HIV-1	inhibitor	yes	target
Tipranavir	Cytochrome P450 3A4	P08684	Human	substrate/inhibitor	unknown	enzyme
Tipranavir	Cytochrome P450 2D6	P10635	Human	substrate/inhibitor	unknown	enzyme
Tipranavir	Cytochrome P450 2C19	P33261	Human	substrate/inhibitor	unknown	enzyme

Continued...

Generic name	Target Name	Target UniProt ID	Target Organism	Target Action	Pharmacological Action	Target Type
Zidovudine	Reverse transcriptase/RNaseH	Q72547	HIV-1	inhibitor	yes	target
Zidovudine	Telomerase reverse transcriptase	O14746	Human	inhibitor	unknown	target
Zidovudine	Cytochrome P450 2A6	P11509	Human	substrate	unknown	enzyme
Zidovudine	Cytochrome P450 2C19	P33261	Human	substrate	unknown	enzyme
Zidovudine	Cytochrome P450 2C8	P10632	Human	substrate	unknown	enzyme
Zidovudine	Cytochrome P450 2C9	P11712	Human	substrate	unknown	enzyme
Zidovudine	Cytochrome P450 3A4	P08684	Human	substrate	unknown	enzyme
Zidovudine	Thymidine kinase, cytosolic	P04183	Human	substrate	unknown	enzyme
Zidovudine	UDP-glucuronosyltransferase 2B7	P16662	Human	substrate	unknown	enzyme
Zidovudine	Solute carrier family 22 member 2	O15244	Human	unknown	unknown	transporter
Zidovudine	Solute carrier family 22 member 6	Q4U2R8	Human	unknown	unknown	transporter
Zidovudine	Solute carrier family 22 member 7	Q9Y694	Human	unknown	unknown	transporter
Zidovudine	Solute carrier family 22 member 8	Q8TCC7	Human	unknown	unknown	transporter
Zidovudine	Solute carrier family 22 member 11	Q9NSA0	Human	unknown	unknown	transporter
Zidovudine	Sodium/nucleoside cotransporter 1	O00337	Human	unknown	unknown	transporter
Zidovudine	Equilibrative nucleoside transporter 2	Q14542	Human	unknown	unknown	transporter
Zidovudine	Multidrug resistance protein 1	P08183	Human	substrate	unknown	transporter
Zidovudine	Multidrug resistance-associated protein 4	O15439	Human	substrate	unknown	transporter
Zidovudine	Multidrug resistance-associated protein 5	O15440	Human	substrate	unknown	transporter
Zidovudine	ATP-binding cassette sub-family G member 2	Q9UNQ0	Human	substrate	unknown	transporter

Table 7.4 Read length and basic statistics for the sample sequences

Sequences	Chr 6			Unmapped		
Name	Reads	Read Length	GC (%)	Reads	Read Length	GC (%)
DNK02	4 697 290	R1:95, R2:101	39	57 276 273	R1:94-95, R2:100-101	39
DNK07	1 291 750	R1:100, R2:100	39	30 964 911	R1:100, R2:100	45
HGDP00927	5 248 739	R1:95, R2:101	39	19 103 905	R1:94-95, R2:100-101	39
HGDP00936	9 569 154	R1:100, R2:100	40	16 088 122	R1:100, R2:100	39
HGDP00982	10 163 030	R1:100, R2:100	40	33 026 884	R1:100, R2:100	40
HGDP01029	9 117 882	R1:100-101, R2:100-101	39	20 710 642	R1:94-95, R2:100-101	39
HGDP01036	7 102 919	R1:100, R2:100	40	36 190 070	R1:100, R2:100	40
HGDP01284	6 629 629	R1:94-95, R2:94-95	39	23 179 894	R1:94-95, R2:100-101	40
HGDP01286	5 963 261	R1:100, R2:100	40	30 371 591	R1:100, R2:100	40
HGDP0456	5 211 203	R1:94-95, R2:100-101	39	18 466 068	R1:94-95, R2:100-101	40
SAHGP010	39 167 920	R1:100, R2:100	39	38 677 793	R1:100, R2:100	42
SAHGP012	43 343 841	R1:100, R2:100	39	44 211 685	R1:100, R2:100	41.5
SAHGP013	41 834 592	R1:100, R2:100	39	40 977 331	R1:100, R2:100	41
SAHGP014	39 558 580	R1:100, R2:100	39	37 525 332	R1:100, R2:100	42
SAHGP015	43 949 258	R1:100, R2:100	39	44 654 978	R1:100, R2:100	41.5
SAHGP016	37 374 573	R1:100, R2:100	39	36 712 619	R1:100, R2:100	42
SAHGP017	39 351 558	R1:100, R2:100	39	37 908 488	R1:100, R2:100	42
SAHGP020	45 199 638	R1:100, R2:100	39	42 523 672	R1:100, R2:100	41
SAHGP021	28 950 691	R1:100, R2:100	39	26 041 431	R1:100, R2:100	41.5
SAHGP024	41 842 895	R1:100, R2:100	39	39 683 327	R1:100, R2:100	42
SAHGP025	43 956 224	R1:100, R2:100	39	42 684 004	R1:100, R2:100	42
SAHGP027	40 752 196	R1:100, R2:100	39	38 948 521	R1:100, R2:100	42
SAHGP029	41 656 383	R1:100, R2:100	39	39 573 245	R1:100, R2:100	42
SAHGP030	38 015 245	R1:100, R2:100	39	35 917 655	R1:100, R2:100	42
SAHGP032	45 674 675	R1:100, R2:100	39	40 108 014	R1:100, R2:100	41
SAHGP034	41 412 201	R1:100, R2:100	39	38 648 049	R1:100, R2:100	42

Table 7.5 Processing times for the multi-coloured de Bruijn graph

		AVG	STDEV	MIN	MAX	MED
MERGE PE READS	Combined pairs (%)	11.16	15.87	2.66	55.59	4.64
	Flash time (s)	2519.18	2174.02	906	13 459	2122
BUILD SAMPLE GRAPHS	Build memory (GB)	33.62	20.50	23.7	94.8	27.6
	kmers created	948 814 584.91	692 268 646.09	521 156 182	3 756 798 314	738 913 594
	Build time (s)	5398.67	5038.92	1726	21 098	3297
FILTER SUBGRAPH THAT CONTAINS KMERS FROM CHR6 READS	Subgraph memory (GB)	26.34	21.47	13.9	100	19.9
	kmers touched	828 886 462.70	737 681 757.76	403 874 821	3 756 798 314	556 225 759
	kmers (%)	0.83	0.10	0.72	1	0.83
	Subgraph time (s)	3561.76	1301.24	1738	6706	3088
CLEAN SAMPLE GRAPHS	Clean memory (GB)	21.96	19.50	10.9	100	14.8
	Cleaning threshold	9.27	0.88	7	11	9
	Kmers removed (%)	73.47	9.48	60.04	95.47	71.4
	Final kmers	161 684 574.21	10 676 165.98	129 103 975	173 180 697	165 753 596
	Clean time (S)	1537.48	1564.22	573	8168	995
MERGE SAMPLE GRAPHS	Merge Memory (GB)	71.5				
	Kmers in merged graph	228 601 888				
	Colours	33				
	Merge time (s)	14 180				

