The University of Manchester

PATHOGENIC VIRUSES AND THEIR INTERACTION WITH HUMAN HOST CELLS

A THESIS
SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY (PHD)
IN THE FACULTY OF LIFE SCIENCES

JAMIE IAN MACPHERSON

2011

CONTENTS

Final word count: 60501

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Virology, like many other biological research topics, has benefited from the application of large-scale data generation and analysis. Particular effort has been applied to developing a greater understanding of prevalent human-pathogenic viruses including type-1 Human immuno-deficiecy virus (HIV-1) and Hepatitis C virus (HCV). For example, host-virus interaction networks have been researched and important factors required for virus replication or innate cellular defence have been elucidated. Thus, large-scale data sources have provided a wealth of information regarding virus replication and virus-host interaction that may directly influence research and development of new antiviral treatments. In this thesis, we present research into the interaction between pathogenic viruses – HCV, HIV-1 and SIV (the simian equivalent of HIV) – and their host cells. Our research is largely integrative, computational research of large-scale data sources. In particular, we employ network models and related modes of analysis, with emphasis on identifying novel drug targets among cellular factors. Initially, we provide relevant background on virology, large-scale data sources and associated computational methods. Following this, we present four research projects that investigate either HIV-1, HCV or SIV interaction with host cells. Finally, we present a detailed analysis of the relationships between large-scale network data and biological function using the *Saccharomyces cerevisiae* model and demonstrate the importance of composite interaction networks. In our research we show that integration of large-scale data, combined with bespoke computational analyses, can provide a means for investigating specific aspects of viral infection. In particular, using this approach we provide insight into host-virus interactions that influence HIV-1, SIV and HCV infection and we infer cellular functions and specific host factors that may be useful in the search for novel antiviral drug targets. Thus, we recommend that computational methods for analysing large-scale data sources continue to be developed, with particular emphasis on methods that integrate data sources, so that results can be derived from multiple types of information and more complete biological models.

DECLARATION

The University of Manchester
PhD by published work Candidate Declaration

Candidate Name: Jamie Ian MacPherson

Faculty: Faculty of Life Sciences

Thesis Title: Pathogenic viruses and their interaction with human host cells

Declaration to be completed by the candidate:

I declare that no portion of this work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signed: Date: December 1, 2011

COPYRIGHT

ACKNOWLEDGEMENTS

This work was funded by the BBSRC and Pfizer Global Research & Development. Many thanks to all of those people who supported and contributed to this work. In particular I would like to thank David Robertson, John Pinney, Marilyn Lewis and Ben Sidders. I would also like to thank my examiners, Professors Paul Kellam and Werner Muller for their valuable time and effort. The Robertson Laboratory, past and present, Dr Simon Lovell and Sam Griffiths-Jones and the wider computational biology community at the University of Manchester provided an enjoyable and supportive working environment, for which I shall always be grateful.

This work is dedicated to my mum, who taught me to read, and my dad, who taught me to write.

LIST OF ABBREVIATIONS

Listed below are abbreviations used in this document. Gene names that involve abbreviations are too numerous to practically include. However, standard gene names and symbols have been used throughout and these can be searched via the National Centre for Biotechnology Information (www.ncbi.nlm.nih.gov).

AGM – African Green Monkey
AIDS – Aquired Immunodeficiency Syndrome
ARACNE – Accurate Cellular Networks
ARD – Argenine-rich domain
BIND – the Biomolecular Interaction Network Database
BioGRID – the Biological General Repository for Interaction Datasets
CDF – Chip description file
DAG – Directed acyclic graph
DAVID – DatabaseforAnnotation, Visualization and Integrated Discovery
dsRNA – double-stranded RNA
ER – Endoplasmic Reticulum
FDA – the Food and Drug Administration
FDR – False discovery rate
FP – False positive
GEO – Gene Expression Omnibus
GLASS – Gene Layout by Semantic Similarity
GO – the Gene Ontology
HAART – Highly Active Antiretroviral Therapy
HCV – Hepatitis C Virus
HCVcc – HCV cell-culture
HDF – HIV-1 (or HCV) dependency factor
HHPID – the HIV-1-Human Protein Interaction Database
HIDT – HIV-1-interacting drug target
HIV – Human Immunodeficiency Virus
HPID – the Human Protein Interaction Database
IC – Information content
IRES – Internal Ribosomal Entry Site

ISG – Interferon-stimulated gene
KEGG – the Kyoto Encyclopedia of Genes and Genomes
LCC – Largest connected component
MCC – Matthews correlation coefficient
MCODE – the Molecular Complex Detection algorithm
MI – Mutual information
MIAME – Minimum Information About a Microarray Experiment
MINT – the Molecular Interaction Database
NLS – Nuclear localisation signal
PCA – Principal component analysis
PIC - Pre-initiation complex
PPI – Protein-protein interaction
PSIMAP – the Protein Structural Interactome Map
RM – Rhesus Macaque
RMA – Robust Multichip Average
RRE – Rev Response Element
siRNA – Small Interfering RNA
SIV – Simian Immunodeficiency Virus
SM – Sooty Mangabey
SQL – Structered query language
TAR – Transactivation-responsive region
TO – Term-overlap
TP – True positive
UCSC – the University of California, Santa Cruz
Y2H – Yeast-two-hybrid

INTRODUCTION TO THE THESIS

Permission was granted to present this thesis in the "alternative" format, i.e., the research chapters (chapters 3-7) appear as journal articles. The alternative format is appropriate for presentation of this thesis for several reasons. First, although the research projects presented form a coherent body of work, each chapter is a self-contained study with unique methods and independent results. Second, the research projects were collaborations, thus, these projects are presented in a manner agreed upon by each author. Last, chapters 3 and 4 have already been published and chapter 5 has recently been accepted for publication. Therefore, chapters 3-5 are presented in a form that has passed a peer review process. Although I am the main author for each chapter, for clarification, the sections below include the author contributions and as appropriate, the publishing journal, or the journal for which a manuscript is currently being prepared.

Chapter 3

JNets: Exploring networks by integrating annotation, Jamie I MacPherson[a], Jonathan E Dickerson[a], John W Pinney[b] and David L Robertson[a], *BMC Bioinformatics* 2009, 10:95.

Author contributions

The project was conceived by JIM, JED, JWP and DLR. The software was written by JIM and supervised by JWP. Project work and manuscript preparation were supervised by JWP and DLR. All other work, including the generation of results and the writing of the manuscript, was carried out by JIM.

Chapter 4

Patterns of HIV-1 protein interaction identify perturbed host-cellular subsystems, Jamie I MacPherson[a], Jonathan E Dickerson[a], John W Pinney[b] and David L Robertson[a], *PLoS Computational Biology* 2010, 6:7.

Author contributions

The project was conceived by JIM, JED, JWP and DLR. Project work and manuscript preparation were supervised by JWP and DLR. All other work, including the generation of results and the writing of the manuscript, was carried out by JIM.

Chapter 5

An integrated transcriptomic and meta-analysis of hepatoma cells reveals factors that inSSuence susceptibility to HCV infection, Jamie I MacPherson[a], Ben Sidders[c], Stefan Wieland[d], Paul Targett-Adams[c], Volker Lohmann[e], Perdita Backes[e], Oona Delpuech-Adams[c], Jin Zhong[f], Francis Chisari[d], Marilyn Lewis[c], Tanya Parkinson[c] and David L Robertson[a], *PLoS One*, accepted for publication.

Author contributions

TP, SW, FC and PTA conceived the microarray experiments. SW, JZ and FC performed HCVcc and hepatoma cell subcloning. PTA, VL and PB experimentally determined the set of HCV replication factors (termed HRFs). ODA performed an initial exploratory microarray analysis. JIM, BS, PTA, ML, ODA and DLR conceived the presented microarray and subsequent computational meta-analysis. JIM performed the presented microarray analysis and the subsequent computational meta-analysis. Method descriptions for laboratory assays were prepared and written by SW, FC, PTA, VL and TP. All remaining sections of the manuscript were written by JIM and FC, BS, SW, PTA, VL, ODA, TP, ML and DLR supervised the preparation of the manuscript.

Chapter 6

A comparative analysis of gene regulation in pathogenic and natural host SIV infections, Jamie I MacPherson[a], Béatrice Jacquelin[g], Arndt Benecke[g], Michaela C. Müller-Trutwin[g] and David L Robertson[a], *Retrovirology*, in preparation.

Author contributions

The project was conceived by JIM, BJ, MCMT and DR. AB performed the original gene expression study. Project work and manuscript preparation were supervised by BJ, AB, MCMT and DR. All other work, including results generation and writing of the manuscript was carried out by JIM.

Chapter 7

Network-driven perspectives of biological function, Jamie I MacPherson[a], Ryan M Ames[a], John W Pinney[b], Simon C Lovell[a] and David L Robertson[a], *Nature Biotechnol-*

ogy, in preparation.

Author contributions

The project was conceived by JIM, RA, SCL and DR. JIM and RA are joint first authors for this study. Specifically, RA performed gene ontology enrichment analysis, produced all GLASS visualisations and identified clusters that best characterise a given function from cluster data. JIM built the interaction networks, produced the network clusters, analysed GO enrichment results to determine ease of "capture" of general or specific functions and identified congruent network clusters. RA carried out the analysis and prepared figures described in the first and third paragraphs of the methods section "Functional enrichment in network clusters". JIM carried out the analysis and prepared figures described in: the second paragraph of the section "Functional enrichment in network clusters"; parts 2-4 of the section "Network Generation"; all of the section "Cluster generation"; and all of the section "Identification of congruent network clusters". Analysis in the remaining methods sections were shared by RM and JIM. The manuscript was written by JIM, RA and DLR. DLR wrote the first two paragraphs of the "Background" section. JIM and RMA co-authored paragraphs one and paragraphs three to seven of the results section "Functional enrichment in network clusters". The remainder of the manuscript, including the "Abstract" and "Discussion" sections, was written by JIM. Project work and manuscript preparation were supervised by SCL and DLR.

Affiliations

(a) Computational and Evolutionary Biology, Faculty of Life Sciences, University of Manchester, UK.

(b) Centre for Bioinformatics, Division of Molecular Biosciences, Imperial College London, UK.

(c) Pfizer Global Research & Development, Sandwich, UK.

(d) The Scripps Research Institute, La Jolla, California, USA.

(e) Department of Infectious Diseases, Molecular Virology, University of Heidelberg, Germany.

(f) Institut Pasteur of Shanghai, Chinese Academy of Science, Shanghai Institutes for Biological Sciences, Shanghai, China.

(g) Institut Pasteur, Unité de Régulation des Infections Rétrovirales, Paris, France.

CHAPTER

ONE

AN INTRODUCTION TO VIRUS BIOLOGY

1.1 Abstract

In this chapter we introduce viruses and briefly discuss basic aspects of their structure and replication. Following this we provide a more detailed description of Human immunodeficiency virus (HIV) and Hepatitis C virus (HCV) molecular biology, pathogenicity and treatment as these human-pathogenic viruses are the subject of research in later chapters of this thesis. Finally, with reference to both HIV-1 and HCV, we discuss the emerging science of identifying host cellular proteins that are required for virus replication and their use as novel antiviral drug targets.

1.2 Virus basics: virion structures and replication cycles

1.2.1 Viruses: not living but thriving

Viruses are not classified as living organisms, as they can not autonomously reproduce and they do not sustain their own metabolism [1]. Rather, viruses are complexes of biomolecules that infect cells and hijack the host metabolism and molecular machinery in order to replicate. Viruses are, however, more abundant than living organisms and are also highly diverse – they are present in every known sector of our ecosystem and account for more genetic sequence diversity than all living organisms put together [2, 3]. In addition, there is also evidence to suggest that viruses, or at least, virus-like particles, predate living organisms and were part of an “ancient Virus World" [4].

1.2.2 Virus structures

Depite their great diversity, the biomolecular structure of virus particles (virions) can be generalised with reference to some essential and common components:

All virions are equipped with a genome made from nucleic acid. In comparison to cellular genomes, viral genomes are small, ranging from about 3kb in the case of hepatitis

B virus [5] to 1200kb in the case of the (unusually) large Mimivirus [6]. Virus genomes can be constructed from double-stranded or single-stranded DNA or RNA. In the case where the genomes are single-stranded, the genes are either positive-sense (like mRNA) or negative-sense (complementary to mRNA). Indeed, it is according to these aspects of genome structure that viruses are often classified, in a system known as the Baltimore classification [7].

Viral genomes are encased in a protein coat known as a capsid. Capsids are formed from a multimeric complex of individual protein capsomers. The shape that the virus particle forms, known as the morphology varies but is usually either rod shaped, e.g., tobacco mosiac virus; polyhedral, e.g., adenovirus; spherical enveloped viruses, e.g., influenza, or complex, such as having a polyhedral head from which a tail apparatus extends, e.g., bacteriophage T4 [8]. Enveloped viruses have an outer membrane envelope, derived from the host cell membrane or an intracellular membrane [8]. On the outer perimeter of virions, for example, inserted into a membrane envelope, may be additional proteins whose roles in virus biology typically include attachment to cellular membranes via cellular receptors and entry into cells [8].

1.2.3 Virus replication

Viral reproductive cycles can be divided into several stages: attachment to the host cell, entry into the cell, uncoating, replication, virion assembly and release from the cell.

Firstly, virions attach to the surface of the cell through interaction between the virion and the host cell membrane. In the case of higher eukaryotic viruses, attachment and entry is typically specific to a certain subset of cells, as viral envelope glycoproteins target a specific cell-surface receptor that is unique to a specific cell type. For example, gp120 proteins on HIV-1 virions bind specifically to CD4 receptors on CD4+ lymphocytes [9] and HCV surface glycoproteins E1 and E2 bind to a small array of hepatocyte receptors including CD81, scavenger receptor SR-BI, and claudin-1 [10].

Next, by fusing with the host membrane, or via endocytosis, viruses enter the host cells. Endocytosis is a broad term that covers four major mechanisms – phagocytosis, macropinocytosis, clathrin mediated, and caveolin mediated endocytosis – whereby cells absorb particles that cannot pass directly through the cell membrane [11]. Clathrin mediated endocytosis is a process by which clathrin-coated cargo-containing vesicle are transferred accross the plasma membrane [12]. Clathrin mediated endocytosis occurs constitutively in most mammalian cells and is the main endocytosis mechanism observed for virus uptake, including cellular uptake of both HIV-1 and HCV [12, 10, 11].

Inside the host cell, the capsid proteins are removed from the viral genomic material, e.g., by enzymes in the host cell [8], in a process known as uncoating. Following uncoating, replication of virus genomes and production of viral proteins occurs within the cell. These processes are quite specific to the virus type and in particular, depend on their type of genome. For example, Influenza viruses, have negative-sense single-stranded RNA

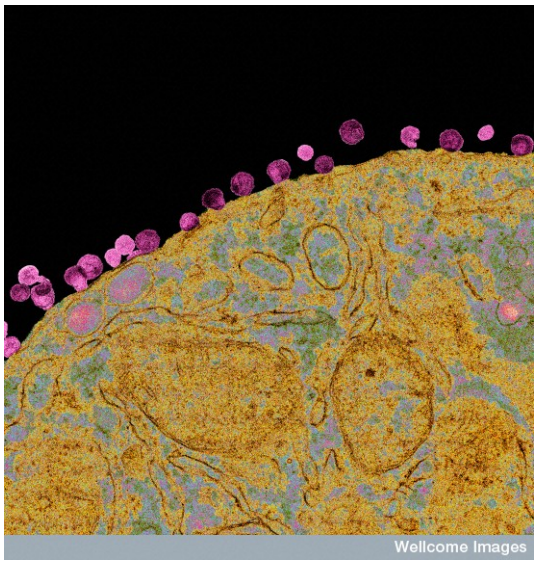

Figure 1.1: HIV virions budding from an infected cell. HIV particles are shown in pink budding from a host lyphocyte cell. Wellcome Images, available under Creative Commons Licence 2.0.

genomes (a type IV virus [7]). In order to produce positive sense RNA, both as a template for new genomes and for translation into viral proteins, the virus requires RNA-dependent RNA polymerase, encoded in its own genome and present also in small amounts as protein within the virion [13, 14]. Conversely, HIV-1 has a positive-sense single-stranded RNA genome from which double-stranded DNA is produced by the viral reverse transcriptase protein. Viral double-stranded DNA is inserted into the host nuclear genome in a process catalyzed by the HIV-1 integrase protein, that may then be transcribed and translated as with cellular DNA [15].

Following the production of new viral genomes and proteins, new virions become assembled and are released from the cell as new infectious particles in a process known as maturation. Though many viruses can self-assemble before release (virus components in buffered solution can form infectious and functional particles [16]), enveloped viruses tend to assemble as part of a scaffold with the cell membrane, where membrane bound virus proteins are located [17]. These enveloped viruses, including both HCV and HIV-1, are released by budding of the cell membrane to form the envelope of the virus, in which viral surface proteins are present. Figure 1.1 shows an image of budding HIV virions. Alternatively, virus release from cells can involve lysis of the cell. For example in Epstein-Barr virus infection, expression of certain ‘late’ genes and viral proteins promotes entry into the lytic cycle [18, 19].

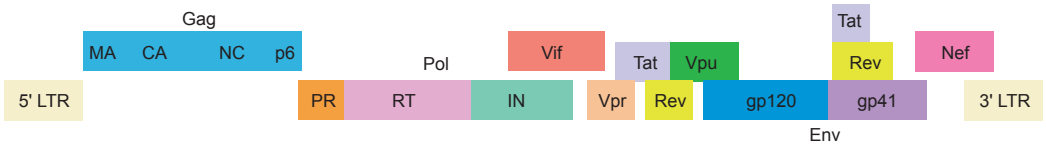

Figure 1.2: Schematic diagram of the HIV-1 genome. The genome has nine open reading frames: Gag, Pol, Vif, Vpr, Tat, Vpu, Rev, Env and Nef. The coding RNA is flanked by a 3' and a 5' UTR. The open reading frames encode 18 main viral proteins: Matrix (MA), Capsid (CA) nucleocapsid (NC), p6, Protease (PR, also known as Retropepsin), Reverse transcriptase (RT), Integrase (IN), Vif, Vpr, Tat, Rev, Vpu, Envelope glycoproteins gp120 and gp41 and Nef; and three polyproteins Gag, Pol and Env (also known as gp160).

1.3 Human-pathogenic viruses HIV-1 and HCV

1.3.1 HIV-1

HIV is a member of the genus *Lentivirus*, part of the *Retroviridae* family. *Retroviridae* are named so due to the 'backward' production of DNA from an RNA genome. Thus, HIV is classified as a type VI virus [7]. HIV infects human CD4+ lyphocytes and is the causative agent of acquired immunodeficiency syndrome (AIDS). The two known virus types, are HIV-1 and HIV-2. HIV-1 is the more potent of the two types, being both more prevalent and pathogenic [20], hence, the majority of scientific research on HIV, including the HIV research in this thesis, is focused on HIV-1.

The roles of HIV-1 proteins in virus replication

Each HIV-1 virion carries two copies of the genome, which is 9.7kb in length and contains nine reading frames that encode 18 proteins, including three functional polyproteins (figure 1.2) [21].