DISTANCE MATRIX	Distance matrix memory (GB)	70.7				
	Distance matrix time (s)	3264				
INFER EDGES	Infer edge memory (GB)	5.8				
	Modified kmers	359 750				
	Modified kmers (%)	0.16				
	Infer edge time (s)	844				
THREAD LINKS		AVG	STDEV	MIN	MAX	MED
	Memory (GB)	24	0	24	24	24
	Paths saved	31 544 334.15	19 034 019.51	10 492 048	99 798 377	30 165 350
	Thread time (s)	4376.39	1578.10	2141	9798	4188
	Cleaning coverage below	9.91	1.10	7	12	10
	Links after cleaning	1 509 611.10	476 188.89	576 120	2 197 422	1 647 883
	Clean time (s)	160.76	197.27	37	884	100
MERGE LINKS	Paths memory (GB)	2.2				
	Paths saved	22 191 474				
	Kmers	1 438 479				
	Merge time (s)	2990				
BUBBLE CALLER	Memory (GB)	11.5				
	Bubbles called	2 400 125				
	Serial bubbles dropped	4 700 752				
	Bubble call time (s)	2075				
5′ FLANK EXTRACTION	Process time	2515				
VCF MAPPING	5′ unmapped	15 169				
	5′ low MAPQ	0				
	3′ multi-hits	604				
	3′ not found	6207				
	flanks overlap too much	23 740				
	Bubbles mapped	2 354 405				
	Bubbles mapped (%)	98.10	2 354 405			
	3′ kmer exact match (%)	91.31	2 149 841			
	3′ alignment found	9.7	228 304			
	Ref allele too long	0.06	1433			
	Alt alleles per call	2.27	5 345 695			
	Alt alleles too long	0.01	523			
	Alt alleles match ref (%)	42.48	2 270 630			
	Alt alleles mapped (%)	57.51	3 074 542			
	ALTs printed	4 017 593				
	VCF mapping time	334				
POST PROCESS VARIANTS	Total	4 017 593				
	Split	0				
	Realigned	460 531				
	Skipped	0				
VCF ANNOTATE (COVERAGE AND GENOTYPING)	Memory (GB)	77				
	ALTs read	4 017 593				
	ALTs used	4 013 925				
	ALTs too long (>100bp)	3668				
	ALTs too dense (>8 within 51bp)	2 144 448				
	ALTs printed with coverage	1 873 145				

	Coverage Time (s)	18 941				
	Genotyped (%)	46.62				
	Ploidy two (%)	100				
	Genotype time (s)	47				
		AVG	STDEV	MIN	MAX	MED
	Kmer Coverages	20.61	4.49	11.5	26.2	22.3
	Read lengths	74.73	62.77	47	251	50
	Sequencing error rates	0.01	3.52×10^{-18}	0.01	0.01	0.01
		AVG	STDEV	MIN	MAX	MED
UNMAPPED READS TOUCHING THE GRAPH	Memory (GB)	3.25	0.23	2.6	3.4	3.3
	Reads loaded	37 788 216.81	9 540 576.64	20 069 428	50 010 036	40 142 336.5
	Reads printed	26 830 051.58	7 588 186.86	10 857 693	37 162 585	29 151 515.5
	Printed (%)	68.26	6.10	50.02	74.31	70.705
	Time (s)	1483.81	451.95	610	2309	1666.5
GFA FORMAT	Memory (GB)	7.3				
	Unitigs	5 268 284				
	Time (s)	264				

Table 7.6 List of properties stored against nodes and relationships in the Neo4j graph database

Property	Node type	Property type	Property description	Data type
Uniprot_id	Proteins	Node identifier	Uniprot protein identifier	text
gene_Names	Proteins	Node identifier	Gene names in Uniprot	text
organism	Proteins	Node identifier	Organism name (HIV-1, Human, MTB CDC1551, Anti-TB Drug, Anti-HIV Drug)	text
MHC	Human proteins	Node filter	Filter to identify MHC proteins (Non-MHC, MHC)	text
Interacts_with_pathogens	Human proteins	Node filter	Filter to identify proteins that interact with pathogens (HIV, MTB, HIV and MTB)	text
protein_names	Proteins	Node identifier	Protein names in Uniprot	text
GO_biological_process	Proteins	Gene ontology	List of GO terms and ids	text
GO_molecular_function	Proteins	Gene ontology	List of GO terms and ids	text
GO_cellular_component	Proteins	Gene ontology	List of GO terms and ids	text
Degree	Proteins	Network centrality measure	Degree	integer
Betweenness	Proteins	Network centrality measure	Betweenness	float
Closeness	Proteins	Network centrality measure	Closeness	float
Pathogenicity_bridging	Proteins	Network centrality measure	Pathogenicity bridging	float
Minimum_dist_to_MHC	Proteins	Centrality	Minimum distance to MHC	integer
Average_dist_to_MHC	Proteins	Centrality	Average distance to MHC	float
Minimum_dist_to_MTB	Proteins	Centrality	Minimum distance to MTB	integer
Average_dist_to_MTB	Proteins	Centrality	Average distance to MTB	float
Minimum_dist_to_HIV	Proteins	Centrality	Minimum distance to HIV	integer
Average_dist_to_HIV	Proteins	Centrality	Average distance to HIV	float
Minimum_dist_to_Drug	Proteins	Centrality	Minimum distance to Drug	integer
Average_dist_to_Drug	Proteins	Centrality	Average distance to Drug	float
Unified_Score	Protein interactions	Interaction score	Unified interaction score	float
interaction_details	All interactions	Interaction type	Describes interaction type, e.g. Human PPI, *Mtb*-Human PPI, HIV-1-Human PPI, *Mtb*-*Mtb* PPI, Drug-target, Variant of	text
human_STRING_neighbourhood	Human PPI	Interaction score	STRING neighbourhood score	float
human_STRING_fusion	Human PPI	Interaction score	STRING fusion score	float
human_STRING_cooccurence	Human PPI	Interaction score	STRING co-occurrence score	float
human_STRING_coexpression	Human PPI	Interaction score	STRING co-expression score	float
human_STRING_experimental	Human PPI	Interaction score	STRING experimental evidence score	float
human_STRING_database	Human PPI	Interaction score	STRING database score	float
human_STRING_textmining	Human PPI	Interaction score	STRING textmining score	float
human_BOSSI_TwoHybrid	Human PPI	Interaction score	Bossi and Lehner two hybrid score	float