The Gag gene encodes for a polyprotein from which the structural proteins of HIV-1, Matrix (MA), Capsid (CA) and nucleocapsid (NC), are derived (figure 1.2) [22]. The core of the virion is composed from the genomic RNA bound to NC proteins, about which is a cone shaped capsid, formed from CA capsomers [22]. Outside the core is the virus matrix, composed from MA proteins that are associated both with the inner layer of a membrane envelope and with the core [23]. Figure 1.3 shows an illustration of a HIV virion. As with most HIV-1 proteins, NC, CA and MA are attributed with regulatory roles in virus reproduction, not all of which are fully understood. For example, MA has both a nuclear localisation and a nuclear export signal, so may have chaperone activity [23]. MA is also important for recruitment of Gag polyproteins to immature virions, where Gag is cleaved by PR, as with other HIV-1 polyproteins, as part of the maturation process [23]. NC also has more than just a structural role: it can cooperate with cellular transcription factors to enhance viral HIV-1 transcription [24].

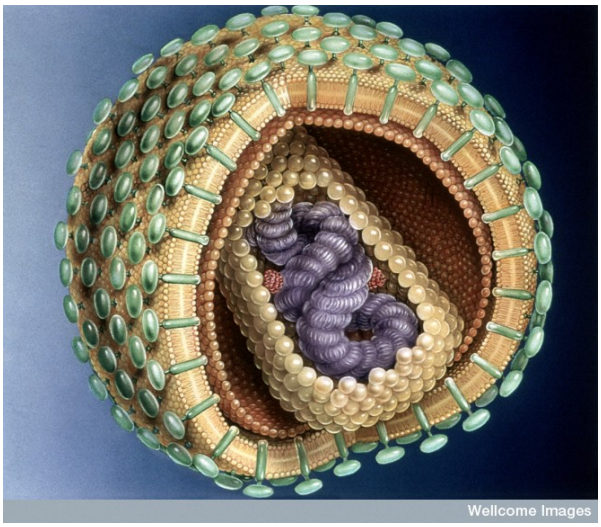

Figure 1.3: A cross-sectional illustration of a HIV virion. Coiled at the centre of the virion, encased in a cone-shaped capsid is the genomic RNA, Outside of the capsid, within a membrane envelope is the matrix layer. At the perimeter of the virion is the membrane envelope into which the envelope glycoproteins complexes are inserted. Wellcome Images, available under Creative Commons Licence 2.0.

Inserted into the membrane envelope of HIV-1 are trimeric complexes consisting of glycoproteins dimers – a gp41 transmembrane protein and a gp120. The primary function of this complex is to facilitate virus attachment and entry into host cells. Entry is achieved by the binding of gp120 to CD4 molecules on the surface of lymphocytes, the complex spike then undergoes a conformational change, allowing gp120 to bind to a secondary chemokine receptor, either CXCR4 or CCR5, depending of the virus tropism. This two-pronged attachment allows gp41 to penetrate the membrane of the host cell [9].

The reverse transcriptase protein is an essential enzyme for production of double-stranded DNA from RNA. Not only does HIV-1 RT catalyse synthesis of negative-sense cDNA (complementary to the HIV-1 genome) it also has an RNaseH domain that degrades genomic RNA before synthesis of the second DNA strand. When the double-stranded DNA has been sythesised, it is translocated to the nucleus and inserted into the host cellular genome in a process catalysed by the IN protein. The inserted DNA is known as a provirus – a template for both new viral genomes and mRNA. During DNA synthesis by RT, recombination is frequent and polymerisation is error-prone. Furthermore, both of the parental RNA genomes can contribute to producing a single DNA copy. These promote rapid evolution of HIV-1 that can have significant implications on the appearance of drug-resistant virus strains. [15]

HIV-1 Tat, is the transactivator of viral transcription. Tat binds to the transactivation-responsive region (TAR) region of in the viral DNA and recruits a transcription activator complex, P-TEFb, forming what is usually termed the “pre-initiation complex”, that in turn activates RNA polymerase II [25]. In addition, Tat has been shown to recruit histone acetyltransferases, in order to modify chromatin structure and promote transcription [25]. As a transcriptional activator, Tat also has many regulatory effects on cellular genes

linked to a range of functions (indeed several hundred such regulatory relationships have been catalogued [26, 27]), including immune system regulation [28, 29, 30, 31], signal transduction [32, 33] and cell survival [34, 35].

Once viral RNA transcripts are produced, in order to act as either viral genomes or as translated messengers, they must be exported out of the cell nucleus and into the cytoplasm. However, these transcripts contain introns, but nuclear export of intron-containing RNA is not a typical cellular activity. Thus, nuclear export of viral RNA can be facilitated by HIV-1 Rev [36]. A nuclear export signal Rev binds to the host protein exportin 1and the argenine-rich domain (ARD) in Rev binds to a Rev-response element (RRE). The exportin-Rev-RNA complex docks at a nuclear pore complex, and in an interaction that is mediated by nucleoporins, RNA passes across the nuclear membrane [36, 37, 38].

HIV-1 also has what are termed "accessory proteins" – Nef, Vif, Vpr, and Vpu. Though these proteins have various roles, their main theme of activity involves manipulation and evasion of the immune responses of the host [39]. For example, a major role of Vif (Viral infectivity factor), is for counteraction of APOBEC3G (apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G). APOBEC3G is a cytidine deaminase that is active as part of the innate immune response to viral infection. By binding to HIV-1 genomic RNA, APOBEC3G can create many cytidine (C) to uridine (U) mutations to cause loss of the encoded signal and also be incorporated into newly formed HIV-1 virions [39]. However, Vif binds to APOBEC3G and recruits an enzyme that ubiquitinates APOBEC3G, targeting it for proteasomal degredation [39]. Thus, Vif inhibits APOBEC3G-mediated innate immunity. Vpu can also abrogate interferon-α dependent tethering of fully formed virions at the cell surface, again, as mechanism to avoid innate immune activity. In the context of acquired immune system avoidance, both Nef and Vpu can alter the ectopic expression of antigen-presenting MHC class molecules [39]. The established role of Vpr is somewhat different: Vpr arrests infected cells in the G2 stage of the cell cycle. G2 arrest is thought to increase viral replication by upregulating transcription and translation of viral genes [40].

Lastly, the p6 protein, found at the C-terminus of the Gag polyprotein, is required for incorporation of Vpr into HIV-1 virions [41].

HIV-1 pathogenicity

HIV-1 infection can be divided into two clear stages, acute and chronic infection. Acute infection follows the initial infection event. During initial acute infection HIV-1 primarily infects resting CD4+ T cells in areas such as the intestinal and genital mucosa. Within two weeks of becoming infected, extreme levels of virus replication occurs, infection plasma viral loads of roughly 10^7 to over 10^8 million copies per ml can be observed and a large proportion of the total CD4+ effector memory T cells in these zones become infected. As a result, CD4+ cell counts are are dramatically reduced. Following this period of intense replication, a necessary switch takes place whereby infection begins to

target activated, dividing T cells, marking the transition into chronic infection. [42]

Though much attention is paid to deterioration in health that is observed in the chronic stages of infection, the massive extent of acute HIV-1 infection is also associated with some severe disease symptoms including fever, fatigue, weight loss, ulceration and neuropathy [43]. Furthermore, the severity of acute infection is thought to be closely linked to the severity of immune system failure during chronic infection [42].

During chronic HIV-1 infection, viral loads become lower [42] and the infection is largely asymptomatic [44]. However, throughout chronic infection the host undergoes increasing general immune activation such as increased pro-inflammatory cytokine production, increased frequency of activated memory T cells [44]. The reasons for chronic immune activation are not entirely clear, however virally-induced innate immune activation could play an important role [42]. Eventually, typically over the course of a number of years, the regenerative capacity of the immune system becomes exhausted and the T cell counts in the host fall below an effective minimum population. This point marks the beginning of AIDS and fatal immune system collapse [44].

Simian Immunodeficiency Virus does not cause AIDS in natural SIV hosts, such as African Green Monkeys and Sooty Mangabeys [45]. However, in some non-human primates that are non-natural hosts, such as macaques, SIV causes AIDS by a similar, if somewhat accelerated pathway to that observed in human HIV-1 infection [46]. This paradigm has allowed some interesting studies (e.g., [47, 48, 49]) to try to identify the molecular determinants of HIV-1 pathogenicity, using SIV infected non-human primates as biological models. These concepts are studied in greater detail in chapter 6.

HIV-1 Epidemiology

According to the Joint United Nations Program on HIV/AIDS report from 2010, over 33 million people are infected with HIV [50]. Furthermore, HIV-1 infection is a truly global phenomenon as more than one in 1000 people are infected in most countries for which data is available; tens of countries from around the world, including Russia, Portugal and the United States have a prevalence of at least one in 500; and in many sub-saharan African countries, more than one in 20 are infected [50].

Due to the widespread infection frequencies, HIV-1 is considered by the World Health Organisation to be a pandemic. The major limiting factor to the spread of HIV-1 is the limited routes for transmission. HIV-1 infection is only transmitted through direct contact of bodily fluids between an infected and uninfected individual, typically through sexual transmission, blood (such as sharing hypodermic needles and infected blood transfusions), of from an infected mother to their child during pregnancy, childbirth or breast feeding [51]. In practical terms, prevention of HIV-1 infection can be achieved by relatively simple methods, e.g., injecting only with sterile needles and using condoms. However, the social and economic requirements for successfully managing HIV-1 infection on a global scale are clearly problematic and as a result millions of new HIV-1 infections

occur every year [50].

1.3.2 HIV-1 treatment

Despite recent advances in HIV-1 vaccine development, there is no effective HIV-1 vaccine [52]. However, there are over twenty antiretroviral drugs that are approved for use in order to prohibit HIV-1 replication and the onset of AIDS.

Antiretroviral drugs can be categorised according to their mechanism of action. The two categories that together account for the majority of antiretrovirals are protease inhibitors and reverse transcriptase inhibitors. Protease inhibitors, of which there are currently ten approved drugs [51], inhibit the HIV-1 protease and prevent cleavage of viral polyproteins to individual proteins. Reverse transcriptase inhibitors, that prevent reverse transcription of viral cDNA from genomic RNA, fall into two subtypes, nucleoside analogues and non-nucleoside analogues. Nucleoside analogues prevent cDNA synthesis by acting as competitive RT substrates that when incorporated into viral DNA cause termination of DNA synthesis [53]. Non-nucleoside/nucleotide analogues bind to a hydrophobic pocket of RT to alter the structure and inhibit the activity of the enzyme [53]. Other types of antiretroviral agents include cell entry and integrase inhibitors.

These antiretroviral agents are, in the current protocol for treatment, typically administered in combination, in a system known as highly active antiretroviral therapy (HAART). HAART is normally started during chronic HIV-1 infection when CD4+ T cell counts become low, e.g., lower than 200 cells/mm^3 [54]. Typically HAART involves two nucleoside analogues and either a non-nucleoside analogue or a protease inhibitor. However, the combination of agents is frequently altered in order to avoid intolerance, toxicity and the appearance of resistance mutations [54]. Although, owing to proviral integration into host cell genomes, HAART cannot eradicate HIV-1, it is a very effective program of treatment and can prevent the onset of AIDS for the lifetime of an infected patient [54].

A major drawback of HAART is the ongoing requirement for access to pharmaceuticals and treatment monitoring. In low- and middle-income countries only an estimated 36% of those patients in need were receiving treatment. However, this percentage has increased in recent years. [50]

1.3.3 HCV

“Hepatitis” means inflammation of the liver [55]. HCV primarily infects hepatocytes found in the human liver [56, 55] and is a major cause of acute and chronic hepatitis, hepatocellular carcinoma and liver cirrhosis. HCV is the soul member of the genus *Hepacivirus*, part of the *Flaviviridae* family. Flaviviridae include several other human-pathogenic viruses including yellow fever virus, dengue fever virus and Japanese and Tick-bourne encephalitis [57]. Like all Flaviviridae, HCV has a positive-sense, single stranded genome. Thus, HCV is classified as a type VI virus [7].

Figure 1.4: Schematic diagram of the HCV genome. The genome has a single main open-reading frame that encodes ten main proteins. In addition, within the Capsid (C) sequence is an alternative reading frame protein, termed Frameshift protein (or F).

The roles of HCV proteins in virus replication

The HCV genome encodes a single polyprotein of approximately 3000aa in length, that is cleaved to form ten main proteins; from the C to N-terminus these are C, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A and NS5B [58] (figure 1.4). Owing to advances in HCV cell culture models, many advances in our understanding of HCV biology have been made in the last ten years [59]. However, our knowledge of HCV replication does not yet match that of HIV-1.

The structural proteins of HCV are C (core) and envelope glycoproteins E1 and E2. The core protein encases the genomic RNA and forms the capsid of the virus. Core is biochemically very basic and can self-assemble to form HCV-like particles [58]. Surrounding the capsid is a membrane envelope, into which E1-E2 heterodimeric glycoproteins are inserted.

Liken to HIV-1, the prominent role of the surface glycoproteins is to mediate fusion and entry into target cells. Perhaps the best etablished HCV cell-entry mediator is CD81. CD81, that binds to the E2 protein to mediate HCV entry, is found on the cell surface of a number of cell types, including hepatocytes. However, not all CD81 expressing cells are susceptible to HCV entry [57], leading to the proposal that other receptors are required in this process. Other identified receptors include scavenger receptor B type I (SR-BI) and heparan sulfate proteoglycans that bind E2 [57], tight-junction proteins Claudin-1 and Occludin and low-density lipoprotein [58]. However, the role of these cellular proteins is far from fully elucidated. HCV particles enter the cell via the clathrin-mediated pathway, following which viral RNA may be released into the cytosol where virus replication can begin [60].

In addition to their structural and entry-mediating properties, HCV structural proteins have other roles. For example, Core is essential for association of HCV virus particles with the endoplasmic reticulum (ER). Core can also translocate into the mitochondria of infected cells, chaperone mitochondrial proteins and effect Ca^{2+} and apoptosis signals [58]. E1 and E2 are also thought to be important for ER localisation [57].

HCV replication occurs entirely in the cytosol [56]. Synthesis of HCV polyproteins is perceptively more straightforward than for HIV-1. Situated near to the 5' end of genomic HCV RNA is the internal ribosomal entry site (IRES). The IRES mediates binding of the RNA to the ribosome in order to form a translation initiation complex, consiting of

viral RNA and proteins, ribosomal subunits and cellular transcriptional activators such as eukaryotic translation initiation factor, eIF3 [57].

The non-structural proteins – NS2, NS3, NS4A, NS4B, NS5A and NS5B – are thought to be involved in genomic replication of HCV [58]. The function of the NS2-NS3 protease is not well understood. Originally a polyprotein of NS2-NS3 was termed an autoprotease, as it was assumed that the proteolytic active site in NS2, that is required for NS2-NS3 cleavage, acts alone through a cis-acting mechanism. However, it has been shown that the complex acts as a dimer. NS2 is required for the production of infectious virus particles, possibly as a late stage in virus reproduction during particle assembly [61, 58]. NS3 is multifunctional, it is an RNA helicase and a serine protease. As a protease, with NS4A as a cofactor, it is responsible for cleavage of viral polyproteins at junctions NS3-NS4A, NS4A-NS4B and NS4B-NS5A. The NS3-NS4A protease may also be important for blocking acute immune activation by inhibiting signals mediated by the double-stranded RNA sensor, RIG-I [62, 55]. As a helicase, NS3 interacts with the NS5B polymerase in order to couple RNA unwinding to ATP hydrolysis [58], presumably as part of RNA replication. HCV replication takes place in what has been termed a 'membranous web' – a matrix of small vesicles that are associated with the rough endoplasmic reticulum. The role of NS4B includes promoting formation of these vesicles and acting as a membrane anchor for the HCV replication complex, though several other putative functions have been suggested including modulation of NS5B activity and interleukin 8 signaling [58, 55]. NS5A is an RNA binding phospho-protein. NS5A is present in hyper- and basally phosphorylated forms. The basally phosphorylated form is associated with increased RNA replication [58]. NS5A and NS5B are also required for assembly of the replication complex, possibly by interacting with vesicle membrane-associated protein hVap-33 [55]. NS5A has also been attributed with other roles including inhibiting protein kinase R, another innate immune system double-stranded RNA sensor [55].

The HCV p7 protein is a membrane protein that can act as a Ca^{2+} ion channel. Little is known about the details of p7 activity, though it is required for effective HCV infection *in vitro* and it may have a role in virus particle maturation and release from the cell [61].

In addition, there is another protein, the frame-shift protein, F, that is encoded in an alternative reading frame to the other HCV proteins, within the capsid-encoding sequence. The F protein has been shown to be present in some chronically infected patients but the function and also the mechanism by which this protein is produced are not well understood, although putative functions have been proposed including prevention of apoptosis of infected cells [63, 64].

HCV pathogenicity

Like HIV-1, HCV infection also has acute and chronic stages. The acute infection usually lasts from two to six weeks and during this time the majority of patients do not experience any noticeable symptoms, though in some cases patients get jaundice or other

non-specific symptoms such as abdominal pain [55]. The symptoms during acute infection are probably caused by virus replication and the associated immune activation, in particular, a HCV-specific CD8 cytotoxic T cell response [65]. During the first six months of infection, some patients clear HCV completely, though this only occurs in 10-25% of cases, the majority progress to chronic hepatitis C [65].

Disease progression during chronic HCV infection is slow. HCV infected individuals may not experience any severe symptoms for over twenty years from contracting the infection [55]. However, on their eventual appearance, HCV disease symptoms are severe and include liver inflammation, liver fibrosis and cirrhosis and an increased risk of hepatocellular carcinoma [65]. Liver fibrosis is the build up of connective tissue caused by repeated scarring of the liver that may eventually lead to loss of function (known as cirrhosis).

Liver damage, as with in acute infection, is linked to host cytotoxic T cell response. However, disease symptoms are also linked to the pathogenic activities of HCV proteins [65]. For example, C, NS3 and NS5A proteins have been linked to an increase in oxidative stress – a contributer to fibrosis and possibly also to DNA damage that induces carcinoma. The core protein is also linked to steatosis (also known as "fatty liver") by inducing the build up of lipid droplets within cells [55]. Steatosis occurs in the majority of chronic HCV infections [65] and increases inflammation and contributes to inflammation and liver damage.

Though persistent inflammation and oxidative stress probably contribute hepatocellular carcinoma in chronic HCV infection, the causes are not entirely understood. However, the activity of some HCV proteins may be oncogenic. For example, HCV core stimulates growth factors and inhibits apoptosis. In addition, core, NS3 and NS5A can influence the activity of tumour suppressor, p53. Liken to this, NS5B downmodulates a second tumour suppressor, pRb. [65]

HCV Epidemiology

Owing to the often asymptomatic nature of HCV infection, an accurate figure for number of HCV infected individuals is difficult to estimate. However, according to World Health Organisation estimates, about 3% of the world population are infected with HCV and over 170 million people are currently at risk from developing liver cirrhosis as a result of their infection [66]. Due to the delayed onset of hepatitis C, unfortunately, in the medium term, the burden these infections have on society is likely to increase [67].

HCV can only be transmitted by direct blood-to-blood contact [57]. Indeed, many transmissions have occurred directly as a result of contaminated blood transfusions and this has impacted on the global distribution of HCV carriers. For example, 15-20% of the population in Egypt are (or have before clearance) been HCV infected and this epidemic can be directly linked to schistosomiasis treatment. Schistosomiasis is a parasitic infection caused by trematode worms, the treatment for which has included mass-treatment via

injection, a significant proportion of which were contaminated [68].

Other global hotspots include Italy, where before HCV screening, injections of nutrients and vitamins were popular [69]; and Romania, where contaminated transfusions and more recently, intravenous drug use have been relatively widespread. Both Italy and Romania have an HCV prevalence of $> 3\%$.

HCV Treatment

As with HIV-1, there is no HCV vaccine. A high level of virion production (estimated as 10^{12} virions per day), combined with the error-prone HCV RNA polymerase, causes frequent mutation of the viral genome (estimated at 8-18 mutations per year) resulting in production of immune escape mutants [70].

Treatment of chronic HCV infection is currently based on a combination of pegylated interferon-α and ribivarin. Interferon-α is a not a virus-specific treatment, rather it evokes a general antiviral immune response that is not fully characterised but includes antigen expression by class I major histocompatibility complexes, activation of immune effector cells, and activation and regulation of cytokine cascades [71]. Ribivarin is a broad spectrum guanosine nucleotide analogue that interferes with elongation of RNA-dependent RNA polymerisation but the mechanism of this drug is not greatly understood [71].

Unlike HIV-1, this program of therapy can cure chronic HCV infection, as there is no provirus integration. However, clearance only occurs in about 40% of cases [71] and in addition, treatment is often interrupted due to the adverse side effects of these drugs, such as flu-like symptoms caused by interferon-α and hemolytic anemia (destruction of red blood cells) caused by ribivarin [72].

1.4 Host cellular factors as targets for antiviral drugs

An emerging train of thought and research into antiviral therapy regards the potential for host-cellular factors, that are essential for viral reproduction, to act as novel drug targets [73, 74, 75]. The rationale behind this idea is that viruses rely on the functions of specific host proteins in order to replicate.

There are two major advantages of targeting host, rather than virus proteins. Firstly, host proteins are not liable to frequently mutate during the course of antiviral therapy – viral escape mutants would have to circumvent use of the cellular cofactor being drugged, perceptibly a more complex route for escape than slight alterations in structure to avoid direct drug binding [73, 74, 75]. Secondly, viruses have very few proteins to target compared with humans. Indeed, in addition to those host factors with established virological roles identified as important through specific (and often multiple) small scale studies, hundreds of novel host-cellular factors have been identified as essential by genome-scale siRNA screens, for several viruses including HIV-1, HCV and Influenza [76].

Currently, our portfolio of antiviral drugs that target host factors is small. For example, from over twenty antiretroviral drugs that are currently approved for use against HIV-1, only one, maraviroc, targets a host-cellular factor [75] – the CCR5 entry co-receptor [77]. Furthermore, this agent is not without problems, as maraviroc drives intra-patient HIV-1 populations towards usage of the alternative CXCR4 coreceptor, rendering maraviroc ineffective [78]. No host-targeting anti-HCV drugs have yet been approved. However, alisporivir, a Cyclophilin A inhibitor, has recently entered phase II trials [79] and inhibitors to microRNA mir-122 are also being investigated [80, 81].

Clearly, as a science, the approach of targeting host cellular factors as an antiviral approach is in its infancy. By researching and understanding more about how these host factors influence viral replication, we further our ability to successfully identify effective targets. By integrating information and conducting data mining we maximise the informative potential of the available information – this lemma embodies the motivation behind the research in this thesis.

CHAPTER
TWO

A REVIEW OF LARGE-SCALE BIOLOGICAL DATA SOURCES AND DATA MINING TECHNIQUES

2.1 Abstract

In this chapter we introduce sources of large-scale data including data derived from laboratory experiments and from computational inference. The methods for obtaining the data are described and examples of data sources are given. Following this we describe methods of data mining that are applicable to the analysis of large-scale biological data. Throughout, particular attention is paid to areas of research relevant to virus-host interaction, particularly the study of interaction networks. Lastly, we briefly discuss the role that computational biology has in furthering our understanding of virus infection.

2.2 Large-scale data sources

Large-scale biological data is most commonly information pertaining to one of the "omic" types of biological information, i.e., genes, transcripts, proteins and metabolites, of which there are many thousands of each entity in a typical human cell. Single large-scale data sources usually provide comparable information across many entities from one source, e.g., the sequences for many genes. Data sets pertaining to entire genomes and proteomes are quite usual, furthermore, recent advances in high-throughput sequencing have also allowed many genomes to be compared. Importantly, these data sets can be sufficiently large (and accurate) such that the data can form a basis for statistical and computational analyses.

There are many sources of large-scale data. Some sources stem from large laboratory experiments, e.g., genome-wide microarray analysis [82]; some are composite sources that store measurements taken by individual (but comparable) studies, e.g., repositories for protein and genetic interaction data, such as the Biological General Repository for Interaction Datasets (BioGRID) [83]; and some comprise meta-data that, biologically speaking, can be very diverse but is present in a standardised structure, e.g., Gene Ontol-

ogy (GO) annotation [84].

In the following subsections sources of large-scale data are briefly described. Not every source is covered, rather, focus is placed on data types that are relevant to work in this thesis, particularly gene expression data, protein and gene interaction data and biological annotation.

2.2.1 Gene expression profiling by DNA microarray

Gene expression profiling is possible using several methodologies including high-throughtput sequencing [85] and reverse-transcriptase polymerase chain reaction (often referred to as RT-PCR) [86]. However, perhaps the most prominent method is the use of DNA microarrays. Indeed, microarray-based gene expression profiling is, perhaps, the most commonly measured of all large-scale biological data. For example, the Gene Expression Omnibus data repository at NCBI hosts over 19 thousand publicly available microarray-based expression profiling experiments [87].

Briefly, DNA microarrays are small solid plates to which multiple strands of DNA are attached (often referred to as gene chips) [88]. The strands (known as probes), are specific sequences, such as sections of known genes, or simply short synthetic sequences that cover the possible sequence-space. Furthermore, probes are grouped into probe sets. Samples of mRNA are harvested from a source of interest and, depending on the specific platform, this mRNA may be used to produce cDNA or cRNA and one, or a mixture of these nucleic acid samples (known as the target sequences) are washed over the gene chip to allow hybridisation against probes of a complementary sequence [88]. Typically, the target DNA is fluorescence-tagged so that the hybridisation reaction causes emission of radiation that can be detected using a confocal microscope, to allow the quantity of specific sequences within the target pool to be determined [88]. Microarray technology has developed greatly over the last 20 years and commercially developed gene chips, kits and methods, from companies such as Affymetrix, Illumina and Agilent have improved the availability, coverage and accuracy of gene expression profiling. Many such methodological advances are reviewed in [89].

A major aim of many DNA microarray experiments concerns the identification of differentially expressed genes [89], that is, genes whose expression changes significantly between two differently treated biological samples. Thus, gene regulation can be linked to a specific treatment (such as a stimulus or a perturbation) and insights into gene functions and cellular responses can be gained. Considering this aim, there are many aspects additional to the biochemical process of microarray analysis that contribute to experimental findings including experimental design, probe-set annotation, filtering, quality control and statistical significance calculations. These aspects are described below.

At a basic level, experimental design involves the selection of samples, including treated samples and untreated controls for comparison. Typically, for reasons of reliability and increased statistical power, several biological samples are prepared and tested, in

addition, measurements may be taken more than once for each sample (known as technical replicates). Further to this, efficient designs have been conceived for specific types of experiment [90].

After raw array intensity measurements have been taken and before differentially expressed genes are identified, computational steps to annotate probe sets and also quality-check, normalise and filter the data, are usually taken.

Probe set annotation concerns linking array probes to known genes or transcripts from public databases. For example, Affymetrix produce annotations that correspond specifically to the arrays that they produce [91]. However, custom annotations for commercial arrays have also been produced that improve upon the precision and accuracy of commercial annotations by using more recent genome builds and by removal of certain "bad" probe sets [92, 93, 94, 95].

Quality control is an important part of microarray analysis as problems can arise at several steps during the labaratory protocol, including uneven hybridisation or fluorescence over the array and RNA degredation [96]. Quality checks vary in their complexity, the simplest being visual inspection of array intensities to identify obvious anomolies, such as scratches. Specific software packages that perform a plethora of more complex metrics are available. For example, the *ArrayQualityMetrix* package [97] produces plots that can be used to assess RNA degradation, spatial distribution of probe-target hybridisation and comparative probe intensity distribution plots. In addition to these microarray-specific quality checks, more general methods, such as principal component analysis (PCA), can be employed [98] (see section 2.3.2 for a more detailed description of PCA).

Due to inconsistency in the preparation of samples and variation of environmental conditions during the laboratory steps of microarray protocols, normalisation is an essential step in analysis pipelines in order to allow samples to be compared without bias [98]. Several normalisation methods are commonly applied but arguably the best is quantile normalisation [99, 100, 101]. Quantile normalisation works by making same-size distributions equivalent. First probes from each array are ordered according to their intensity value. Second, the values at each equivalent position from all arrays are changed to the mean those values. Last, the probes are put back in the original order. Hence, when expression values are summarised for each probe set, their distributions will typically be very similar (figure 2.1). Indeed, as this method is simple, it is also computationally quick to calculate [100]. Quantile normalisation is the normalisation step used to calculate Robust Multichip Average (RMA) expression measures [99].

Microarrays can provide expression values for thousands of genes. For each gene for which the significance for differential expression is to be identified, a statistical test is required to calculate a probability (P). However, this leads to thousands of statistical tests being carried out and thus, a correction for multiple statistical tests must be applied. Indeed, the more tests that are performed, the more severe the correction. Therefore, by filtering out genes that appear to be insignificant before testing, potentially more genes

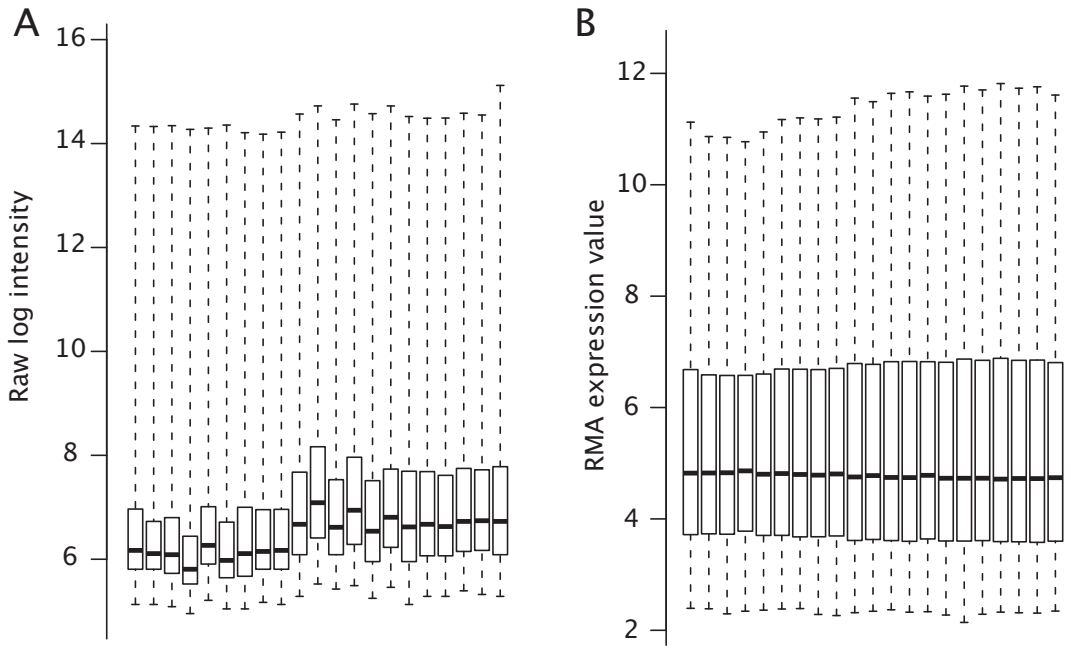

Figure 2.1: Example of microarray probe intensity normalisation. Box-and-whisker plots display probe intensity distributions for (A) raw sample intensity values and (B) the same samples after their intensity distributions have been normalised using RMA, a process that includes quantile normalisation. Sample distributions are clearly more similar to one another after normalisation.

attain statistical significance [102, 103]. Filtering methods utilise a variety of metrics including aspects of inter- and intra-sample variation [102], assessment of perfect versus mis-matched probe intensities [104] or simply removal of probe sets that do not meet a minimum expression value [102].

The probability for differential expression can be calculated using a statistical test, the classical choice for a simple two-sample test being a T test. In a T test, the replicates from one treatment are tested against the replicates from another treatment, for a single gene. The T test lacks power because in each individual test the majority of information available on the array is disregarded. Hence, more complex models for differential expression analysis have been developed that utilise more information to calculate significance. For example the *Limma* method [105] utilises a bayesian method that takes data values from other genes on the array into account to calculate a moderated T-statistic. As a result Limma generally outperforms the traditional T test as a method for identifying differentially expressed genes from the results of microarray analysis [106].

When differentially expressed genes have been identified, further analysis can take place in order to interpret the results. In particular, it is popular to perform: (i) functional annotation enrichment analysis using tools such as Database for Annotation, Visualization and Integrated Discovery (DAVID) functional enrichment tools [107, 108]; and (ii) perform clustering analysis, using methods such as hierarchical clustering to group genes (and also similar treatments) that share a similar profile of expression [98] (section 2.3.2

has greater detail on clustering methods). Although other methods that make use of other data types, such as protein interaction networks have also been developed [109].

Microarray analysis has been applied for the detection of gene-regulatory changes induced by the infection of pathogenic viruses, including HIV-1 [110] and HCV (e.g., [111, 112, 113]). The experimental designs utilised in these studies can compare preparations taken from uninfected cells to those from infected cells, or alternatively, time-course experiments can assess changes in gene-regulation over the course of infection.

2.2.2 Computationally inferred gene-regulatory interactions

Using the microarray platform it is possible to observe the effect that specific proteins have on the regulation of genes (gene-regulatory interactions). For example, a set of over 300 interferon-stimulated genes was verified using microarray analysis [114]. Likewise, the effects of virus proteins can be assessed. Izmailova *et al.* [115] monitored the effects of the HIV-1 Tat protein on the expression of several thousand cellular genes from dendritic cells and found that over 30 genes are differentially expressed. However, by this methodology, it is practically infeasible to elucidate the effect that every protein has on every other gene, i.e., an all-by-all analysis, for a human system. In an attempt to obtain such a universal view of gene regulatory relationships, computational inference methods have been developed that can reverse engineer regulation networks from expression data [116].

There are two major gene-regulatory inference methodologies, Bayesian networks and information theoretic approaches. Bayesian networks utilise joint probability distributions to infer causal relationships between genes. Gene regulatory relationships derived using Bayesian networks are represented by directed acyclic graphs (DAGs), where the nodes of the graph represent genes and the edges of the graph represent dependencies between those genes. In this model, the expression level of a gene is dependent on the other genes on which there is a dependency. The network structure can be estimated from gene expression data by a machine learning algorithm, typically by searching for the best network, according to a statistical scoring function [117].

Information theoretic approaches operate on the assumption that genes with similar expression profiles are likely to be linked by some form of regulatory event. Indeed, clustering genes of similar expression using a method such as hierarchical clustering, is also an information theoretic approach that can coarsely group genes that may be co-regulated. However, such clustering approaches make no attempt to separate direct regulatory interactions from those that are the indirect result of gene-regulatory interaction cascades [118]. In order to infer direct relationships, information theoretic approaches firstly calculate similarity, or mutual information (MI), between all pairs of genes. Secondly, by a data-processing step, the most direct, causal relationships are estimated. Established statistical measures, such as Spearman's correlation can be used as the similarity measure, or alternatively, certain specific MI statistics are also appropriate [119]. A variety of methods

for the aforementioned data processing step have been developed [118, 120, 107, 121], and in comparison to inferring Bayesian networks, the processing steps are relatively simple and computationally inexpensive. For example, the Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) is based on the theory of Data Processing Inequality [122], that states that if gene X, interacts with gene Y in a second order interaction through gene Z (rather than directly), then the MI value between X and Y, $MI(X;Y)$, will be smaller than either $MI(X;Z)$ or $MI(Y;Z)$. Using this rationale, weak edges within triplets and also edges with MI cutoff less than a given cutoff are removed [119]. The advantage of MI network approaches, is that large networks of many thousand genes can be inferred [107], an attribute not yet demonstrated for Bayesian networks.

A major problem with computationally inferred gene-regulatory interactions is the error rate, both in terms of precision and recall. For example, Altay *et al.* [123] demonstrate that even when the number of samples is as high as 200 (one sample being equivalent to a single microarray), the very best MI network inference method achieves an F score of roughly 0.45, i.e., the accuracy value, taking into true and false positives and also true and false negatives, is less than half of the potential value.

2.2.3 Protein-protein interactions (PPIs)

Whereas genes store the code that "programs" cellular biology, proteins are a major effector of this code. In order to carry out their tasks, proteins act in unison, dynamically regulating and physically transporting one another, coming together in specific combinations to form functional complexes and providing intracellular and extracellular signals to react to ever-changing environmental factors. All of these protein activities rely on a complex and dynamic network of interactions between proteins [124]. By elucidating what protein take part in interactions with one another, we can gain an insight into these activities. Though earlier technologies, such as GST pulldown and co-immunoprecipitation, allow PPIs to be discovered, large-scale (and even genome-wide) screens are now quite common using methods such mass spectrometry to identify complexes and yeast-two-hybrid assays to identify interacting protein pairs [124].

To identify protein complexes, proteins can be affinity-tagged and used as a bait for protein complex formation. Tagged complexes may then be purified and the subunits identified by mass-spectometry. A benefit of this method is that protein complexes can be allowed to form in a natural physiological environment, however, this also poses a drawback as the environment tested may not allow every true complex to form. [124]

Yeast-two-hybrid works using hybrid fusion proteins in a yeast system, one protein being fused to a DNA-binding domain (known as the *bait*), the other protein fused to a transcriptional activating domain (known as the *prey*). When the bait and the prey bind one another, they form a DNA-binding transcriptional activator that promotes expression of a reporter gene [125]. Variations and developments that use the theme of bait-prey

binding have been developed to allow greater numbers of interactions to be detected using large libraries of prey constructs and pooling of many preys to reduce the number of tests required [125]. However, yeast-two-hybrid studies have drawbacks, in particular, that the proteins being tested are not always in their native physiological environment, indeed they may be from a different organism than yeast. The second drawback of yeast-two-hybrid, as with all protein interaction assays, is that it is prone to both false-positive and false-negative errors. For example, it is estimated that protein interaction assays of *Drosophila melanogaster* proteins suffer from roughly 17% false positives and 28% false negatives [125].

The largest available data set of protein-protein interactions between HCV and host-cellular proteins was mainly produced using the yeast-two-hybrid system (and also partly by curating interactions from previous literature). This data set includes 314 protein interactions identified by yeast-two-hybrid. Follow up analysis of the implicated proteins revealed that HCV targets proteins from specific functions, for example, proteins involved in insulin signaling that can be directly linked to steaosis, thereby providing a molecular insight into the pathogenicity of HCV infection. [126]

2.2.4 Computationally inferred protein-protein interactions

Like gene-regulator interactions, protein-protein interactions can also be computationally inferred. Most of these prediction methods use sequence or structural data to predict PPIs, as it is likely that interactions have influenced the evolution of the protein and corresponding gene. Four examples follow, though there are many more methods that have been developed [125].