Continued…

Property	Node type	Property type	Property description	Data type
human_BOSSI_Interolog	Human PPI	Interaction score	Bossi and Lehner interolog score	float
human_BOSSI_Coexpression	Human PPI	Interaction score	Bossi and Lehner co-expression score	float
human_STRING_combined	Human PPI	Interaction score	STRING combined score	float
human_BOSSI_combined	Human PPI	Interaction score	Bossi and Lehner combined score	float
human_REACTOME	Human PPI	Interaction score	Reactome score	float
hiv_NumberOfPublications	Human-HIV-1 PPI	Interaction evidence	Number of publications reporting the interaction	int
hiv_Regulatory	Human-HIV-1 PPI	Interaction evidence	Flag indicating if the interaction was labeled as regulatory (0/1)	binary
hiv_Gene_reg	Human-HIV-1 PPI	Interaction evidence	Flag indicating if the interaction was labeled as gene regulation (0/1)	binary
hiv_Protein_reg	Human-HIV-1 PPI	Interaction evidence	Flag indicating if the interaction was labeled as protein regulation (0/1)	binary
hiv_Physical	Human-HIV-1 PPI	Interaction evidence	Flag indicating if the interaction was labeled as physical (0/1)	binary
hiv_Binding	Human-HIV-1 PPI	Interaction evidence	Flag indicating if the interaction was labeled as binding (0/1)	binary
hiv_Degradation	Human-HIV-1 PPI	Interaction evidence	Flag indicating if the interaction was labeled as degradation (0/1)	binary
hiv_Modification	Human-HIV-1 PPI	Interaction evidence	Flag indicating if the interaction was labeled as modification (0/1)	binary
mtb_neighbourhood	*Mtb-Mtb* PPI	Interaction evidence	STRING neighbourhood score	float
mtb_genefusion	*Mtb-Mtb* PPI	Interaction evidence	STRING fusion score	float
mtb_cooccurrence	*Mtb-Mtb* PPI	Interaction evidence	STRING co-occurrence score	float
mtb_coexpression	*Mtb-Mtb* PPI	Interaction evidence	STRING co-expression score	float
mtb_experimental	*Mtb-Mtb* PPI	Interaction evidence	STRING experimental evidence score	float
mtb_database	*Mtb-Mtb* PPI	Interaction evidence	STRING database score	float
mtbhum_DDI	*Mtb*-Human PPI	Interaction evidence	Flag indicating if the interaction was identified via DDI by Rapanoel et al. (2013) (0/1)	binary
mtbhum_DIP	*Mtb*-Human PPI	Interaction evidence	Flag indicating if the interaction was identified via DIP by Rapanoel et al. (2013) (0/1)	binary
mtbhum_Huo	*Mtb*-Human PPI	Interaction evidence	Flag indicating if the interaction was identified by Huo et al. (2015) (0/1)	binary
mtbhum_Interolog	*Mtb*-Human PPI	Interaction evidence	Flag indicating if the interaction was identified via interologs by Rapanoel et al. (2013) (0/1)	binary
mtbhum_Literature	*Mtb*-Human PPI	Interaction evidence	Flag indicating if the interaction was described in the literature (0/1)	binary
drug_target_action	Drug-target interaction	Interaction evidence	Any combination of: substrate, inhibitor, and inducer	text
drug_target_pharmacology	Drug-target interaction	Interaction evidence	Whether or not pharmacology was listed in DrugBank (Yes/No)	text

Continued…

Property	Node type	Property type	Property description	Data type
drug_target_type	Drug-target interaction	Interaction evidence	Enzyme, transporter, carrier, target	text
DE_LTB_CO	Human proteins	Differential expression	Differentially expressed in HIV-TB co-infection vs. Latent TB infection (0/1)	binary
DE_LTB_HIV	Human proteins	Differential expression	Differentially expressed in HIV vs. Latent TB infection (0/1)	binary
DE_LTB_PTB	Human proteins	Differential expression	Differentially expressed in Pulmonary TB vs. Latent TB infection (0/1)	binary
DE_HIV_CO	Human proteins	Differential expression	Differentially expressed in HIV-TB co-infection vs. HIV infection (0/1)	binary
DE_PTB_CO	Human proteins	Differential expression	Differentially expressed in HIV-TB co-infection vs. Pulmonary TB infection (0/1)	binary
MEAN_CO	Human proteins	Differential expression	Average expression in HIV-TB co-infected individuals	float
SD_CO	Human proteins	Differential expression	Standard deviation of expression in HIV-TB co-infected individuals	float
MEAN_HIV	Human proteins	Differential expression	Average expression in HIV-infected individuals	float
SD_HIV	Human proteins	Differential expression	Standard deviation of expression in HIV-infected individuals	float
MEAN_PTB	Human proteins	Differential expression	Average expression in Pulmonary TB infected individuals	float
SD_PTB	Human proteins	Differential expression	Standard deviation of expression in Pulmonary TB infected individuals	float
MEAN_LTB	Human proteins	Differential expression	Average expression in Latent TB infected individuals	float
SD_LTB	Human proteins	Differential expression	Standard deviation of expression in Latent TB infected individuals	float
ADJ_PVAL_HIV_CO	Human proteins	Differential expression	Adjusted P-value for HIV vs. co-infection expression comparison	float
LOG_FC_HIV_CO	Human proteins	Differential expression	Log fold change in expression between HIV-TB co-infection and HIV infection	float
ADJ_PVAL_LTB_CO	Human proteins	Differential expression	Adjusted P-value for Latent TB vs. co-infection expression comparison	float
LOG_FC_LTB_CO	Human proteins	Differential expression	Log fold change in expression between HIV-TB co-infection and Latent TB infection	float
ADJ_PVAL_LTB_HIV	Human proteins	Differential expression	Adjusted P-value for Latent TB vs. HIV expression comparison	float
LOG_FC_LTB_HIV	Human proteins	Differential expression	Log fold change in expression between HIV and Latent TB infection	float
ADJ_PVAL_LTB_PTB	Human proteins	Differential expression	Adjusted P-value for Latent TB vs. Pulmonary TB expression comparison	float

Continued…

Property	Node type	Property type	Property description	Data type
LOG_FC_LTB_PTB	Human proteins	Differential expression	Log fold change in expression between Pulmonary TB and Latent TB infection	float
ADJ_PVAL_PTB_CO	Human proteins	Differential expression	Adjusted P-value for co-infection vs. Pulmonary TB expression comparison	float
LOG_FC_PTB_CO	Human proteins	Differential expression	Log fold change in expression between Pulmonary TB and HIV-TB co-infection	float
priority_score_random_walk	Proteins	Prioritisation score	Prioritisation score calculated using the Random walk with restart method	float
priority_score_page_rank	Proteins	Prioritisation score	Prioritisation score calculated using the Page rank with priors method	float
priority_score_network_propagation	Proteins	Prioritisation score	Prioritisation score calculated using the Network propagation method	float
priority_score_netCombo	Proteins	Prioritisation score	Prioritisation score calculated using the NetCombo method	float
variant_chromosome	Human Variants	Variant details	Variant chromosome	text
variant_chromosome_position	Human Variants	Variant details	Chromosome start position	int
variant_type	Human Variants	Variant details	Type of variant	text
reference_allele	Human Variants	Variant details	Reference allele	text
alternate_allele	Human Variants	Variant details	Alternate allele	text
variant_id	Human Variants	Variant details	db SNP rs identifier where available, else the identifier in Clinvar or chromosome and position for novel variants	text
gene	Human Variants	Variant details	The gene the variant maps to	text
gene_position	Human Variants	Variant details	The position in the gene	text
aa_change	Human Variants	Variant details	Amino acid change for non-synonymous SNPs	text
sift_score	Human Variants	Variant predicted effect score	SIFT score	float
polyphen_score	Human Variants	Variant predicted effect score	Polyphen score	float
minor_allele	Human Variants	Variant frequency	Minor allele	text
global_maf	Human Variants	Variant frequency	Global Minor allele frequency (reported in dbSNP)	text
1000Genome_AFR	Human Variants	Variant frequency	Minor allele frequency for African populations in 1000 Genomes	float
1000Genome_EAS	Human Variants	Variant frequency	Minor allele frequency for East Asian populations in 1000 Genomes	float
1000Genome_EUR	Human Variants	Variant frequency	Minor allele frequency for European populations in 1000 Genomes	float
gnomADgenome_AFR	Human Variants	Variant frequency	Minor allele frequency for African populations in gnomAD Genomes	float
gnomADgenome_EAS	Human Variants	Variant frequency	Minor allele frequency for East Asian populations in gnomAD Genomes	float