Phylogenetic profiling of proteins can be used to predict interactions. In this method, the co-existance of proteins in multiple species is considered, the rationale behind the method being that if certain proteins only operate as a functional unit a functional complex, then these the set will tend to be evolutionarily conserved in multiple genomes. The drawbacks of this method are that the method relies on accurate determination of presence or absence of similarly functioning homologues between species, thus relying on whole-genome sequences. Furthermore, this method fails for essential proteins that are present in all species [125].

PPIs can also be assigned directly from interactions that have been experimentally verified in another organism, by assuming that conserved proteins take part in conserved interactions. Such assumptions can also be made at the level of protein domains. For example, the protein structural interactome map (PSIMAP) classifies these interactions [127]. The Human Protein Interaction Database (HPID) has used both of these methods to predict interactions between human proteins. They utilise protein interactions verified in yeast and also domain assignments made by the Structural Classification of Proteins project [128].

Proteins that interact with one another have a tendency to be co-expressed [129]. How-

ever, relying on co-expression alone for assignment of interactions is clearly problematic, as indirect functional relationships (as well as spurious relationships) will also be present in the data. Despite this, co-expression can aid the prediction of PPIs [129].

Machine-learining methods are able to combine data such as co-expression, domain interactions and sequence information in order to make predictions [130], a methodology that has also been applied to prediction of HIV-1 protein interactions with host proteins [131]. The advantage of these predictions is that they are inexpensive and quick to perform. However, while these methods perform better than random the predictions are still subject to a considerable level of inaccuracy [130].

2.2.5 Composite interaction data sources

A study by von Mering *el al.* [124] compared the coverage attained by different large-scale protein-protein interaction detection methods, including both laboratory assays and computational inference, from a yeast system. Their findings indicated that the methods are biased towards detection of interactions between proteins of specific functional categories, cellular localisations and factors such as protein abundance can also have a significant effect. In every case, the overall coverage of each method is incomplete and accuracy is surprisingly poor, e.g., as low as $\sim$10% for early Y2H screens [124]. Clearly therefore, it is useful to consider data from a variety of sources in order to get a universal set of interactions between proteins. Indeed, using a composite source, in order to gain a set of interactions with greater confidence, it is possible to select interactions that have been verified by independent studies, or ideally, by more than one method.

There are a number of databases that provide interaction data, from a variety of sources, in an open-source and amenable format. These databases increase the potential usefulness of the data for the scientific community. Examples of such databases are The Biomolecular Interaction Network Database (BIND) [132], The Molecular Interaction Database (MINT) [133], BioGRID [134] and the HIV-1, Human Protein Interaction Database (HHPID) [26, 27].

BioGRID currently holds approximately 400 000 reports of physical protein-protein, gene-regulatory and gene-knockout interactions from a variety of organisms. These interactions are deposited by the scientific community from both high-throughput and small-scale detection methods in addition to interactions that are manually curated from literature. BIND contains networks, pathways, complexes and interactions [132]. Like BioGRID, BIND contains data from both high throughput and small-scale sources and is designed to be an open source, inclusive library of interactions. The MINT database is a collection of $>$95 000 PPIs, involving $>$27 000 proteins from 325 organisms [133]. These PPIs are collected from experimentally verified sources; although 90% are from high-throughput methods, the remaining 10% represents a large number of curated interactions.

The HHPID contains experimentally verified interactions between HIV-1 and human

proteins. The HHPID contains over 2589 unique HIV-1-human protein interactions, including both direct physical and indirect regulatory events and all interactions were manually are curated from primary literature – a seven year effort that involved the screening of over 100 000 journal articles [27]. Furthermore, the interactions are not simply simply binary events between two proteins, as every interaction includes a short description of the biological outcome of the event, such as "upregulates" or "inhibits", from a vocabulary of 68 such descriptions.

Though composite data sources provide a holistic, inclusive picture of biological interactions, they are not without their drawbacks. Firstly, those studies that include data from small-scale studies are subject to study bias. As, unlike a whole-genome screen, small-scale derived interactions, though perhaps rigorously proven, are a product of human choice. For example, the p53 protein has been extensively studied, particularly for the role it plays in preventing human cancers [135]. The p53 protein is the human protein in 19 records in the HHPID, involving five different HIV-1 proteins. While these may be *bona fide* interactions, the identification of such a large number of interactions is probably, in part, due to the interest the scientific community has in this protein. When testing is carried out without correction for this study bias, the results are potentially liable to report aspects of bias rather than of biology [136].

2.2.6 Biological annotation

Biological annotation is a highly valuable form of information. Genome and proteome resources such as the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) provide plethora of information about many thousands of genes. On a small-scale this information is useful for obtaining specific gene, transcipt and protein information, links to publications and further sources of information.

In large-scale analyses, biological annotation can also provide powerful insights. For example, gene sets identified by large-scale experiments are almost invariably subject to enrichment analysis, i.e., performing a statistical test in order to identify whether certain biological features (and associated annotating terms) are over-represented. Indeed many specific tools and packages have been developed to perform this exact type of analysis, these software typically requiring little more than a list of gene identifiers to obtain detailed results (for example [137, 138, 139]).

Probably the most widely used class of biological annotation in statistical and other computational analysis is GO annotation [84]. GO is a structured method of gene annotation, using a controlled vocabulary of annotating terms. Every GO term fall into one of three ontologies: biological process, referring to a theme of activity; cellular component, referring to a physical location that is associated with the gene product; and molecular function, refering to the biochemical and enzymatic activity of the gene product. The structure of each ontology is as a directed acyclic graph (DAG), where the vertices of the graph are the annotating terms. The root of an ontology is a node referring to one of bio-

logical process, cellular component or molecular function. Below the root node are more specific 'child' terms from which extend further child terms. The nodes may have multiple parent and child relationships but importantly, as a DAG, a node is never an ascendent and descendent of another node, ensuring that the structure is acyclic. [84]

Genes can be assigned with any number of GO terms and the details of the method for assignment of each term are available as one of about twenty *evidence codes*. These evidence codes include a computational inference, experimental validation and hand-curated term assignment. Annotation sets are available for a wide range of organisms, including humans. The human annotation set, compiled by the European Bioinformatics Institute [140], currently includes over 18 000 genes and over 200 000 gene-term associations and certainly is, therefore, a large-scale source of biological data.

2.2.7 Short Interfering RNA technology

RNA interference screening was discovered in 1991 following experiments in which double-stranded RNA (dsRNA) was injected into nematode worms and found to silence genes with complementary sequence [141, 142]. Since that time the technology has developed, in particular, the use of shorter dsRNA strands (less than 30 base pairs) circumvents activation of the innate antiviral immune response and has allowed RNA interference studies of mammalian cells [142]. Gene silencing involves processing of dsRNA into short-interfering RTNA (siRNA) by the RNase enzyme, Dicer. The siRNA is incorporated into an RNA-induced silencing complex whose activity is to silence complementary messenger RNA [142].

Using scaled up, genome-wide siRNA screens, phenotypes that are associated with the silencing of specific genes can be investigated. This screening procedure is highly appropriate for the study of host-virus systems, as it enables identification of host factors that have a direct influence on virus replication (these are referred to as "dependency factors"). Of particular interest are those dependency factors that are essential for replication of pathogenic viruses, as these factors present a clear opportunity for use as novel host-cellular drug targets [143, 144].

A recent boom in siRNA screening of host-virus systems have identified dependency factor sets for HIV-1 [144, 145, 146], Influenza [147, 148, 149], HCV [150, 151, 152, 153, 154] and Western Nile Virus [155]. A meta-analysis of three genome-wide screens of the HIV-1 host system [144, 145, 146] by Bushman *el al.* [156] showed that the intersection in the identified host genes, from all studies, was surprisingly small at just three genes. Despite the apparent lack of overlap, through integration of PPI data and biological annotation Bushman *el al.* identify functional components, such as subunits of the proteasome, that are significantly represented by multiple studies and are therefore important for HIV-1 replication. Their study illustrates that differences in experimental procedure can lead to quite different results, but perhaps most importantly, that the greatest biological insight is gained through integrative analysis of large-scale data.

2.3 Data mining in biology

Large biological data sets allow robust computational analyses of biological processes and cellular activity, including the study of virus infection. At a simple level this can involve the use of classical statistical tests to determine aspects such as significance of enrichment of biological annotation. Indeed, large experiments, such as whole genome microarray, can not be uncoupled from computational analysis because the amount of data created is so great that large-scale statistical and mathematical analysis are essential.

However, the data types currently presented, including genome-scale screens, networks of protein and regulatory interactions, structured systems of annotation and potentially multiple layers of primary experimental and secondary inferred data, are unequivocally the domain of intricate, integrative models of biology and methods of analysis that are confined to computational studies. This type of study is often termed *data mining*. Here, we describe some important biological data-mining concepts. Though data mining encompasses a large variety statistical and algorithmic methods, here we focus primarily on methods of data representation and data analysis that concern, or are compatible with biological networks.

2.3.1 Biological networks

In the case where relationships can be defined between multiple entities, the resulting data structure can be modelled by a network, otherwise known as a graph, where the nodes (or vertices) of the network represent the entities and the edges between those nodes represent relationships between the entities [157]. This concept is common in many varied fields outside of biology, for example, diagrams of electrical circuits and topological maps of railway stations are frequently represented by networks (figure 2.2).

The terms *graph* and *network* are, in this field, interchangeable – the former stressing a mathematical concept and the latter stressing the application [157]. A particular advantage of modeling biological systems using graphs is that are then amenable to graph theoretical analysis, an established field of mathematics. The following definitions describe some of the concepts of graph theory as describe in a review by Huber *et al.* [157]:

A graph is specified by a set of nodes V and edges E; each edge from E connects two nodes from V (except in the case of hypergraphs, where edges can connect multiple nodes). Node interactions (i.e., edges) can be *directed* or *undirected*. For example a transcription activation complex acting to upregulate a certain protein could be considered a directed interaction. Nodes of a graph are said to be *adjacent* if they are connected by an edge.

An edge ending at a node is said to be an *incident* at node. The *degree* of a node is the number of edge incidences at that node; a hub node has a relatively high *degree*. The degree includes incidences caused by *self-loops*; where an edge has both ends at the same node. A *complete graph* is one where all nodes are joined to all other nodes by an edge.

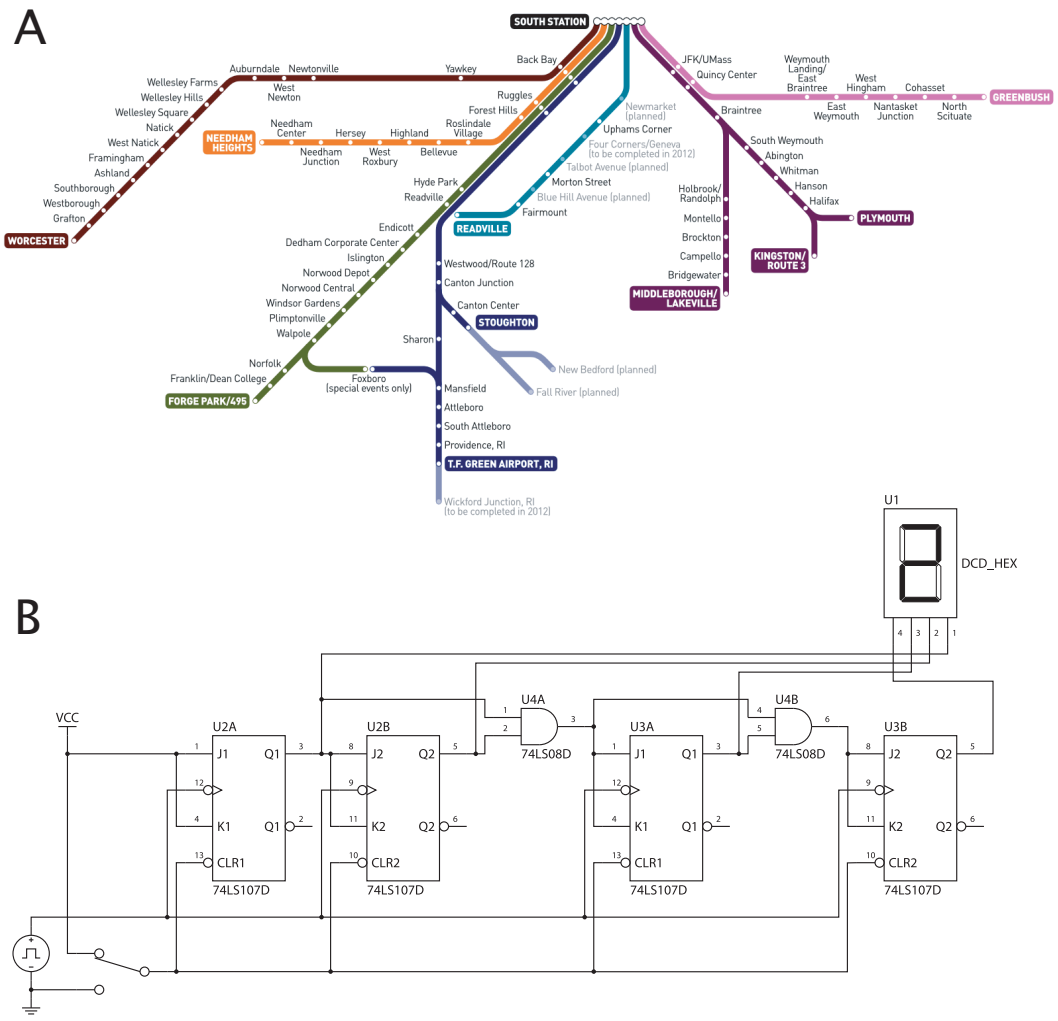

Figure 2.2: Examples of networks. (A) Topological diagram representing a rail network (Source: The Port of Authority, Creative Commons Attribution-ShareAlike 3.0 License, GNU Free Documentation License). (B) Electronic circuit diagram. Both are examples of networks, where nodes represent stations or electrical components and edges represent rail links or flow of electricity, respectively. Neither diagram is to scale but importantly, connections between nodes are clearly represented.

Two nodes are *reachable* from one another if one can traverse a set of alternating nodes and edges between the two nodes in question; the alternating set of nodes and edges between a pair of reachable nodes is called a *walk* between nodes. A graph is *connected* if a walks exists between every node pair. Types of walks include; *paths*, walks with no repeated nodes; *trails*, walks with no repeated edges and the *distance* is the shortest possible walk between two nodes.

The *connectivity* of a graph is the minimum number of edges that can be removed before one node or nodes becomes disconnected from the rest of the graph , i.e., one or more nodes are not reachable from the other nodes. The set of edges removed to disconnect a graph is referred to as a *cut*. Therefore, to work out the connectivity one works out the

minimum cut. Graphs may have a connectivity of 0; as they may already have nodes that are not reachable from others (a *disconnected* graph) – this is a fairly common feature of current biological networks, as our knowledge of interactions is incomplete.

A well described feature of PPI networks is their “small world", scale free structure (e.g [158]). A small world network is one where any one node can be reached from any other node in a few steps; therefore, the mean shortest path length is small [158, 159]. In addition, the average clustering coefficient in small world networks of nodes is higher than would be found in a random network [158]. The clustering coefficient is a mathematical term for measuring how close a node and all adjacent nodes are from forming a clique. Therefore, a property of small-world networks is that they tend to have a high number of cliques and cohesive-subgroups.

PPI networks also tend to have a few well connected ‘hub’ proteins with a high degree and many proteins with a low degree. This property remains the same no matter what the size of the network – the degree distribution is independent of the scale. As a result, biological networks are often described as “scale free” [160]. The nature of these scale free networks may explain some of their robust properties. For example, it has been suggested that because most proteins are not highly connected, these proteins, unlike hubs, may be removed without failure of the overall system [161].

Areas within graphs that have a high incidence of internal connections can be referred to as *cohesive subgroups*. A *clique* is a type of cohesive subgroup, formed by a set of nodes, where every pair of nodes is connected via an edge. A clique that cannot be made larger by the addition of more nodes is called a *maximal clique*. However, among biological networks the presence of true cliques may not be common (especially given incomplete interaction data). Therefore, other definitions for describing cohesive subgroups have been developed. For example a *k-plex* is a subgraph containing x nodes where each node is adjacent to at least n other nodes, where $n = x - k$.

Computational analysis using graph theoretical methods can be used to to make inferences regarding the function and structure of biological networks. For example, network module identification has attracted significant study. Modules are subnetworks that perform specific biological function, thus forming ‘building blocks’ for whole biological systems. Therefore, the ability to predict these modules *in silico* from network topology and network parameters is potentially valuable for understanding biological systems. One of the main problems with studying multi-body subgroups, is the computational complexity. For example, there are about 1×10^{23} different 10-node subsets within a 1000-node network [162]. Hence, predicting functional modules in biological networks is a challenging research topic [163] though, successful studies in this field have been published.

For example, Spirin and Mirny [162] examined yeast protein interaction networks in order to identify important functional modules. In their work they present an algorithm to mathematically investigate cohesive clique-like subnetworks that are highly connected within the subnetwork but poorly connected to the remaining network. By applying this

algorithm, they discover over fifty statistically significant subnetworks, including some modules with an established function and some novel modules. Most modules fell into one of two categories; protein complexes, such as large transcription factors, and functional units. The members of functional units perform the steps of a particular cellular process but not necessarily at the same time or in the same subcellular location, such as the proteins of a signaling cascade [162].

Some computational tools include analysis methods that specifically facilitate network analysis. Perhaps the most used and versatile network analysis software package is Cytoscape [164]. Cytoscape supports the integration of externally developed 'plug-ins'. For example, the NetworkAnalyzer plug-in [165] can calculate many network topology parameters including clustering coefficients, path lengths, degree distributions and neighbourhood connectivities. Furthermore, results from these calculations can be visualised in a network context for ease of interpretation.

Networks in biology do not have to represent proteins or genes. For example Yildrim *el al.* [166], present a network-based analysis of drugs and their targets, where nodes represent either a drug or a drug-target and edges represent a targeting event. By inclusion of drugs currently under development in their analysis, Yildrim *el al.* discover a trend towards the targeting of more diverse proteins, that have not been previously targeted. Thus, networks are useful for exploring a variety of biological data types.

Although mathematical distinctions can thoroughly describe features of graphs, the method for their accurate and useful application to real biological systems is non-trivial. Firstly, there are no rules on how to identify biological phenomena (such as functional modules) using graph-theoretic methods, particularly as this is a relatively new field. In addition, our knowledge of biological interactions is incomplete and may also include error or study bias. Secondly, consolidation of significant graph theoretical results with actual biological significance is problematic, as, by necessity, the network models are abstractions of reality. Therefore, during results interpretation, limitations of the data and the network model should be respected and should be performed with reference to established biological knowledge.

2.3.2 Clustering and distance metrics

Clustering is an important tool in data-mining for inspection and analysis of large data sets, as it has the ability to systematically arrange data points, of which there are likely to be too many to consider individually, into fewer groups (i.e., clusters) that share similar properties. By necessity, clustering must employ a distance measure in order to define how similar, or associated individual data points are to one another.