Continued…

Property	Node type	Property type	Property description	Data type
gnomADgenome_EUR	Human Variants	Variant frequency	Minor allele frequency for European populations (non-Finnish) in gnomAD Genomes	float
gnomADexome_AFR	Human Variants	Variant frequency	Minor allele frequency for African populations in gnomAD Exomes	float
gnomADexome_EAS	Human Variants	Variant frequency	Minor allele frequency for East Asian populations in gnomAD Exomes	float
gnomADexome_EUR	Human Variants	Variant frequency	Minor allele frequency for European populations (non-Finnish) in gnomAD Exomes	float
ClinVar_type	Human Variants	Clinical variant annotation	Clinvar variant classification type (e.g. protective, benign)	text
ClinVar_phenotype	Human Variants	Clinical variant annotation	Phenotypes associated with the SNP in ClinVar	text
Clinvar_HIVTB	Human Variants	Clinical variant annotation	Binary flag indicating whether the SNP was listed as clinically associated with HIV or TB in ClinVar	binary
common_african	Human Variants	Variant frequency	Binary flag indicating whether the SNP had a minor allele frequency of at least 0.01 in an African population	binary
afr_higher	Human Variants	Variant frequency	Binary flag indicating SNPs where Africans had a significantly higher minor allele frequency than another population	binary
afr_higher_populations	Human Variants	Variant frequency	List of populations where Africans had a significantly higher minor allele frequency	text

Table 7.7 MHC proteins and their expression values and prioritisation scores

Uniprot Id	Gene	Page rank Rank (score)	Random Walk Rank (score)	Network Propagation Rank (score)	NetCombo Rank (score)	Average expression during HIV-TB coinfection (Standard deviation)	Average expression during Latent TB infection (Standard deviation)	Log fold change	Adjusted *p*-value
P06681	C2	98.2 (0.031)	76.129 (6.23E-04)	69.125 (6.13E-04)	99.5 (0.523)	219.70 (134.89)	56.12 (36.08)	-1.92	<0.001*
Q9BW19	KIFC1	93.4 (0.019)	76.59 (4.38E-04)	69.55 (4.41E-04)	98.5 (0.424)	78.17 (79.26)	24.80 (29.37)	-1.37	<0.001*
P00751	CFB	92.6 (0.018)	76.62 (4.12E-04)	69.58 (4.15E-04)	98.2 (0.409)	92.43 (49.42)	40.89 (25.91)	-1.26	<0.001*
P62269	RPS18	96.5 (0.025)	76.120 (2.23E-04)	69.116 (2.15E-04)	95.0 (0.320)	5667.18 (2789.70)	8195.94 (3178.88)	0.60	<0.001*
Q03518	TAP1	93.4 (0.019)	76.108 (3.88E-04)	69.104 (3.87E-04)	97.6 (0.383)	12014.15 (4409.85)	5092.03 (1968.48)	-1.22	<0.001*
O14931	NCR3	90.7 (0.016)	76.13 (4.49E-04)	69.9 (4.57E-04)	98.9 (0.454)	167.21 (116.77)	397.55 (186.67)	1.43	<0.001*
P28065	PSMB9	91.7 (0.017)	76.31 (3.68E-04)	69.27 (3.71E-04)	98.1 (0.404)	576.99 (240.61)	265.57 (131.26)	-1.15	<0.001*
Q9Y333	LSM2	97.0 (0.026)	76.95 (1.20E-04)	69.91 (1.02E-04)	88.0 (0.253)	2756.77 (773.18)	3216.86 (634.44)	0.25	<0.001*
P10321	HLA-C	99.0 (0.038)	76.93 (1.76E-04)	69.89 (6.45E-05)	85.4 (0.233)	-	-	-	-
Q13838	DDX39B	94.8 (0.021)	76.7 (1.53E-04)	69.3 (1.23E-04)	89.8 (0.266)	9.19 (13.68)	5558.91 (1003.66)	0.32	<0.001*
P17693	HLA-G	93.4 (0.019)	76.10 (1.85E-04)	69.6 (1.66E-04)	91.2 (0.277)	5661.03 (2127.65)	4041.58 (1406.89)	-0.49	<0.001*
P16188	HLA-A	97.3 (0.027)	76.81 (9.81E-05)	69.77 (5.53E-05)	85.8 (0.236)	-	-	-	-
P07437	TUBB	94.8 (0.021)	76.36 (1.24E-04)	69.32 (1.04E-04)	88.5 (0.257)	301.66 (131.71)	4803.10 (865.77)	0.25	<0.001*
Q06643	LTB	86.2 (0.013)	76.26 (2.58E-04)	69.22 (2.64E-04)	96.4 (0.345)	2997.16 (1045.88)	4963.26 (1028.36)	0.78	<0.001*
Q31610	HLA-B	97.0 (0.026)	76.111 (9.84E-05)	69.107 (5.52E-05)	85.8 (0.236)	-	-	-	-
P30511	HLA-F	90.7 (0.016)	76.46 (1.97E-04)	69.42 (1.82E-04)	91.2 (0.277)	21337.43 (6280.90)	14530.03 (3887.79)	-0.54	<0.001*
Q03519	TAP2	88.2 (0.014)	76.51 (2.34E-04)	69.47 (2.33E-04)	93.0 (0.294)	1795.82 (592.03)	1055.63 (329.55)	-0.73	<0.001*
P67870	CSNK2B	89.6 (0.015)	76.88 (1.27E-04)	69.84 (1.19E-04)	90.6 (0.273)	3151.57 (889.91)	3835.62 (670.40)	0.32	<0.001*
O15533	TAPBP	88.2 (0.014)	76.116 (2.37E-04)	69.112 (2.37E-04)	93.2 (0.296)	314.08 (162.70)	122.05 (70.44)	-0.74	<0.001*
Q14676	MDC1	86.2 (0.013)	76.121 (1.78E-04)	69.117 (1.81E-04)	94.7 (0.316)	634.74 (262.37)	847.97 (262.71)	0.53	<0.001*
Q06587	RING1	88.2 (0.014)	76.8 (1.14E-04)	69.4 (1.08E-04)	90.8 (0.274)	3433.61 (948.15)	4046.08 (617.87)	0.28	<0.001*
P01374	LTA	84.1 (0.012)	76.139 (2.37E-04)	69.135 (2.40E-04)	95.5 (0.328)	426.95 (177.20)	661.55 (184.97)	0.70	<0.001*
P46379	BAG6	94.8 (0.021)	76.96 (1.09E-04)	69.92 (3.87E-05)	82.6 (0.220)	855.90 (515.09)	824.70 (396.46)	-0.02	0.95
Q92759	GTF2H4	90.7 (0.016)	76.20 (1.31E-04)	69.16 (1.26E-04)	88.2 (0.254)	336.97 (119.68)	418.79 (89.40)	0.36	<0.001*
P06340	HLA-DOA	84.1 (0.012)	76.71 (2.29E-04)	69.67 (2.30E-04)	93.7 (0.302)	438.38 (186.01)	679.14 (226.52)	0.70	<0.001*
P48634	PRRC2A	84.1 (0.012)	76.32 (2.28E-04)	69.28 (2.26E-04)	94.4 (0.311)	253.48 (115.01)	377.54 (111.15)	0.68	<0.001*
O95872	GPANK1	82.0 (0.011)	76.143 (3.24E-04)	69.139 (3.25E-04)	96.8 (0.356)	40.17 (36.30)	19.23 (23.43)	-1.04	<0.001*
Q96KQ7	EHMT2	95.3 (0.022)	76.29 (8.30E-05)	69.25 (3.15E-05)	83.0 (0.222)	84.77 (50.41)	123.73 (59.90)	0.52	0.04
Q29974	HLA-DRB1	95.8 (0.023)	76.65 (6.71E-05)	69.61 (4.07E-05)	81.7 (0.217)	-	-	-	-

Continued…

Uniprot Id	Gene	Page rank Rank (score)	Random Walk Rank (score)	Network Propagation Rank (score)	NetCombo Rank (score)	Average expression during HIV-TB coinfection (Standard deviation)	Average expression during Latent TB infection (Standard deviation)	Log fold change	Adjusted *p*-value
P01375	TNF	99.3 (0.042)	76.138 (1.73E-04)	69.134 (5.29E-05)	79.1 (0.212)	509.51 (183.88)	514.21 (156.44)	-0.01	0.97
P28062	PSMB8	86.2 (0.013)	76.107 (1.83E-04)	69.103 (1.74E-04)	91.9 (0.284)	3442.84 (691.57)	1673.71 (412.70)	-0.50	<0.001*
P13765	HLA-DOB	84.1 (0.012)	76.110 (2.22E-04)	69.106 (2.23E-04)	93.3 (0.298)	469.88 (189.64)	756.13 (287.86)	0.68	<0.001*
Q30154	HLA-DRB5	94.2 (0.020)	76.76 (6.04E-05)	69.72 (3.84E-05)	81.3 (0.216)	3076.91 (5264.90)	2336.46 (4032.63)	-0.43	0.86
O60231	DHX16	86.2 (0.013)	76.48 (7.40E-05)	69.44 (7.17E-05)	85.9 (0.237)	1403.38 (256.93)	1578.02 (252.80)	0.18	<0.001*
O95870	ABHD16A	79.7 (0.010)	76.142 (1.46E-04)	69.138 (1.43E-04)	90.9 (0.275)	1180.50 (336.48)	863.72 (155.23)	-0.42	<0.001*
Q5SSG8	MUC21	86.2 (0.013)	76.57 (3.18E-05)	69.53 (2.79E-05)	82.6 (0.220)	-9.81 (12.32)	-10.45 (13.72)	1.31	0.08
Q5ST30	VARS2	79.7 (0.010)	76.123 (1.55E-04)	69.119 (1.56E-04)	89.8 (0.266)	265.71 (111.54)	349.87 (127.66)	0.46	<0.001*
Q9Y676	MRPS18B	82.0 (0.011)	76.122 (1.27E-04)	69.118 (1.23E-04)	86.1 (0.238)	1669.70 (446.55)	2093.73 (373.57)	0.36	<0.001*
O15213	WDR46	86.2 (0.013)	76.135 (1.10E-04)	69.131 (1.05E-04)	81.7 (0.217)	418.18 (171.90)	490.38 (117.81)	0.27	<0.001*
Q9UBS5	GABBR1	89.6 (0.015)	76.64 (3.41E-05)	69.60 (2.07E-05)	76.5 (0.207)	16.09 (15.15)	20.29 (14.79)	0.24	0.46
Q99519	NEU1	77.1 (0.009)	76.9 (1.63E-04)	69.5 (1.65E-04)	87.8 (0.252)	806.11 (235.80)	577.58 (145.07)	-0.46	<0.001*
P26640	VARS	77.1 (0.009)	76.80 (1.38E-04)	69.76 (1.40E-04)	86.9 (0.244)	548.58 (246.08)	693.69 (184.70)	0.40	<0.001*
Q5SQ64	LY6G6F	71.3 (0.007)	76.35 (2.10E-04)	69.31 (2.11E-04)	93.1 (0.295)	202.54 (184.30)	109.82 (72.21)	-0.67	<0.001*
P0C0L5	C4B	84.1 (0.012)	76.37 (2.89E-05)	69.33 (1.72E-05)	79.7 (0.213)	-10.70 (11.30)	-12.26 (11.37)	-0.37	0.77
O43189	PHF1	74.5 (0.008)	76.112 (1.10E-04)	69.108 (1.13E-04)	89.4 (0.263)	57.78 (27.54)	384.74 (89.61)	0.32	<0.001*
P46695	IER3	71.3 (0.007)	76.21 (1.97E-04)	69.17 (1.98E-04)	91.2 (0.277)	-0.19 (16.81)	429.26 (169.66)	-0.63	<0.001*
P0C0L4	C4A	82.0 (0.011)	76.38 (2.45E-05)	69.34 (1.47E-05)	78.0 (0.210)	72.26 (41.90)	76.71 (48.64)	0.09	0.64
Q8NE71	ABCF1	74.5 (0.008)	76.89 (1.24E-04)	69.85 (1.26E-04)	84.2 (0.227)	623.97 (200.46)	788.90 (177.38)	0.37	<0.001*
O15212	PFDN6	74.5 (0.008)	76.92 (1.21E-04)	69.88 (1.21E-04)	83.7 (0.225)	82.30 (34.24)	63.65 (24.35)	-0.37	<0.001*
Q29980	MICB	71.3 (0.007)	76.52 (2.08E-04)	69.48 (2.08E-04)	87.6 (0.250)	1164.54 (372.65)	714.02 (179.17)	-0.67	<0.001*
Q9UER7	DAXX	82.0 (0.011)	76.69 (3.01E-05)	69.65 (2.59E-05)	75.5 (0.205)	538.97 (160.86)	604.06 (166.22)	0.17	0.02
P01909	HLA-DQA1	82.0 (0.011)	76.23 (2.52E-05)	69.19 (1.99E-05)	72.9 (0.199)	-	-	-	-
P01911	HLA-DRB1	82.0 (0.011)	76.90 (2.47E-05)	69.86 (1.99E-05)	71.3 (0.196)	-	-	-	-
P01903	HLA-DRA	82.0 (0.011)	76.25 (2.60E-05)	69.21 (2.01E-05)	71.7 (0.197)	-	-	-	-
P04440	HLA-DPB1	82.0 (0.011)	76.130 (2.59E-05)	69.126 (2.09E-05)	71.7 (0.197)	684.79 (416.05)	604.35 (309.50)	-0.09	0.78
Q9UBC1	NFKBIL1	88.2 (0.014)	76.27 (1.56E-04)	69.23 (5.24E-05)	64.6 (0.185)	-3.29 (15.37)	-10.35 (13.81)	-0.99	0.04
Q9GIY3	HLA-DRB1	79.7 (0.010)	76.84 (2.44E-05)	69.80 (2.00E-05)	71.7 (0.197)	-	-	-	-
P01906	HLA-DQA2	79.7 (0.010)	76.24 (2.45E-05)	69.20 (2.00E-05)	72.4 (0.198)	-19.18 (15.79)	-18.40 (16.88)	1.70	0.27
P20036	HLA-DPA1	79.7 (0.010)	76.109 (2.46E-05)	69.105 (2.00E-05)	72.4 (0.198)	12611.17 (5006.02)	11121.41 (2454.12)	-0.08	0.62