Clustering methods are used frequently during analysis of large numerical data sets, such as gene expression levels identified by microarray [98, 167], and also defining cohesive subgroups and functional modules in biological networks [168, 157]. Perhaps the most popular type of image associated with gene expression data is a heatmap showing

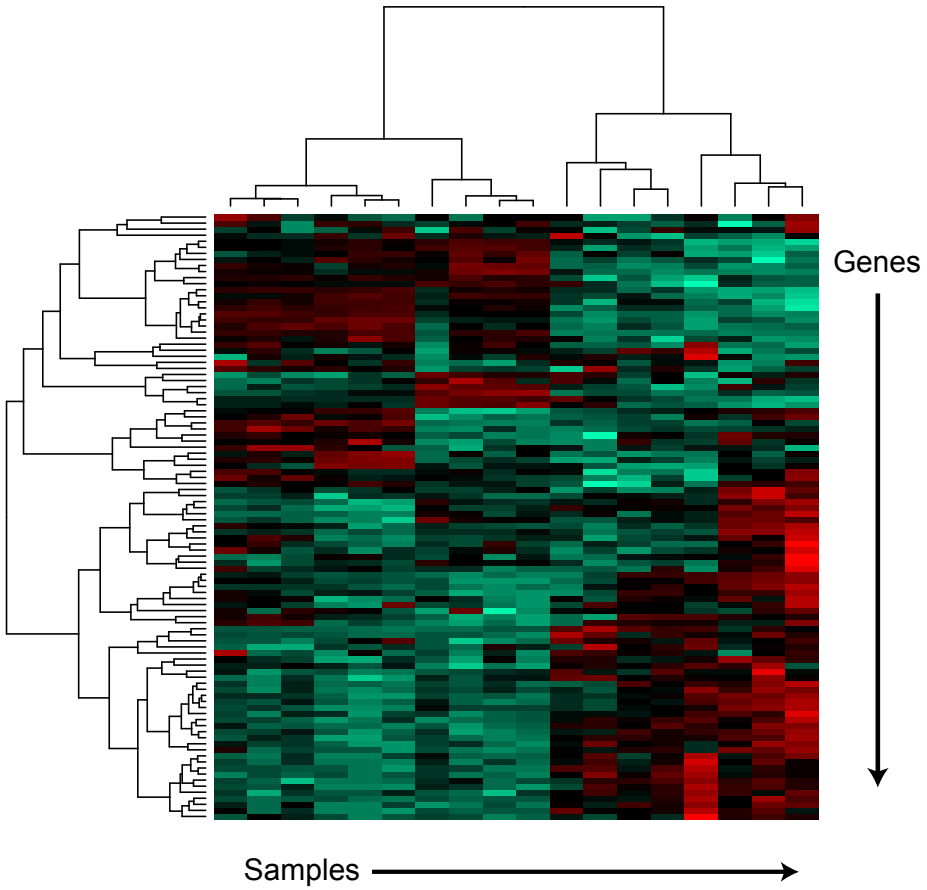

Figure 2.3: Heatmap of clustered gene expression profiles. Samples are represented by columns and genes by rows. Individual cells show the level of gene expression for the given sample, where green represents a lower expression value and red a higher expression value, the intensity of the colour denoting a lesser or greater value, respectively. Dendograms above and to the left show the clustering of samples and genes, respectively, that were produced using hierarchical clustering.

expression levels of genes across different samples (as in figure 2.3). In these plots, the genes tend to be clustered, according to their expression profile. Most examples of this type of clustering first require the calculation of a distance (or similarity) matrix that enumerates the similarities between gene expression profiles. Similarities can be measured using correlation statistics, such as Pearson's or Spearman's correlation, or alternatively using an Elucidean distance measure [167].

Clustering of genes is often performed using hierarchical clustering, an agglomerative approach whereby genes are put into groups and those groups are repeatedly merged until a single cluster is reached. The groupings are, thus, hierarchical and can be displayed using a dendogram. An alternative to the hierarchical method is k-means clustering, whereby the number of clusters to be created, k, is predefined by the user. In this method, genes are randomly assorted between k clusters and the mean inter- and intra-cluster distances are calculated. Multiple iterations are then performed where genes are moved between clusters to minimise intra-cluster distances and maximise inter-cluster distances. Though a relatively simple algorithm, k-means clustering is quite computa-

tionally expensive. In addition, it may be difficult to know what value for k will give the most intuitive results. [98]

Principal component analysis (PCA) can also be used to define distances between genes or samples. Importantly, PCA can reduce the dimensionality of the data, so that differences may be more readily displayed and interpreted. For example, multiple-sample microarray data can be reduced to two dimensions, so that samples can be plotted as points on a regular scatter plot – a useful method for checking that replicate array samples cluster together (figure 2.4A).

Network clustering algorithms can be used to partition networks into smaller subnetworks, in order to identify cohesive subgroups of nodes. This approach has been applied to successfully identify functional modules and protein complexes from yeast PPI networks [162] (described in greater detail in section 2.3.1). Network clustering has also been applied to host factors essential to HIV-1 replication identified from seemingly discordant siRNA sceens. In this study, host factors from different screens were found to be part of specific functional modules that are, therefore, essential components of HIV-1 replication [156].

A key aspect of cohesive subgroups within networks, is that they have a high internal density of edge connections and relatively few edge connections to the remaining network [168]. Hence algorithms to identify network clusters attempt to optimise these, or related characteristics. Despite this shared aim, a diverse array of algorithms for identification of network clusters have been developed. Two such algorithms are briefly described below.

Dunn *et al.* [169] developed a method known as edge-betweenness clustering, by decomposing the network. In edge-betweenness clustering, edges through which many shortest paths between nodes cross (known as a high *betweenness* coefficient) are repeatedly removed to cut the network into smaller subnetworks. This method tends the removal of edges that connect dense subnetworks, resulting in their separation from the network. Using GO enrichment analysis, these subnetworks were shown to be functionally cohesive units [169]. One drawback of this method is that the betweenness coefficient for every remaining edge must be recalculated after each edge-removal step. For large networks this is a computationally expensive process but a speed-up is possible by utilising an heuristic algorithm to the calculate shortest paths between nodes, such as that proposed by Dijkstra [170].

The Molecular Complex Detection algorithm (MCODE) [171] operates in a very different way to betweenness clustering, through a cluster building approach. MCODE has three main steps. First, all nodes are scored based on their clustering coefficient. Second, a seed node with the highest score is selected. Following this, neighbours of that seed, that have a score greater than a given percentage of the seed node score, are added to the cluster. This process continues until no more nodes can be added to the cluster. A new high-scoring seed that has not been clustered is then selected and the process is repeated. Post-processing steps remove clusters that fail to meet a criteria and, depending

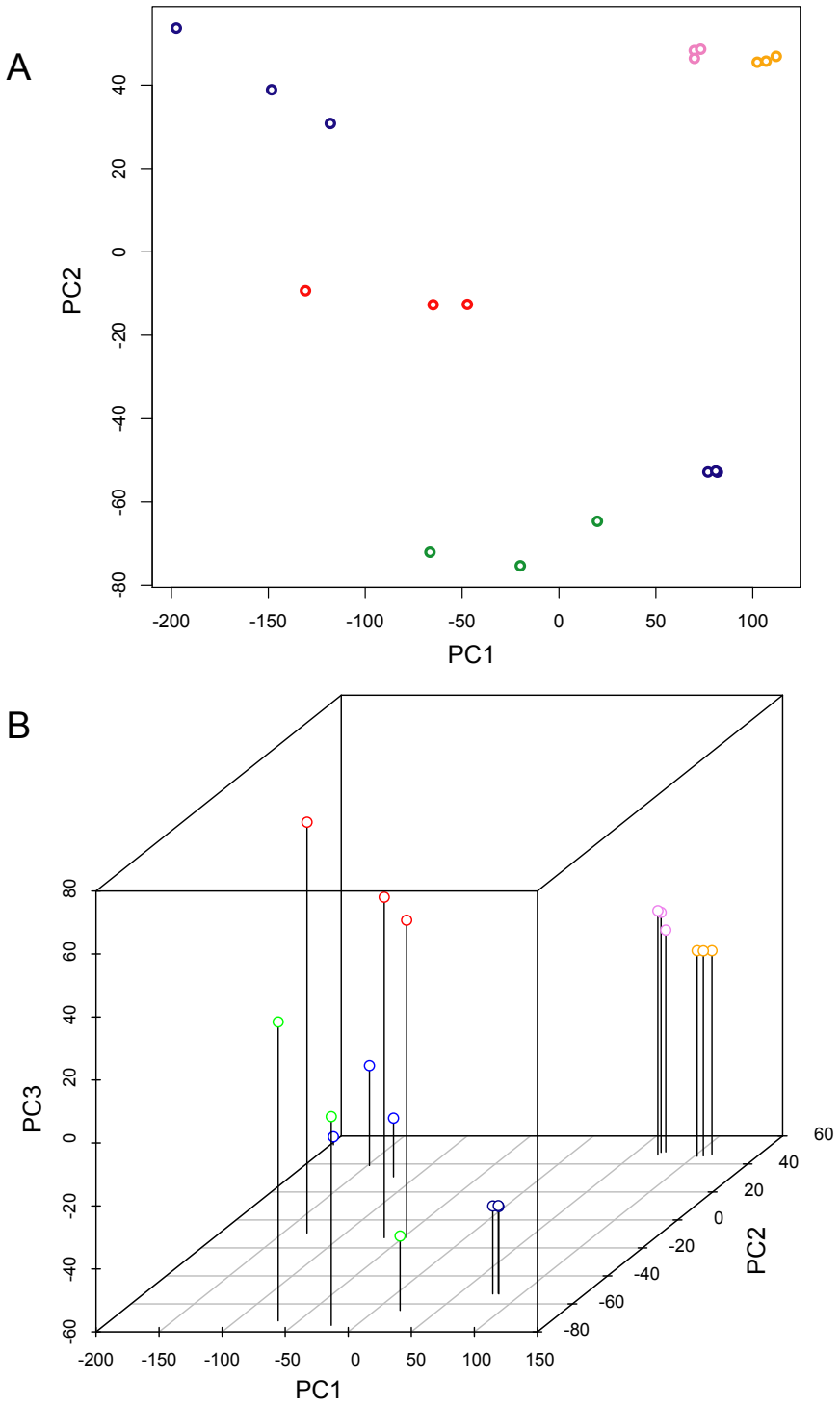

Figure 2.4: PCA plots of microarray samples plotted in both two and three dimensions. Samples are represented by points on the plots and biological replicates are plotted in the same colour. In (A), the first two principal components (PC1 and PC2) are plotted. In (B), the first three principal components (PC1, PC2 and PC3) are plotted on a three-dimensional scatterplot. Importantly, the plots highlight that the biological replicates tend to cluster closer to one another than to other samples.

on the user specifications certain unclustered nodes are added (termed "fluff") and certain poorly connected nodes are removed (termed "a haircut"). [171]

Because GO annotation conforms to a specific DAG structure, it is amenable to computational analysis additional to gene set enrichment calculations. In particular, methods that define a distance between annotating GO terms, and also between genes, have been developed [172]. These measurements are termed semantic distance measures. Several of these consider the information content (IC) of GO terms, where terms that are rarely assigned have a higher IC. Other methods apply a vector-space model, where each term corresponds to a dimension of the vector, on which a calculation is performed [172]. Perhaps the simplest method is the term-overlap (TO) method [173], whereby the distance between two genes is simply the sum of terms they have in common. Despite the simplicity of TO, it has been shown to perform as well as other methods including those that employ IC and vector-space calculations, at enumerating cohesiveness among sets of biologically related genes [173].

Semantic distance measures have two obvious applications. First, the distances can be used to define whether a given set of genes is more functionally cohesive than would be expected by random chance and second, relatedly, whether one gene set is more or less similar to a second gene set than would be expected by random chance. In addition, these distances have also been used to inform gene layout algorithms for visualisation of HIV-1 interacting proteins, so that genes of similar function are rendered close together [174]. Similarly, Holden *el al.* [175] show that an evolutionary distance measure between proteins can be used to produce informative layouts for PPI networks, in this case proteins with greater sequence conservation appear more closely together than those proteins that are more divergent.

2.3.3 Data visualisation

Data visualisation techniques can facilitate improved exploration, understanding and editing of datasets by taking advantage of two aspects of human vision: firstly, the broad bandwidth of the eye allows a vast amount of visual data to be dealt with at one time. Secondly, the visual system of humans is exceptional at spotting patterns in terms of size, space, colour and time [176]. Visualisation techniques are commonly used in the field of bioinformatics, particularly as data sets are often large and can be complex. For example, the Human Genome Browser at UCSC [177] allows customisable sections of chromosomes or individual genes to be visually explored with reference to intron/exon boundaries, cross-species homologies, expressed sequence tag alignments and more. Another commonly used visualisation technique in bioinformatics is viewing (and editing) multiple alignments with the aid of colours to show columns of aligned elements, e.g., Jalview [178].

Results from clustering analysis, such as clustering of gene expression profiles, are frequently displayed using heatmaps and graphical plots (figures 2.3 and 2.4) both for ease

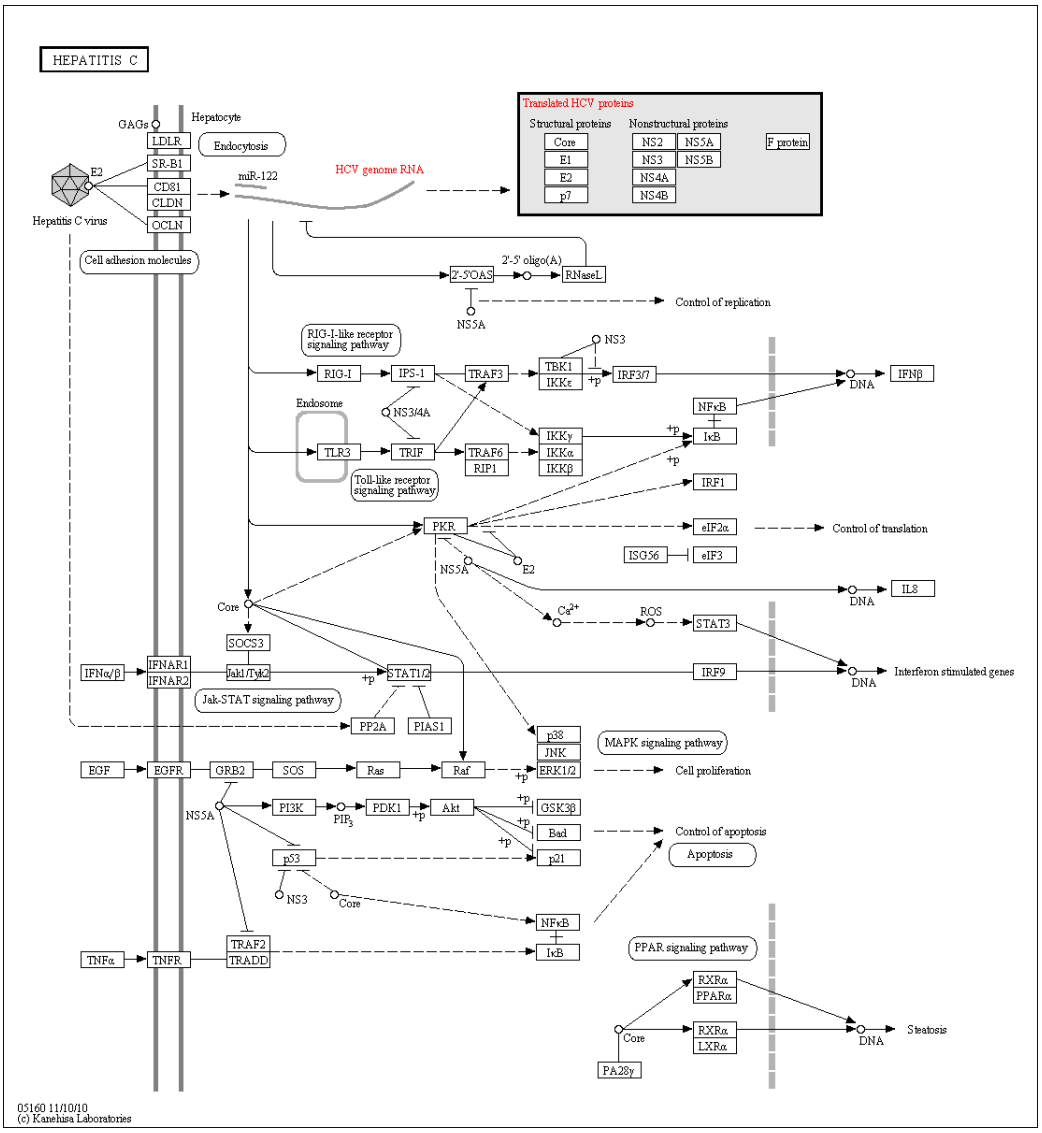

Figure 2.5: KEGG network representation of HCV replication [182]. This representation shows cellular interactions that occur during HCV infection. Though this network does not include every host factor that has been implicated in HCV infection, it does outline established and important aspects of host-virus interaction.

of interpretation and presentation in scientific work [98]. Indeed, software applications that perform clustering, be that for gene expression, network or general use, often provide results graphically as well as in the form of raw data, such as gCLUTO [179], BicACT [180] and jCLUST [181].

Dynamic and interactive visualisation techniques are frequently used to display biological networks, as these structures are difficult to envisage simply from source data. For example, resources such as the Kyoto encyclopedia of genes and genomes (KEGG) [182] and Reactome [183] are composite data sources that provide information on pathways and metabolic networks with corresponding visualisations, for reference material to allow users to better assimilate this information. In addition, both KEGG and Reactome provide such representations for relatively well established aspects of virus-host interaction data for HIV-1 [182, 183] and HCV [182] (figure 2.5).

KEGG and Reactome visualisations are provided in the form of a library of manually drawn, static networks. For creating user-defined, bespoke network visualisations, specific software applications are available. In addition to providing interactive network visualisations, these packages provide: (i) methods for analysis that are coupled with the visualisation; (ii) automated network layouts (an essential feature for all but small networks); and (iii) integration of secondary data to enrich the information content of the visualisations [184]. Figure 2.6 highlights the importance of network layouts and illustrates methods by which secondary data can be incorporated into network visualisations.

Software suites that are dedicated to network visualisation include Cytoscape, Osprey [186] and Pathway Studio [187]. Cytoscape [164] is a widely used open-source Java application. Cytoscape offers a wide range of viewing options, network layouts and visualisation of additional data. Cytoscape can render large networks of 100K nodes [184]. Perhaps the greatest benefit of Cytoscape is the wide array of available plug-ins, created by Cytoscape and third-party developers. These plug-ins fall into five categories; (i) network analysis of existing networks; (ii) I/O plug-ins for importing or exporting network data; (iii) network inference plug-ins for inferring new networks from data; (iv) communicating and scripting plug-ins and (v) functional enrichment plug-ins. In total there are over 35 plug-ins available for download, making cytoscape an incredibly versatile tool. Pathway Studio integrates data from several databases such as BIND [188], providing customizable network visualisations and a flexible SQL-like query language for filtering nodes and their attributes. However, Pathway Studio is not open-source software and is only available as a Microsoft Windows application, making it less accessible to the public than other resources, such as Cytoscape.

Biological networks are not necessarily just displayed simply as nodes and edges. Holden *el al.* [175] present a variety of network visualisation techniques that incorporate phylogenetic data and adjacency matrices: Terminal nodes of phylogenetic trees can be joined by arc-shaped edges to signify an interaction. Adjacency matrices can display networks using a grid system, where hubs and densely connected subnetworks are highlighted in by linear and block structures, respectively [175].