Continued…

Uniprot Id	Gene	Page rank Rank (score)	Random Walk Rank (score)	Network Propagation Rank (score)	NetCombo Rank (score)	Average expression during HIV-TB coinfection (Standard deviation)	Average expression during Latent TB infection (Standard deviation)	Log fold change	Adjusted *p*-value
P05538	HLA-DQB2	79.7 (0.010)	76.33 (2.45E-05)	69.29 (2.00E-05)	71.7 (0.197)	-	-	-	-
P04229	HLA-DRB1	79.7 (0.010)	76.19 (2.44E-05)	69.15 (2.00E-05)	71.7 (0.197)	-	-	-	-
O77932	DXO	63.6 (0.005)	76.12 (1.15E-04)	69.8 (1.19E-04)	86.3 (0.239)	197.79 (69.32)	247.43 (67.25)	0.35	<0.001*
Q99466	NOTCH4	82.0 (0.011)	76.113 (3.18E-05)	69.109 (1.81E-05)	68.2 (0.191)	18.28 (29.56)	23.73 (25.51)	-0.17	0.71
P55008	AIF1	63.6 (0.005)	76.150 (1.41E-04)	69.146 (1.44E-04)	83.0 (0.222)	2375.31 (1114.31)	1705.31 (715.20)	-0.45	<0.001*
Q9H633	RPP21	67.6 (0.006)	76.145 (1.13E-04)	69.141 (1.15E-04)	78.0 (0.210)	2616.93 (682.12)	3263.01 (778.80)	0.33	<0.001*
O00299	CLIC1	67.6 (0.006)	76.131 (1.63E-04)	69.127 (1.63E-04)	74.5 (0.203)	10474.72 (2745.34)	7255.83 (1777.44)	-0.52	<0.001*
Q8N1B4	VPS52	67.6 (0.006)	76.98 (1.20E-04)	69.94 (1.21E-04)	72.4 (0.198)	481.53 (123.86)	609.71 (124.61)	0.36	<0.001*
Q99943	AGPAT1	67.6 (0.006)	76.83 (1.13E-04)	69.79 (1.14E-04)	71.3 (0.196)	444.31 (177.52)	335.93 (95.91)	-0.35	<0.001*
O60888	CUTA	67.6 (0.006)	76.102 (1.66E-04)	69.98 (1.66E-04)	71.3 (0.196)	27.93 (42.36)	2200.07 (507.50)	0.53	<0.001*
Q29836	HLA-B	74.5 (0.008)	76.70 (1.03E-05)	69.66 (9.32E-06)	60.8 (0.179)	-	-	-	-
P13746	HLA-A	74.5 (0.008)	76.44 (1.03E-05)	69.40 (9.32E-06)	60.8 (0.179)	-	-	-	-
Q31612	HLA-B	74.5 (0.008)	76.11 (1.03E-05)	69.7 (9.32E-06)	60.8 (0.179)	-	-	-	-
Q01860	POU5F1	77.1 (0.009)	76.103 (2.32E-05)	69.99 (1.67E-05)	57.8 (0.174)	34.02 (19.73)	35.68 (22.63)	0.14	0.60
P30480	HLA-B	74.5 (0.008)	76.124 (1.03E-05)	69.120 (9.32E-06)	60.8 (0.179)	-	-	-	-
P04439	HLA-A	74.5 (0.008)	76.63 (1.07E-05)	69.59 (9.36E-06)	59.5 (0.177)	-	-	-	-
P01889	HLA-B	74.5 (0.008)	76.126 (1.05E-05)	69.122 (9.36E-06)	59.5 (0.177)	-	-	-	-
O43196	MSH5	54.4 (0.003)	76.97 (7.80E-06)	69.93 (1.17E-05)	75.5 (0.205)	-	-	-	-
Q08345	DDR1	63.6 (0.005)	76.54 (2.50E-05)	69.50 (1.80E-05)	66.3 (0.188)	261.58 (257.61)	316.41 (185.67)	0.47	0.00
P14373	TRIM27	54.4 (0.003)	76.79 (5.74E-06)	69.75 (7.49E-06)	72.9 (0.199)	331.75 (135.63)	329.21 (92.37)	0.05	0.77
Q9P1U0	ZNRD1	67.6 (0.006)	76.75 (1.70E-05)	69.71 (1.32E-05)	58.8 (0.176)	949.12 (176.29)	923.70 (178.95)	-0.05	0.55
Q15477	SKIV2L	63.6 (0.005)	76.85 (1.35E-05)	69.81 (1.63E-05)	62.2 (0.181)	985.13 (271.37)	1045.89 (272.55)	0.09	0.32
P25440	BRD2	48.7 (0.002)	76.22 (1.39E-05)	69.18 (1.79E-05)	78.0 (0.210)	5065.36 (1206.73)	5418.89 (801.47)	0.12	0.03
P18615	NELFE	71.3 (0.007)	76.14 (1.26E-05)	69.10 (1.15E-05)	54.5 (0.169)	353.75 (105.93)	310.43 (101.44)	-0.21	0.01
P40425	PBX2	77.1 (0.009)	76.6 (3.83E-05)	69.2 (1.82E-05)	48 (0.158)	1314.26 (369.20)	1447.23 (369.39)	0.15	0.05
P28067	HLA-DMA	59.2 (0.004)	76.106 (1.16E-05)	69.102 (1.23E-05)	66.3 (0.188)	5413.15 (2286.10)	4674.33 (1317.69)	-0.14	0.23
Q15109	AGER	59.2 (0.004)	76.68 (1.53E-05)	69.64 (1.44E-05)	64.0 (0.184)	25.05 (23.11)	172.65 (73.63)	0.32	0.33
P28068	HLA-DMB	59.2 (0.004)	76.105 (1.15E-05)	69.101 (1.23E-05)	64.6 (0.185)	7585.76 (2951.07)	7557.14 (1524.68)	0.08	0.55
Q9Y330	ZBTB12	40.7 (0.001)	76.94 (1.00E-05)	69.90 (2.95E-05)	81.3 (0.216)	-14.93 (23.95)	-18.55 (22.35)	0.59	0.56
O75955	FLOT1	48.7 (0.002)	76.115 (8.81E-06)	69.111 (1.39E-05)	70.7 (0.195)	3169.12 (1520.44)	2598.52 (826.12)	-0.20	0.10
Q96PV0	SYNGAP1	59.2 (0.004)	76.40 (9.65E-06)	69.36 (9.02E-06)	57.1 (0.173)	-21.11 (25.65)	-20.16 (22.23)	-0.94	0.52