2.4 A role for computational biology in understanding host-virus systems

A major aim of studying pathogenic viruses, such as HIV-1 and HCV, is for discovery of novel, or improved antiviral therapies. An idealistic pipeline for achieving this aim can be summarised in three steps: (i) Identify factors (e.g., proteins) that are important in virus infection; (ii) Identify the relevant functional roles that those important factors have; and (iii) develop antiviral therapies that successfully exploit those factors or functions. Indeed, this is not only an idealistic but also an extremely simplified pipeline – there are many complicated parts in each of these steps and in reality this process is unlikely to be linear,

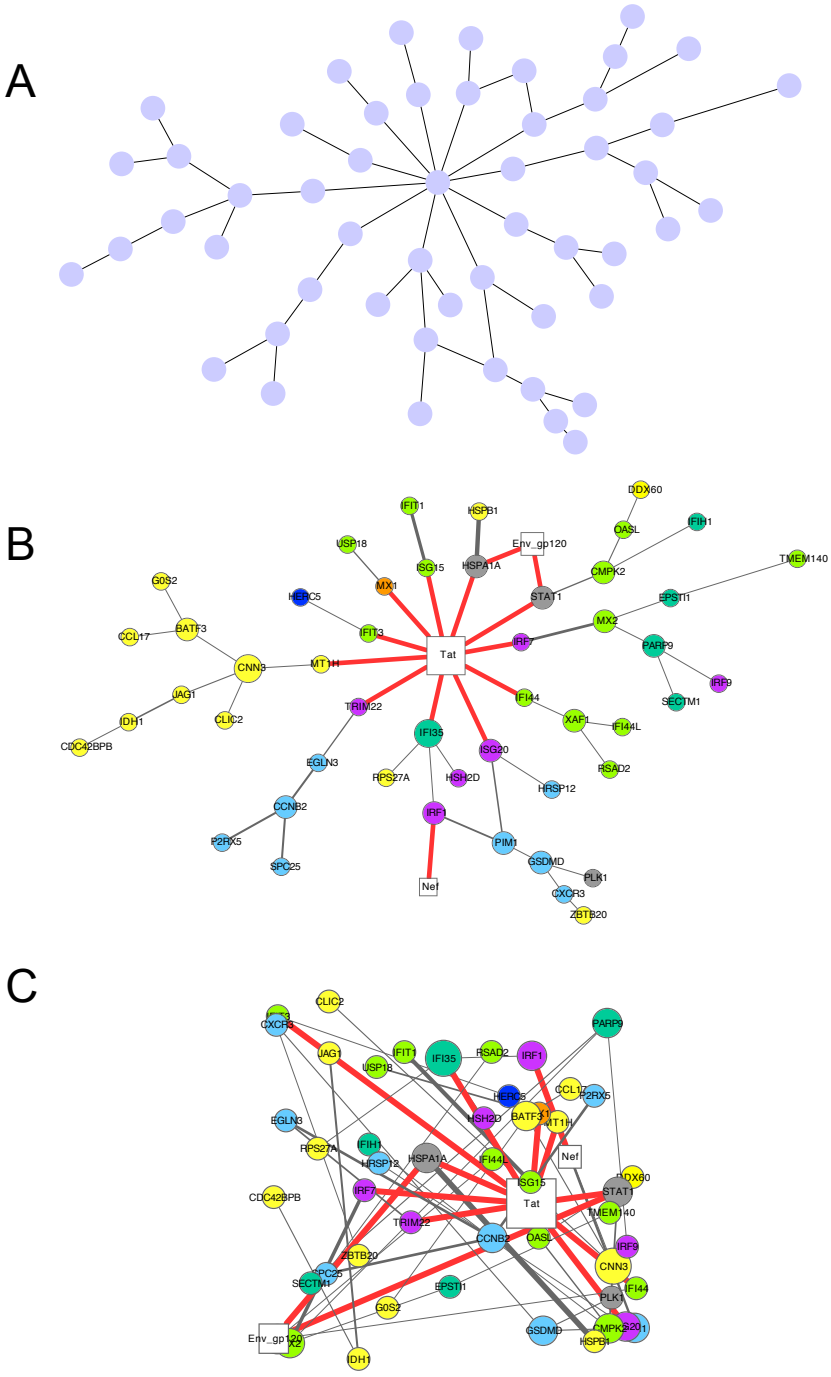

Figure 2.6: Network visualisation examples. The visualisations displayed all represent the same network. (A) A basic "ball-and-stick" network representation, where nodes are drawn as circles and edges are drawn as lines between the nodes. (B) An information-rich network visualisation. Here, nodes are labelled according to the biological entity that they represent. Both nodes and edges are subdivided into categories based both on their colour and shape (e.g., square nodes in this diagram represent HIV-1 proteins). Furthermore, edge widths and node sizes represent continuous data about nodes and interactions (e.g., node size represents the degree of the node). (C) A network where the nodes have been positioned at random. Very little information can be discerned from this network, highlighting the importance of the network layout. The networks in (A) and (B) are much clearer, having been laid out by a network layout algorithm and then manually edited. All visualisations were produced using Cytoscape [185].

as results from a later step are likely to inform knowledge of a previous step. However, contribution to this, or a not dissimilar scheme of discovery, is frequently proposed as the rationale behind many HIV-1 and HCV studies, e.g., [113, 126, 144, 175, 156, 26, 112, 154, 189].

The scientific community has become very adept at the first step in this pipeline of discovery: gene expression studies, that have been carried out for both HIV-1 [110] and HCV [111, 112, 113], can give a broad insight into host responses to viral infection and provide an insight into both virus replication and pathogenicity; siRNA screens have allowed detection of host factors that are essential for virus replication [144, 145, 146]; and PPI screens provide us with knowledge of the interface between the virus and the host cell [126]. Furthermore these technologies are possible on a whole-genome scale. In addition to large-scale experimental data, a vast amount of information is present in small-scale studies of virus-host interaction. Therefore, literature-curated resources, such as the HHPID, that catalogues HIV-1-host PPIs, are also hugely valuable to the research community.

The second step is much more difficult to achieve, as there is no universal assay to determine the function of a protein or gene. In fact, to determine the exact functions of a given gene or protein, bespoke laboratory assays are required and these can not be performed on a large-scale. Thus, the relevant function of many host factors linked to virus infection, identified by large-scale analyses, remains unknown. However, computational analyses are certainly not powerless, as they allow inference of active functions on a more general scale. Indeed it is commonplace for articles that describe results from large-scale experiments to have performed a meta-analysis of their results. Meta-analyses can help to discern relevant protein functions through integration of secondary data, often including functional annotation, PPI networks and results from other experimental studies. In addition, following the publication of new large-scale data sets, it is not unusual for additional computational meta-analysis to be performed, in order to draw together current data sets to gain a new perspective and insight into a biological system. A good example of such a computational meta-analysis performed on host-virus data was that by Bushman *et al.* [156], who successfully consolidate HIV-1 siRNA screen data to provide genuine biological insight into cellular components and subsystems that are essential for virus replication through use of PPI network data and biological annotation.

An alternative computational analysis of human-pathogen data was that performed by Dyer *et al.*, who perform a network based analysis of host-pathogen interactions. Their study includes data from both viral and bacterial systems (although the vast majority of this regards HIV-1, coming from the HHPID). Dyer *et al.* assess the graph-theoretical characteristics of pathogen-interacting human proteins and discover that they have a propensity for having high degree and betweenness coefficients. In addition, they identify enrichment for certain biological functions, such as involvement in the cell cycle. While the study by Dyer *et al.* was both original and interesting, providing a high-level

view of human-pathogen interacting proteins, the practical use of the results are somewhat limited. Firstly, the choice to pool data from more than one pathogen means that the study is not specific to a certain disease. Secondly, it is difficult to assess the biological relevance of graph-theoretical measures, such as degree and betweenness, especially when no correction is made for study-bias [136]. For example it is unclear from the study by Dyer *et al.* whether the pathogen-interacting proteins have a high degree due to the amount of studies that document their activities, or for a genuine biological reason. Although Dickerson *el al.*, taking study bias into account, show that HIV-1 interacts with cellular functions that are genuinely highly connected in the human PPI network [136].

In this thesis we perform integrative computational analyses using a variety of data types. Broadly, our aim is to provide specific biological insight into mechanisms that are important during HIV-1 infection and HCV infection. In order to obtain biologically significant results we use bespoke computational methods and incorporate multiple types of large-scale data. We intend that the following research chapters contribute to the overall knowledge and ability required to develop new or improved antiviral strategies.

CHAPTER

THREE

JNETS: EXPLORING NETWORKS BY INTEGRATING ANNOTATION

3.1 Abstract

JNets is a network visualization tool that incorporates annotation to explore underlying features of interaction networks. The software is available as an application and a configurable applet that provides a dynamic online interface to a wide variety of network data. As a case study, we use JNets to investigate approved drug targets present within the HIV-1 Human protein interaction network. Our software highlights the intricate influence that HIV-1 has on the host immune response.

3.2 Rationale

Interaction networks can be studied to gain a greater understanding of the biological system that they represent [190, 161, 162, 163]. A common method for studying interaction networks is through visualization, typically by representing a network as a 'ball-and-stick' graph [164, 187, 175, 191, 192]. Interactive visualizations can enhance our understanding of networks and allow new patterns and trends to be discerned [176], particularly when these tools offer network analysis capabilities. However, most published network visualizations are static representations that do not permit the user to view associated annotation let alone integrate other biological information in a useful manner.

The development of JNets was motivated by the need for an online, interactive protein-protein interaction (PPI) network viewer for the HIV-1, Human Protein Interaction Database (HHPID) [27, 26]. The HHPID is a valuable resource for the study of HIV-1 infection. HHPID data is manually curated and in addition to the pairs of interacting HIV and human genes, contains details of the interaction *type* (e.g., 'phosphorylates' or 'complexes with'). A tool is required that can be deployed from a website as an applet and that is also capable of useful network manipulation and analysis. This will aid in understanding the mechanism of infection and the host-viral interactions involved in the HIV-1 life cycle.

JNets thus has greater functionality than pure visualization software, such as InterView [175], but remains 'lightweight' and easy to use.

JNets is available as a stand-alone application and a web-deployable applet and is applicable to any type of biological or non-biological network data. Analysis in JNets is achieved by overlaying node and edge annotation on to the network. Groups of nodes and edges can be created by filtering accompanying annotation, and properties of groups can be explored, in terms of annotation, both visually and statistically. In addition, JNets is configureable to allow web-deployed visualizations to be customized by a vendor. Specifically, preset network visualizations can be defined and the JNets user interface altered. Furthermore, JNets is Java software, so is platform independent.

3.3 The JNets system

JNets is available in two forms: a stand-alone application and a web-deployable applet. The latter has some features disabled (such as the 'File menu') due to the security requirements of Java applets. Certain advantageous features in JNets were inherited from InterView [175], the software on which JNets is based. These include the animated spring layout, a container layout [175], interactive 'clickable' nodes and the facility to export network visualizations in PDF and PNG formats. In addition, the Java libraries responsible for graph layout, network rendering and the legend panel also come from InterView. InterView uses libraries from the TouchGraph package. These drive the interactive network display in JNets. The following sub-sections describe the main features of JNets in detail. Where appropriate, examples are given using network data from the HHPID [27, 26]. In addition, a diagrammatic summary that shows the organization of JNets is given in figure 3.1. JNets is available from http://www.bioinf.manchester.ac.uk/jnets, where an applet can be launched to visualize and browse the HHPID network. Also available at this site is a download package, including source code, documentation and example data file. The JNets software package is also supplied in supplementary data S3.1. The main JNets interface is shown in figure 3.2.

3.3.1 Subgroup Creation

An integral feature of JNets is the ability to edit and investigate subgroups of elements. In the other lightweight viewers available, JSquid [192] and InterView [175], subgroups are possible but are not dynamic, as they are predetermined in the input file. However, in JNets, users can create novel subgroups of elements using a simple, flexible system of filtering element annotation. To create a subgroup, three main components are considered. Firstly the input group. This can be the whole network or a previously created edge group, or node group. Note, the input group is the set of elements that will be filtered. Secondly, the element filter. There are two types of filter in JNets: *automatic* filters and *manual* filters. *Automatic* filters create an array of subgroups by taking the input

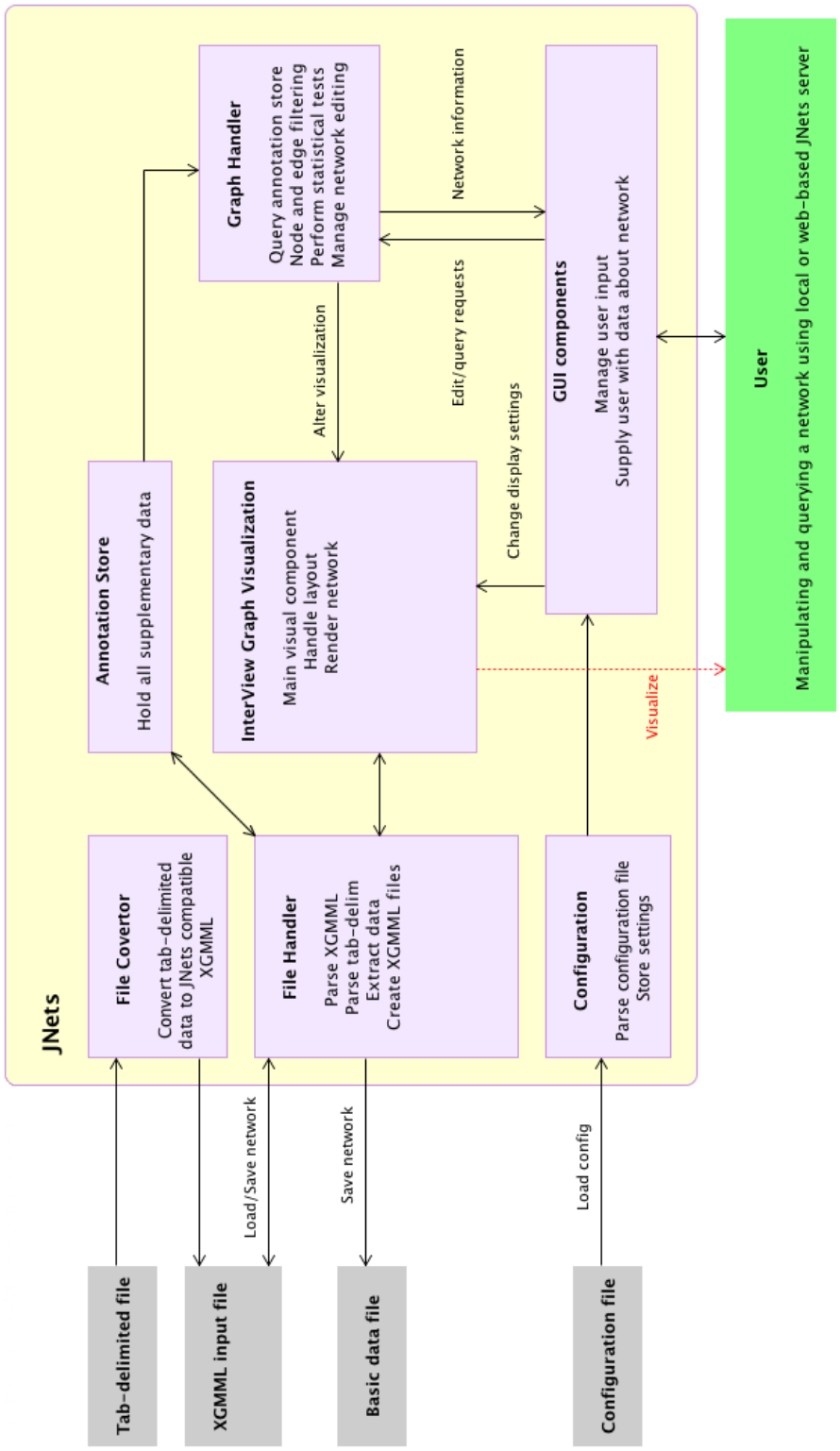

Figure 3.1: Diagrammatic representation of the the JNets system. his diagram describes the conceptual flow of information through the software.

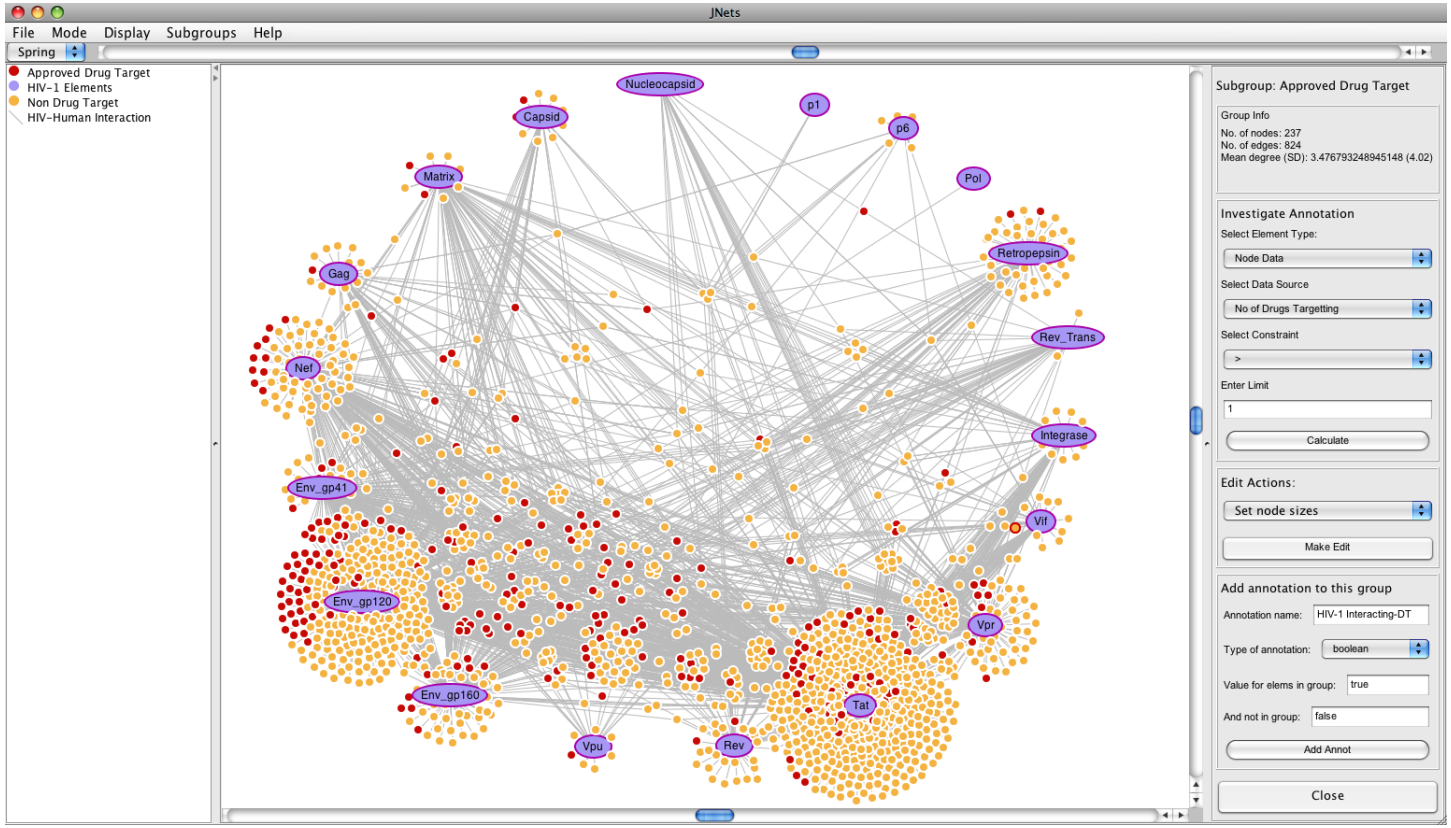

Figure 3.2: The JNets user interface. The main interface and network visualization panel from the JNets application is shown. JNets is displaying the HIV-1 human PPI network, from the HHPID[27, 26]. On the left is the legend panel showing the node and edge groups. The menu bar at the top is customizable. Standard drop-down menus can be enabled or disabled and new menus with user-defined network views can be added, defined in the configuration file. On the right the subgroup edit and analysis panel is shown, through which the user can tailor the visualization, explore subgroup annotation and add further annotation to the network.

group and dividing elements according to the value that they hold for a *single* annotation. For example, in the HHPID, proteins could be automatically filtered according to their taxonomic identifier. This would produce two subgroups of proteins: human and HIV-1. *Manual* filters operate like SQL select statements, where the input group is filtered according to any number of constraints concatenated with AND and OR operators, the output being elements that are true for that statement. For example, the interactions in the HHPID could be filtered to find those that are annotated as binding OR activation but NOT phosphorylation interactions. JNets also allows nodes to be filtered according to edge annotation (and vise versa) by taking into account edge incidences at each node. This greatly increases JNets flexibility in creating subgroups. For example, by this method, human *proteins* could be filtered according to whether they take part in an *interaction* that is annotated as binding OR activation but NOT phosphorylation. The third component is the output type. This can be nodes, edges or both nodes and edges. The output type determines the type of element that is subject to regrouping. For example, HIV-1 proteins could be filtered to find the HIV-1 accessory proteins Vpr and Vpu. Given an output type of nodes *and* edges this would result in the creation of a new node group consisting of Vpr and Vpu and a new edge group of interactions involving Vpr and Vpu. Therefore, JNets allows users great freedom to create subgroups of interest. The subgroup creation interface is shown in figure 3.3.

3.3.2 Network Manipulation

A key feature of JNets is the ability to manipulate the network to improve visualizations. Basic but very useful manipulations available in JNets include setting node and edge colors, setting node sizes, turning node labels on or off, collapsing parts of the network to form composite nodes, deleting elements and sticking nodes to prevent the spring layout from acting upon them. Most of these features were previously available in InterView, but acted upon the entire network only. With JNets these actions can also be applied to novel subgroups of nodes or edges to tailor the visualization to the requirements of the user.

3.3.3 Customizing JNets

JNets can also be customized by providing a configuration file. The configuration file is also in XML format. There are three main aspects of JNets configuration. Firstly, simple user-interface customization. Much of the user interface can be disabled including drop-down menus and right-click functions and the legend panel can be hidden or shown. The ability to create novel subgroups of elements can be disabled, as can specific parts of the subgroup edit and analysis panel. Secondly, the configuration file can determine what data JNets will use from the input XGMML file, in terms of nodes, edges and annotations. This configuration feature allows a single input file to be used, with a number of configu-

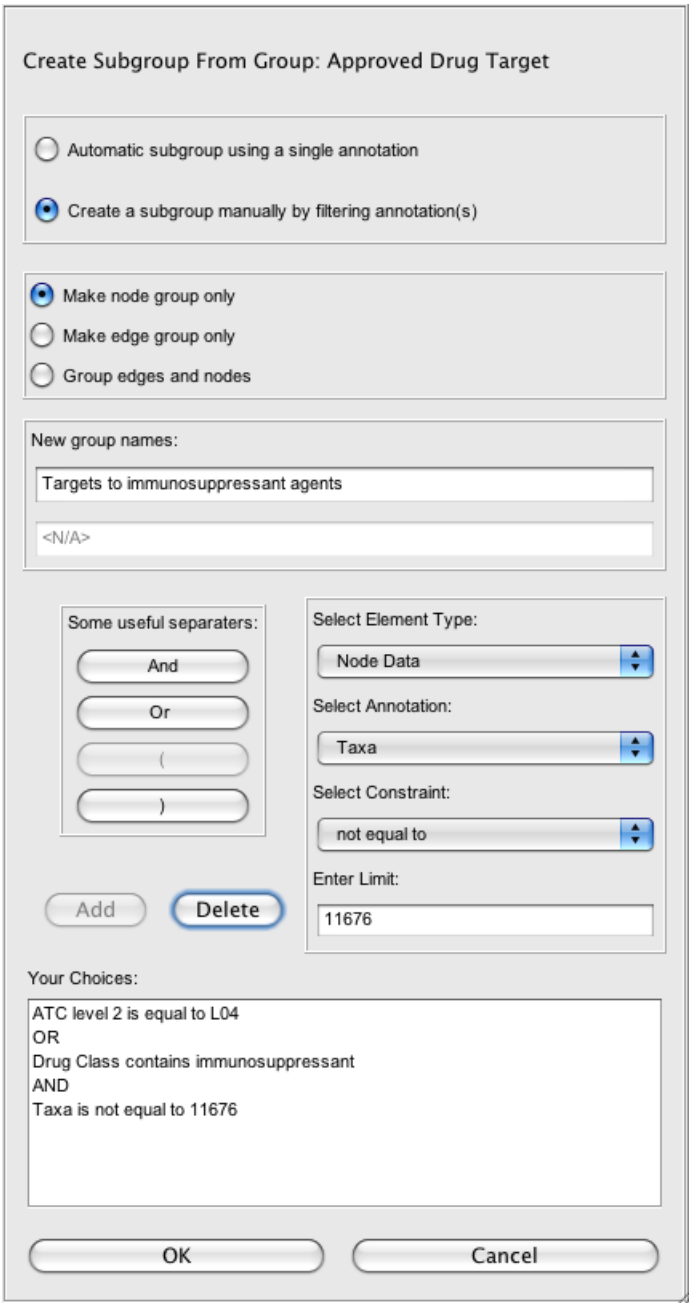

Figure 3.3: The JNets subgroup creation interface. This interface is used to create a new subgroups from already existing ones. The upper half of this panel is used to select some simple options about the subgroups being created, such as whether the grouping should be made automatically or manually, what the name of new subgroups should be and whether new a new node group, edge group, or both should be created. The lower half of this panel is used to select the annotation and set the filters through which new subgroups will be created.

rations, to produce a number of different networks. Thirdly, and most importantly, preset network views can be determined in the configuration file. Network views are defined in the configuration file as sets of annotation filters that encode the creation of new node and edge subgroups in the network. These presets allow groups of nodes and edges to be hidden, so that specific parts of the network can be highlighted. The presets appear in new drop down menus on the main menu bar and on a mouse click will execute the filters to alter the network view.

3.3.4 Network Analysis

In addition to visualization, JNets provides simple but powerful methods for analyzing networks and the annotations that are attributed to network elements. The main analytical capability in JNets is to examine subgroups of elements and assess annotation content using statistical significance methods. For example, using only the HHPID data and JNets, it is possible to create a subgroup of HIV-1-human interactions that involve the Tat protein and discover that there are 774 such interactions from a total of 2,588 unique PPIs. Furthermore, the most statistically detectable over-represented interaction type in this group is 'enhances', with a total of 53 interactions ($p = 3.3 \times 10^{-15}$), using all interactions as the general population, including a correction for multiple tests. To produce such results, JNets will either perform a two-tailed Fisher's Exact test, or Chi-Square approximation test, depending on the population size. The hypergeometric and chi-squared distributions required to do these tests are calculated using classes from the open source Colt 1.2.0 Java package. Where appropriate, correction for multiple significance testing is implemented automatically, using the Benjamini-Hochberg false discovery rate method [193].

This system of analysis allows JNets to remain both flexible (as both the annotation and the subgrouping used in the analysis are determined entirely by the user) and lightweight (as these calculations are computationally simple), particularly important for deployment as an applet. The following case study demonstrates the power of these analysis methods.

3.4 Case study: Drug targets in the HIV-host network

To demonstrate the utility of JNets as a network visualization and analysis tool, human genes that code for products that are both HIV-1 interacting and approved drug targets were examined, using the HHPID as a source for HIV-1-host interactions.

3.4.1 Case study: Introduction

To date, only one U.S. Food and Drug Administration (FDA) approved therapeutic agent, maraviroc, developed by Pfizer, directly targets a host, rather than a viral protein in the treatment of HIV-1 infection. Maraviroc is an antagonist of the CCR5 chemokine

receptor and prevents CCR5-tropic strains of HIV-1 from binding this co-receptor and entering host immune cells [77]. Other HIV-1 therapeutic agents may follow the lead of maraviroc and target host, rather than viral proteins, in the treatment of HIV-1 infection. For example, HMG-CoA reductase inhibitors (statins) are normally used to treat high cholesterol but have been shown to increase CD4+ cell counts and decrease viral load in HIV-1 infected patients. This action is thought to be due to the negative regulation of Rho GTPase activity by statins, impeding viral entry and budding from the cell [194], further evidence that targeting host proteins can disrupt HIV's biology. It is also possible to target HIV-host interactions directly. For example, disrupting the Vif-APOBEC interaction with a small-molecule, RN-18, is a strategy that is currently being pursued [195]. Vif mediates the degradation of the potent anti-retroviral APOBEC proteins (see review [196]) and in the absence of Vif, HIV-1 virions are non-infective [197].

In this case study we use JNets to explore the HIV-1-interacting proteins that are approved drug targets, as this may provide a useful insight for researchers investigating new therapeutic strategies to treat HIV-1 infection and, at the very least, indicate that a human protein that HIV interacts with is to some extent "druggable". Moreover, close examination of the drug classes and drug targets in this intersect may highlight ways that HIV-1 acts to perturb the host system. We consider both the details of the HIV-1 interactions and the types of drugs that target these human proteins.

3.4.2 Case study: Methods

HHPID data were downloaded from the National Centre for Biotechnology Information (NCBI) on May 1st, 2008. These data included the update to the HHPID that includes Env gene-product interactions (available November 13th, 2007). The HIV-1 interaction type was taken from the 'interaction' column of this data file. FDA approved drug and drug-target information was taken from the downloadable file 'drugcard_set.txt' from the DrugBank database [198, 199] on June 3rd, 2008. Drug categories were extracted from the 'Drug_Category' field and the Anatomical Therapeutic Chemical (ATC) classification system code was taken from the 'ATC_Codes' field of the same file.

A single HIV-1-host, drug-target network was created to conduct the investigation. The nodes in this network consisted of all FDA-approved drugs and all FDA-approved drug-target genes, all human genes that code for an HIV-interacting product and all HIV-1 elements (the term 'elements' is used because some of these are genes and proteins). Genes present both in the drug-target and HIV-1 interacting groups were only represented by a single node. There were two types of edges in this network: HIV-host interactions and drug-target interactions. HIV-host interactions came from the HHPID. There were 3,939 distinct interactions between the nodes of HIV-1 and the host, distinct on the HIV-1 element, the host gene and the interaction type. Therefore by this definition, interactions that are reported in multiple source articles, only count as one interaction. However, multiple interactions between the same nodes are possible as these may have significantly dif-

ferent interaction types, e.g., 'upregulates' and 'downregulates'. The nodes in the network were annotated, where relevant, with drug classifications and ATC codes, the number of distinct HIV-1 interactors and the number of HIV-1 interactions. The edges in the network were also annotated where relevant, with information derived from the HHPID, such as the interaction type and whether the interaction is agonistic, antagonistic or neither.

All results were produced by visualizing and analyzing the HIV-1-host, drug-target network using the JNets application.

3.4.3 Case study: Results

The HIV-1-host, drug-target interaction network contained 3,492 nodes and 7,374 edges (figure 3.4). Of the nodes, 19 were HIV-1 elements, 2,391 were human genes and 1,082 were drugs. Of the 2,391 human genes, 1,434 coded for HIV interacting products, 1,194 coded for approved drug targets and 237 coded for products that were both HIV-1 interacting and drug targets (throughout, we will refer to this gene set as HIDTs, figure 3.5). HIDTs account for 17% of human genes that code for HIV-1 interacting products and 20% of the genes that code for drug targets. Of these 237 genes, we found gene-products of 178 to have one or more direct physical interactions with HIV-1, based upon the interaction type. This network also contained 7,374 edges: 3,939 HIV-host interactions and 3,435 drug-target interactions. The whole HIV-1-host, drug-target interaction network is displayed in figure 3.4, where HIDTs are highlighted. A second network view is shown in figure 3.5 in which JNets has been used to filter the network to show HIV-1 interactions with HIDTs.

Of the 237 HIDTs, the gene-products from 114 interact with more than one distinct HIV-1 element. Using JNets, we showed that this is significantly greater than expected (Fisher's exact test, $p < 0.001$), given that from 1,434 human genes whose products are HIV-1 interacting, 529 have interactions with more than one distinct HIV-1 element. From 237 HIDTs, the gene products of 125 are targeted by more than one drug. By performing a Fisher's exact test on the drug-target network, we showed that this is a significantly greater proportion than expected at random ($p < 0.001$), given that 1,194 approved drug-target genes were identified in DrugBank and only 501 of these are targeted by more than one drug.

From 3,939 HIV-host interactions, 820 are between HIV-1 and HIDTs. We examined the HIV-1 elements that are responsible for these 820 interactions. We found significantly more Env-gp120 and Env-gp41 interactions and significantly fewer Tat and Integrase and Rev interactions among drug target genes, than would be expected, given the 3,939 HIV-human interactions as a parent population. See Table 3.1 for more details.

Next, we examined the drugs that target the products of HIV-1-Interacting-DTs, compared to other drug target genes in the human genome. We found that drugs from certain drug classes, according to level-2 ATC classifications denoted in DrugBank, are more likely to target gene products of HIDTs than would be expected, given all human drug

Table 3.1: HIV-1 interactions with approved drug target genes, by HIV-1 element. Here, we show the proportion of interactions that HIV-1 has with genes that encode drug targets, compared to total interactions, grouped by HIV-1 element. The statistical significance of the difference in expected and actual figures, indicated by p-values, were calculated in JNets, using Chi-square tests, corrected for multiple testing. We found significantly more Env-gp120, Env-gp41, Nef and Capsid interactions and significantly fewer Tat, Integrase, Rev and Vif interactions among drug target genes, than would be expected at random.

HIV-1 interactions with drug-target genes, by HIV-1 element			
HIV-1 element	Total interactions	Interactions with drug target genes	Corrected p-value
Tat	1394 (35%)	249 (30%)	0
Env-gp120	856 (22%)	239 (29%)	3.35×10^{-13}
Nef	311 (8%)	82 (10%)	4.42×10^{-2}
Vpr	275 (7%)	47 (6%)	Not significant
Env-gp160	213 (5%)	49 (6%)	Not significant
Env-gp41	190 (5%)	58 (7%)	1.14×10^{-2}
Rev	109 (3%)	7 (1%)	4.37×10^{-3}
Integrase	102 (3%)	5 (1%)	2.54×10^{-3}
Matrix	95 (2%)	15 (2%)	Not significant
Retropepsin	89 (2%)	15 (2%)	Not significant
Vif	77 (2%)	8 (1%)	Not significant
Gag	72 (2%)	15 (2%)	Not significant
Reverse transcriptase	45 (1%)	8 (1%)	Not significant
Capsid	45 (1%)	16 (2%)	Not significant
Vpu	27 (1%)	5 (1%)	Not significant
Nucleocapsid	26 (1%)	4 ($< 1\%$)	Not significant
p6	18 ($< 1\%$)	2 ($< 1\%$)	Not significant
p1	3 ($< 1\%$)	0 (0%)	Not significant
Pol	1 ($< 1\%$)	0 (0%)	Not significant

Table 3.2: HIV-1 interacting drug target genes, by drug category. Here, we show the proportion of drug target genes that are HIV-1 interacting, compared to the total number of drug target genes, grouped by level-2 ATC categories. We have only shown the top ten most statistically significant categories. The statistical significance of the difference in expected and actual figures,indicated by p-values, were calculated in JNets, using two-tailed Fisher's Exact tests, corrected for multiple testing. All categories are statistically over-represented in the HIV-1 interacting group, except for 'Vitamins', that are under-represented.

HIV-1 interacting drug target genes, by drug category			
Level-2 ATC description	Target genes	HIV-1 interacting	Corrected p-value
Immunosuppressive agents	36 (3%)	29 (12%)	1.86×10^{-13}
Antineoplastic agents	100 (8%)	47 (20%)	7.87×10^{-9}
Anti-inflammitory and antirheumatic products	27 (2%)	18 (8%)	3.02×10^{-6}
Stomatological preparations	30 (3%)	19 (8%)	3.35×10^{-6}
Lipid modifying agents	46 (4%)	23 (10%)	4.62×10^{-5}
Antithrombotic agents	65 (5%)	29 (12%)	4.08×10^{-5}
Drugs for obstructive airway diseases	25 (2%)	13 (5%)	3.44×10^{-3}
Vitamins	91 (8%)	6 (3%)	5.38×10^{-3}
Immune sera and immunoglobulins	10 (1%)	7 (3%)	$6.81 \times 10^{-}3$
Immune sera and immunoglobulins	10 (1%)	7 (3%)	$6.81 \times 10^{-}3$

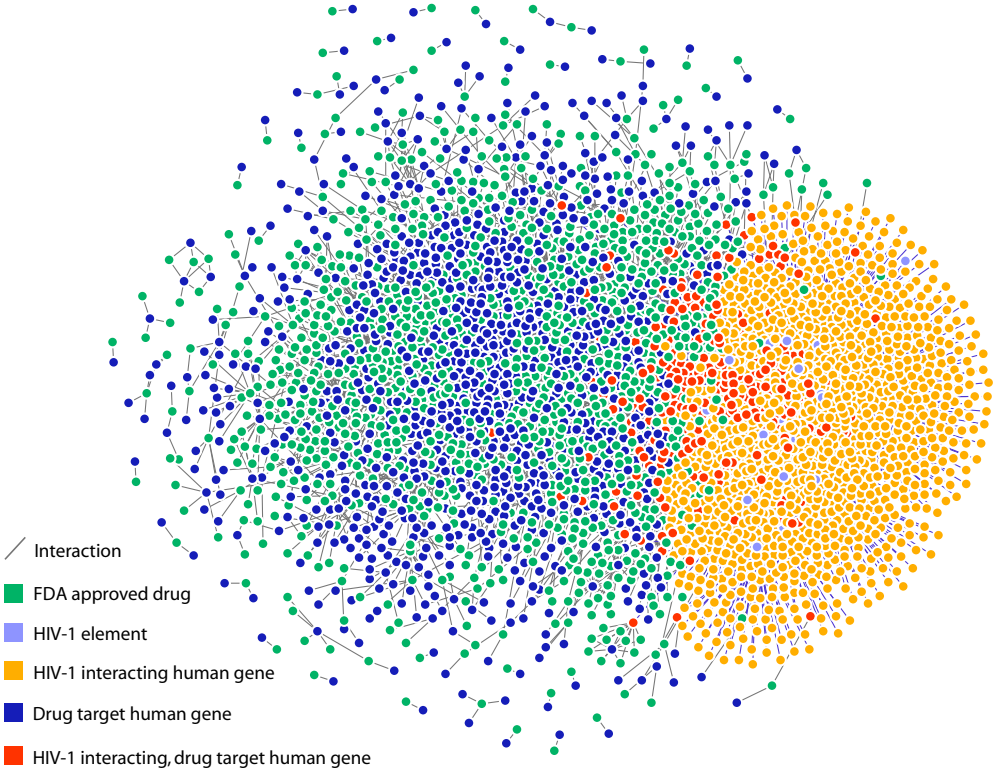

Figure 3.4: The HIV-1-host, drug-target interaction network. This is the whole network that was used for all analyses in our case study. Using the annotation that accompanies nodes and edges, JNets can filter this network to create more focussed visualizations, for example, the networks in figures 3.5-3.7.