Continued…

Uniprot Id	Gene	Page rank Rank (score)	Random Walk Rank (score)	Network Propagation Rank (score)	NetCombo Rank (score)	Average expression during HIV-TB coinfection (Standard deviation)	Average expression during Latent TB infection (Standard deviation)	Log fold change	Adjusted *p*-value
P28702	RXRB	59.2 (0.004)	76.137 (6.94E-06)	69.133 (8.58E-06)	55.2 (0.170)	1375.10 (289.17)	1515.82 (264.71)	0.15	0.00
P34931	HSPA1L	71.3 (0.007)	76.91 (1.98E-05)	69.87 (1.57E-05)	40.6 (0.143)	485.69 (183.71)	514.70 (124.93)	0.12	0.25
Q9UMR5	PPT2	40.7 (0.001)	76.149 (3.47E-06)	69.145 (8.79E-06)	71.3 (0.196)	43.65 (27.99)	-1.97 (15.31)	0.65	0.20
P59796	GPX6	63.6 (0.005)	76.58 (1.28E-05)	69.54 (1.52E-05)	47.0 (0.156)	10.47 (15.67)	24.88 (20.11)	0.49	0.09
O15211	RGL2	59.2 (0.004)	76.134 (3.86E-05)	69.130 (2.74E-05)	49.0 (0.160)	759.25 (256.57)	727.85 (194.19)	-0.02	0.88
P13942	COL11A2	54.4 (0.003)	76.28 (7.02E-06)	69.24 (8.15E-06)	52.1 (0.165)	-20.18 (11.67)	49.22 (24.67)	-1.46	0.36
O15205	UBD	48.7 (0.002)	76.72 (6.15E-06)	69.68 (1.12E-05)	56.5 (0.172)	5.74 (13.88)	3.23 (12.52)	-0.33	0.49
O95670	ATP6V1G2	59.2 (0.004)	76.34 (8.26E-06)	69.30 (9.15E-06)	46.4 (0.155)	45.74 (31.17)	55.02 (34.76)	0.48	0.04
Q92506	HSD17B8	67.6 (0.006)	76.15 (1.64E-04)	69.11 (1.64E-04)	37.8 (0.136)	409.03 (200.06)	564.95 (181.50)	0.52	<0.001*
P22105	TNXB	40.7 (0.001)	76.61 (4.23E-06)	69.57 (9.97E-06)	61.6 (0.180)	-11.44 (15.80)	-13.22 (14.02)	-0.05	0.97
Q99942	RNF5	54.4 (0.003)	76.82 (9.19E-06)	69.78 (1.01E-05)	46.4 (0.155)	312.27 (79.75)	26.93 (15.75)	-0.07	0.85
P36915	GNL1	59.2 (0.004)	76.47 (1.08E-05)	69.43 (1.21E-05)	41.4 (0.145)	442.73 (116.20)	505.08 (125.38)	0.20	0.00
Q9HCM9	TRIM39	48.7 (0.002)	76.141 (7.72E-06)	69.137 (9.64E-06)	50.8 (0.163)	782.28 (172.53)	68.95 (30.92)	0.16	0.56
P49842	STK19	40.7 (0.001)	76.56 (1.04E-06)	69.52 (2.85E-06)	54.5 (0.169)	422.36 (132.28)	1069.77 (161.40)	-0.07	0.57
A0A096LP39	TRIM39-RPP21	48.7 (0.002)	76.67 (1.02E-05)	69.63 (1.53E-05)	44.0 (0.150)	-	-	-	-
Q9UIM3	FKBPL	40.7 (0.001)	76.74 (1.30E-06)	69.70 (3.89E-06)	47.0 (0.156)	72.95 (44.96)	70.57 (22.93)	0.01	0.96
B4E1Z4	NONE	40.7 (0.001)	76.128 (1.35E-06)	69.124 (3.95E-06)	47.4 (0.157)	-	-	-	-
O75715	GPX5	40.7 (0.001)	76.43 (2.50E-06)	69.39 (5.71E-06)	42.4 (0.147)	2.60 (11.19)	25.89 (17.90)	0.39	0.49
Q99941	ATF6B	40.7 (0.001)	76.60 (3.79E-06)	69.56 (7.81E-06)	39.6 (0.141)	127.54 (53.42)	108.86 (39.19)	-0.14	0.49
Q92504	SLC39A7	59.2 (0.004)	76.16 (4.18E-05)	69.12 (1.89E-05)	21.6 (0.076)	107.08 (36.64)	9.74 (20.92)	-0.04	0.81
O60927	PPP1R11	40.7 (0.001)	76.45 (1.37E-06)	69.41 (2.39E-06)	35.6 (0.129)	4465.70 (1616.83)	3861.72 (1013.27)	-0.16	0.09
Q9BZY9	TRIM31	40.7 (0.001)	76.100 (9.57E-07)	69.96 (1.85E-06)	34.6 (0.126)	10.75 (16.09)	5.34 (19.84)	-0.76	0.10
O00453	LST1	48.7 (0.002)	76.118 (3.88E-06)	69.114 (4.10E-06)	25.8 (0.095)	3214.71 (887.30)	3536.95 (899.70)	0.16	0.03
Q53GD3	SLC44A4	40.7 (0.001)	76.133 (2.91E-06)	69.129 (5.13E-06)	33.3 (0.121)	4.09 (14.44)	3.04 (14.27)	0.03	0.51
Q8TD31	CCHCR1	40.7 (0.001)	76.49 (8.03E-07)	69.45 (1.61E-06)	30.9 (0.110)	38.61 (21.27)	35.39 (21.12)	-0.09	0.78
A0A024RCV8	MSH5	40.7 (0.001)	76.136 (1.03E-06)	69.132 (1.65E-06)	27.7 (0.103)	84.13 (42.15)	66.87 (25.34)	-0.30	0.01
O95873	C6orf47	0.0 (0.000)	76.144 (1.78E-06)	69.140 (7.67E-06)	63.4 (0.183)	812.84 (200.68)	691.08 (152.08)	-0.17	0.22
Q16653	MOG	0.0 (0.000)	76.117 (9.62E-07)	69.113 (5.19E-06)	52.8 (0.166)	20.39 (13.17)	17.44 (13.72)	-0.18	0.62
Q6NYC8	PPP1R18	0.0 (0.000)	76.114 (5.42E-07)	69.110 (2.63E-06)	44.0 (0.150)	9978.04 (2299.78)	9380.45 (1587.05)	-0.07	0.42
Q96QC0	PPP1R10	0.0 (0.000)	76.55 (5.97E-07)	69.51 (2.22E-06)	39.2 (0.140)	706.21 (202.15)	646.26 (145.84)	-0.11	0.16

Continued…

Uniprot Id	Gene	Page rank Rank (score)	Random Walk Rank (score)	Network Propagation Rank (score)	NetCombo Rank (score)	Average expression during HIV-TB coinfection (Standard deviation)	Average expression during Latent TB infection (Standard deviation)	Log fold change	Adjusted *p*-value
O95868	LY6G6D	0.0 (0.000)	76.132 (5.95E-07)	69.128 (1.65E-06)	38.8 (0.139)	-15.81 (43.61)	-35.62 (8.56)	NA	NA
O95866	G6B	0.0 (0.000)	76.77 (5.25E-07)	69.73 (1.03E-06)	32.9 (0.119)	-15.37 (13.79)	-10.45 (14.96)	-1.28	0.03
O95445	APOM	0.0 (0.000)	76.39 (3.89E-07)	69.35 (8.07E-07)	31.5 (0.113)	31.31 (34.74)	25.19 (20.80)	-0.18	0.64
Q9UIG4	PSORS1C2	0.0 (0.000)	76.140 (6.19E-07)	69.136 (1.34E-06)	31.3 (0.112)	18.12 (16.65)	22.95 (16.91)	0.37	0.22
Q9H2S5	RNF39	0.0 (0.000)	76.73 (3.81E-07)	69.69 (7.35E-07)	27.1 (0.100)	-18.51 (17.31)	-10.71 (27.93)	1.83	0.14
Q6UXA7	C6orf15	0.0 (0.000)	76.127 (3.61E-07)	69.123 (6.96E-07)	26.1 (0.096)	-6.60 (13.01)	-4.64 (13.07)	0.25	0.81
Q5SRN2	C6orf10	0.0 (0.000)	76.66 (4.88E-07)	69.62 (1.09E-06)	23.4 (0.084)	-40.51 (8.71)	-41.08 (10.01)	NA	NA
O15209	ZBTB22	0.0 (0.000)	76.104 (5.01E-07)	69.100 (1.24E-06)	22.9 (0.082)	1005.73 (205.80)	1008.36 (129.29)	0.02	0.80
Q12899	TRIM26	0.0 (0.000)	76.53 (3.48E-07)	69.49 (6.65E-07)	21.3 (0.074)	1352.02 (389.24)	1327.39 (274.45)	0.00	1.00
P08686	CYP21A2	0.0 (0.000)	76.30 (3.92E-07)	69.26 (6.50E-07)	21.1 (0.073)	11.82 (16.52)	5.01 (14.65)	-0.63	0.05
O95918	OR2H2	0.0 (0.000)	76.99 (3.57E-07)	69.95 (4.96E-07)	19.3 (0.059)	6.90 (26.08)	2.31 (18.43)	0.25	0.67
Q9UDY6	TRIM10	0.0 (0.000)	76.101 (3.49E-07)	69.97 (5.08E-07)	20.0 (0.064)	301.86 (418.44)	355.85 (400.49)	-0.52	0.04
Q9UGF7	OR12D3	0.0 (0.000)	76.87 (3.57E-07)	69.83 (4.96E-07)	19.3 (0.059)	-6.75 (19.26)	-6.62 (15.02)	0.37	0.74
Q9UGF6	OR5V1	0.0 (0.000)	76.86 (3.57E-07)	69.82 (4.96E-07)	19.3 (0.059)	-11.97 (13.38)	-16.39 (11.75)	-0.37	0.85
Q9UGF5	OR14J1	0.0 (0.000)	76.78 (3.57E-07)	69.74 (4.96E-07)	19.3 (0.059)	-37.07 (9.69)	-36.14 (10.23)	NA	NA
O76000	OR2B3	0.0 (0.000)	76.50 (3.57E-07)	69.46 (4.96E-07)	19.3 (0.059)	3.38 (14.58)	4.62 (12.48)	0.05	0.96
O76001	OR2J3	0.0 (0.000)	76.42 (3.57E-07)	69.38 (4.96E-07)	19.3 (0.059)	-15.54 (12.47)	-14.64 (10.46)	-1.35	0.11
O76002	OR2J2	0.0 (0.000)	76.41 (3.57E-07)	69.37 (4.96E-07)	19.3 (0.059)	9.79 (15.88)	12.22 (17.02)	0.26	0.46
Q9Y3N9	OR2W1	0.0 (0.000)	76.18 (3.57E-07)	69.14 (4.96E-07)	19.3 (0.059)	-17.09 (12.38)	-19.23 (10.74)	0.05	0.99
P58182	OR12D2	0.0 (0.000)	76.17 (3.57E-07)	69.13 (4.96E-07)	19.3 (0.059)	14.99 (16.97)	9.01 (15.61)	-0.41	0.29
Q9GZK4	OR2H1	0.0 (0.000)	76.148 (3.57E-07)	69.144 (4.96E-07)	19.3 (0.059)	-35.41 (9.05)	-37.04 (8.57)	NA	NA
Q9GZK7	OR11A1	0.0 (0.000)	76.147 (3.57E-07)	69.143 (4.96E-07)	19.3 (0.059)	55.62 (59.73)	1.35 (13.88)	0.50	0.02
Q9GZK6	OR2J1	0.0 (0.000)	76.146 (3.57E-07)	69.142 (4.96E-07)	19.3 (0.059)	-	-	-	-
Q96KK4	OR10C1	0.0 (0.000)	76.119 (3.57E-07)	69.115 (4.96E-07)	19.3 (0.059)	5.31 (17.35)	5.47 (14.35)	-0.09	0.88
Q5SQI0	ATAT1	0.0 (0.000)	76.125 (4.50E-07)	69.121 (5.39E-07)	18.2 (0.034)	-14.46 (11.46)	-14.20 (11.58)	0.22	0.11

Table 7.8 Description of interactions with and variants found in MHC proteins that have novel variants

Uniprot Id	Gene	DE during co-infection	HIV genes	Bridge Proteins	Drug targets	Total Variants	Novel variants	Higher frequency in African populations	Common in African populations	Deleterious	Average Prioritisation Percentile Rank
Q9BW19	KIFC1	Yes	-	-	-	19	6	7	1	4	97.13
P30511	HLA-F	Yes	Nef	-	-	12	2	3	2	6	90.93
P13765	HLA-DOB	Yes	Nef	CTSD	-	9	1	3	3	5	90.58
Q92759	GTF2H4	Yes	Tat	-	-	5	1	1	0	3	87.43
Q5ST30	VARS2	Yes	-	EPRS	-	27	3	11	10	6	86.7
Q30154	HLA-DRB5	No	Env, Gag, Nef	CTSD, LCK, PRKCQ, RAB7A	-	132	63	43	52	20	84
Q5SSG8	MUC21	No	-	-	-	68	2	33	30	9	80.35
Q9UER7	DAXX	No	-	MAP3K5	-	10	6	2	1	1	77.03
P01903	HLA-DRA	-	Env, Gag, Nef	CTSD, LCK, PRKCQ	-	39	39	0	0	0	74.73
P01909	HLA-DQA1	-	Nef	CTSD, LCK, PRKCQ	-	114	53	26	44	23	74.58
P01906	HLA-DQA2	No	Nef	CTSD, LCK, PRKCQ	-	25	11	3	5	5	74.15
P05538	HLA-DQB2	-	Nef	CTSD, LCK, PRKCQ	-	10	1	5	3	5	73.93
P04229	HLA-DRB1	-	Env, Gag, Nef	CTSD, FYN, LCK, PRKCQ, RAB7A	-	116	116	0	0	0	73.9
Q99466	NOTCH4	No	-	-	-	37	2	16	10	16	73.7

Continued…

Uniprot Id	Gene	DE during co-infection	HIV genes	Bridge Proteins	Drug targets	Total Variants	Novel variants	Higher frequency in African populations	Common in African populations	Deleterious	Average Prioritisation Percentile Rank
Q08345	DDR1	No	-	-	PTGS2 (Anti-TB: Aminosalicylic Acid)	11	5	4	3	3	67.63
P25440	BRD2	No	-	-	-	15	1	7	4	3	64.75
Q31612	HLA-B	-	Gag, Nef	-	-	65	65	0	0	0	60.33
O75955	FLOT1	No	-	CAV1, FYN	-	4	2	1	0	1	58.98
P59796	GPX6	No	-	-	CYP2C8 (Anti-TB: Isoniazid, Rifampicin; Anti-HIV: Ritonavir, Zidovudine, Saquinavir)	314	299	7	6	10	58.43
P14373	TRIM27	No	-	-	-	3	1	1	0	1	55.65
Q96PV0	SYNGAP1	No	-	-	-	5	1	4	2	2	55.2
A0A096LP39	TRIM39-RPP21	-	-	-	-	4	4	0	0	0	53.38
P13942	COL11A2	No	-	-	-	35	1	14	6	17	51.38
P22105	TNXB	No	-	-	-	135	1	57	40	83	50.2
Q99941	ATF6B	No	-	-	-	9	1	5	0	3	42.5
O75715	GPX5	No	-	-	-	15	5	5	3	6	40.95
Q5SRN2	C6orf10	No	-	-	-	1	1	0	0	0	22.4
P58182	OR12D2	No	-	-	-	26	3	9	10	12	13.98
Q9UGF6	OR5V1	No	-	-	-	10	1	4	2	5	13.98
Q9UGF7	OR12D3	No	-	-	-	8	2	4	3	4	13.98