target genes in DrugBank as a parent population. In total, we identified six such classes that showed over-representation ($p < 0.001$, Table 3.2). We investigated three particular over-represented groups of genes in greater detail: (i) HIDTs that code for gene-products targeted by immunosuppressive agents, this was the most overrepresented drug category ($p = 1.87 \times 10^{-13}$); (ii) HIDTs that code for gene products targeted by anti-neoplastic agents, the second most over-represented category ($p = 7.87 \times 10^{-9}$); and (iii) HIDTs that code for gene-products targeted by statins – a specific subset of the category 'lipid modifying agents'. To identify as many HIDTs as possible whose products are targeted by these three drug types, we widened our search to take into account level-2 ATC codes *and* drug classifications from DrugBank. Immunosuppressants were defined with drug classification 'immunosuppressive agents', or, with the corresponding ATC level-2 classification 'L04'. Antineoplastic agents were defined with a drug classification containing the word 'antineoplastic', or, with the corresponding ATC level-2 classification 'L01'. Statins were defined with a drug classification of 'HMG CoA reductase inhibitors' or 'Hydroxymethylglutaryl-CoA Reductase Inhibitors', or, one of the level-3 ATC classifiers: 'C10AA' (HMG CoA reductase inhibitors), 'C10BA' (HMG CoA reductase inhibitors in combination with other lipid modifying agents) or 'C10BX' (HMG CoA reductase inhibitors, other combinations). After this redefinition, all three chosen

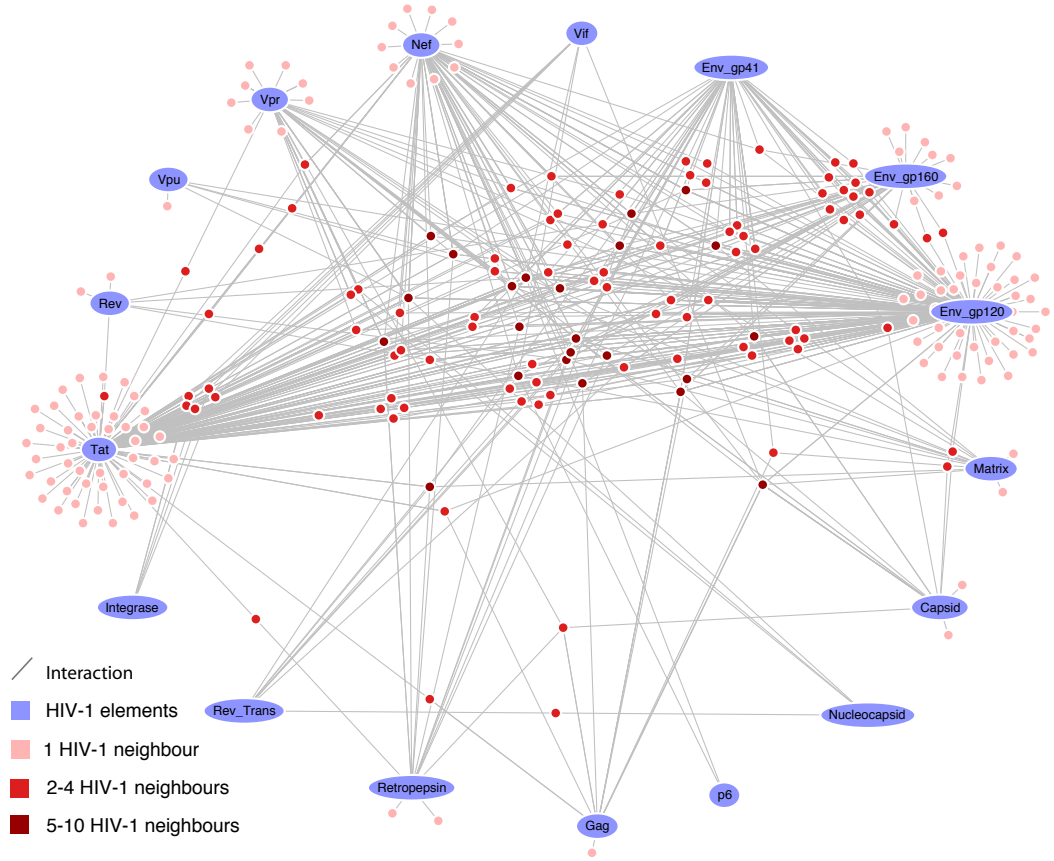

Figure 3.5: HIV-1 interacting drug target genes. This network shows 237 human genes that encode products that are HIV-1 interacting and FDA-approved drug targets. HIV-1 elements are labelled. Human genes are colored according to the number of distinct HIV-1 elements with which they share an interaction (darker = more). 114 from 237 of these human genes interact with more than one HIV-1 element. MAPK1 (mitogen-activated kinase-1), interacts with 10 different HIV-1 elements in 23 distinct interactions. The layout of this network was achieved by manually positioning the HIV-1 nodes, locking them in position and allowing the JNets spring layout to reposition the remaining human gene nodes. As a result, human gene nodes with multiple interactors are drawn to the centre of the network.

drug groups were still statistically over-represented among drugs targeting the products of HIDTs, compared to other human drug target genes ($p < 0.001$).

We found 21 HIDTs whose gene products are a target for HMG-CoA reductase inhibitors. This is 66% of all statin target genes given in DrugBank. This group shows a relatively high degree in the HIV-1-host network; the average degree of a host node is 2.8 but for this group of nodes the average degree was 6.7. From these 21 genes, 13 code for products that interact with Nef (over-represented where $p < 0.001$), 15 with Env gp120 and five with the Matrix protein (both over-represented where $p < 0.05$). In addition, these 21 HIDTs have certain over-represented interaction types with HIV-1 ($p < 0.05$); HIV-1 is activated by three, induces the release of five and upregulates the gene-products of 12 HIDTs in this group. Among these 21 HIDTs were the HMG-CoA reductase gene, whose gene-product is modulated by Nef and the ras homolog, member A (RhoA), whose

gene-product is upregulated by Vpr and activated by Tat.

We found 73 HIDTs whose gene-products are a target for 81 different antineoplastic agents. This is 43% of all antineoplastic agent target genes given in DrugBank. From these 73 genes, the notable HIV-1 elements that their gene-products interact with are Env-gp120 (41 genes) and Nef (23 genes), which are over-represented ($p < 0.05$). From this group, no particular HIV interaction type was found to be statistically over-represented, after correcting for multiple tests. However, the most abundant classes were 'interacts with' (25 genes), upregulates (24 genes), activates (21 genes), inhibits (19 genes), binds (19 genes) and downregulates (12 genes). The 81 antineopastic agents that target the gene-products of HIV-1-Interacting-DTs were from a wide range of ATC classifications. From those with level 2 ATC class 'L01', corresponding specifically to antineoplastic agents, the most abundant were drugs from the 'L01X' level 3 ATC class, corresponding to 'other antineoplastic agents'. Therefore, a specific group of antineoplastic agents that target the same gene-products as HIV-1, was not easily identifiable.

Given the nature of HIV-1 infection, the immune cells that become infected and the detrimental effect that this infection can have on the immune response of the host, we expect to see a high proportion of immune system proteins among HIV-1 interacting drug targets. Indeed we found 35 genes in the HHPID whose products are a target for immunosuppressants. This is 63% of all immunosuppressant target genes given in DrugBank. From these 35, the outstanding HIV-1 elements that they interact with, as for antineoplastic agent target genes, are Env-gp120 (25 genes) and Nef (15 genes), which are over-represented ($p < 0.001$). However, the products of these genes also interact with Env-gp41, 10 genes, and the capsid protein, four genes ($p < 0.05$). From this group of 35 genes, HIV-1 acts to downregulate the products of 16 (over-represented, $p < 0.001$). This was the only statistically significant HIV-1 interaction type detectable among this group of genes after correcting for multiple tests. However, HIV-1 also inhibits 12, upregulates 15 and activates the products of five genes from this group. Figure 3.6 shows the drug-target network, filtered to show only HIV-1-interacting immunosuppressant target genes, the immunosuppressive drugs and the HIV-1 elements that target the products of these genes. The human genes are colored according to the general action that HIV-1 infection has upon their products. Three categories are defined: agonised (e.g., activate, upregulate and enhance), antagonised (e.g., inhibit, downregulate and degrade), both agonised and antagonised, or neutral/unspecified (e.g., binds, modulates and interacts with). The agonised group contained 7 genes, the antagonised 11 genes, the agonised and antagonised 12 genes and five neutral genes, respectively. We found 25 immunosuppressive drugs that target gene products that are HIV-1 interacting, 18 of these drugs have more than one such target. The majority (14) of these 25 drugs have a level 4 ATC classification 'L04AA', meaning that they are selective immunosuppressants. Figure 3.7 shows the same results as in figure 3.6, however, for clarity, the drug nodes and human genes from the neutral group have been removed.

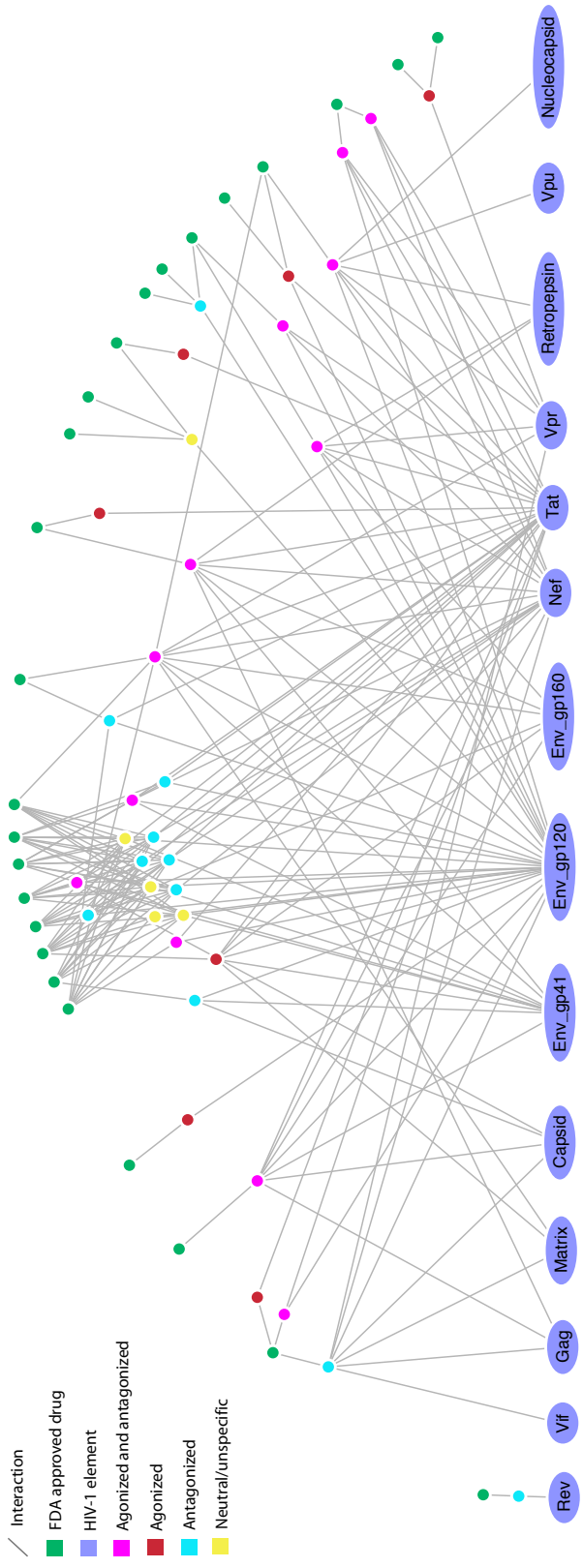

Figure 3.6: Drug-target network showing all immunosuppressant target HIV-1 interacting genes. The drug nodes lie in an arc around the top of the network. Between these two groups are the human gene nodes, colored according to the action HIV-1 has upon them. Four types of action are defined: agonised (7), antagonised (11), agonised and antagonised (12) and neutral/unspecified (5). These distinctions are derived from the interaction description supplied with each HHPID HIV-1-host interaction. HIV-1 elements are shown at the bottom and are labelled. The human genes shown in this network are likely to perform a significant role in the immune system to be targeted by immunosuppressive drugs. HIV-1 targets many of the same proteins as immunosuppressant drugs.

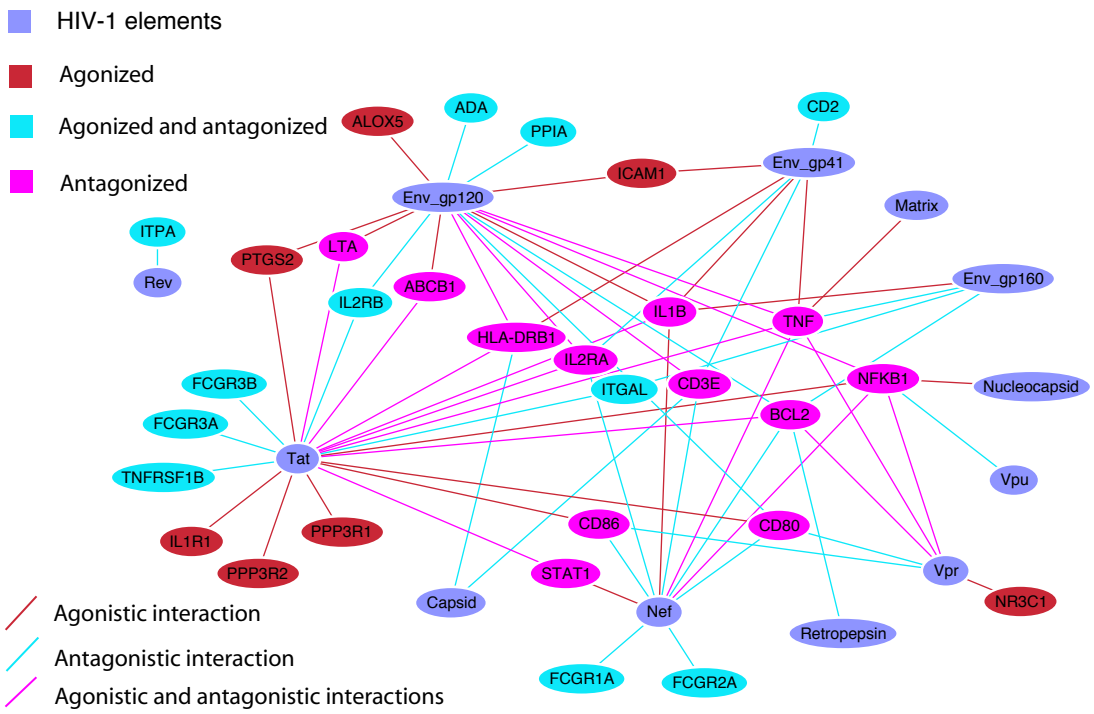

Figure 3.7: HIV-1 host network showing immunosuppressive agent target genes. Human genes that are both HIV-1 interacting and targeted by immunosuppressive agents are shown. Only those genes that are explicitly agonised or antagonised by HIV-1 have been included; human genes are colored according to this action. Seven host genes are agonised, eleven are antagonised and twelve are both agonised and antagonised. These distinctions are derived from the interaction description supplied with each HHPID HIV-1-host interaction.

3.4.4 Case study: Discussion

As an example of the use of JNets, we have taken data from the HHPID and DrugBank to research the cross section of genes that encode gene products that are both HIV-1 interacting and drug targets. We have shown that a significant proportion (237 genes, 17%) of genes that encode an HIV-1-interacting product also encode an approved drug target (HIDTs), of which 178 code for products that have direct physical interactions with HIV-1. In addition, a significant proportion of these interact with more than one HIV-1 element. This suggests that products of these genes are involved in important interactions with HIV-1, rather than being incidental effects of HIV-1 infection. Thus, drugging HIV-1 interacting proteins could disrupt the HIV-1 life cycle. These host proteins could be explored in the search for new HIV-1 treatments.

Many of the HIV interactions involving HIDT gene products are with proteins from the Env complex, Nef and Tat. Tat, the *trans*-activator of transcription, promotes viral transcription and elongation [200]. As a consequence, Tat induces operation of the host transcription machinery and indirectly, the production of other viral proteins. Therefore, as a result of some direct interactions, Tat is also responsible for indirect, downstream

responses in the host cell, many of which will be more directly due to the activity of transcription and the presence of other viral proteins. Moreover, Tat interactions were actually underrepresented among HIDTs. Therefore, it may not be Tat interacting human proteins that are most interesting from a drug-discovery perspective. In contrast, interactions with the Env complex were found to be over-represented among HIDTs. Env proteins and Nef are known to interact with host proteins located in the membrane. For example, the Envelope complex binds CD4 receptors and several co-receptors including CCR5 [201]. Nef, or negative factor, is known to downregulate CD4 [202] and class I MHC molecules [203]. Around 70% of drugs are believed to target membrane proteins [204], which may explain the prominance of Nef and Env interacting drug targets and the over-representation of Env interacting drug targets in our network.

Using JNets, we have shown that human genes that code for products targeted by HMG-CoA reductase inhibitors, antineoplastic agents and immunosupressive agents are over-represented in the HIV interaction data. HIV-1 interacts with the majority of targets of HMG-CoA reductase inhibitors, including HMG-CoA reductase itself. The HIV-1 virion requires clustering of host lipid 'rafts', also known as DIGs (detergent-insoluble glyco-lipid-enriched microdomains), for entry to and budding from the host cell [194]. Raft formation is believed to be controlled by Rho GTPases and remodelling of the actin cytoskeleton. HMG-CoA reductase is required to prenylate and activate these GTPases. Nef is thought to associate with these rafts and prime T-cells for activation and may promote HIV-1 replication [205]. However, due to the extensive crossover we have observed between the targets of HMG-CoA reductase inhibitor and HIV-1 interacting proteins, it seems likely that other mechanisms are active that give HMG-CoA reductase inhibitors the ability to decrease viral load in infected patients. One indicator of this is that the HMG-CoA inhibitor target genes have a particularly high degree in the HIV-1-host protein interaction network, suggesting that they are particularly important in the HIV-1 life cycle. Env-gp120 has interactions with the products of 15 of these genes, which is unsurprising when we consider that Env associates with many host membrane proteins and that these 15 host genes are likely to be involved in membrane raft formation and HIV-1 entry and exit from the cell. One notable interaction that Env has with a host protein is with coagulation factor II (thrombin). Thrombin has been shown to activate the Env complex and enhance fusion of the virus to the host cell [206]. Thrombin is also a target to the HMG-CoA reductase inhibitor simvastatin – a possible mechanism through which statins may be effective drugs in the treatment of HIV-1 infection.

From our results it is not clear why antineoplastic agents and HIV-1 share many human targets. By manually examining the group of drugs that target the products of these genes, it was noted that many are monoclonal antibodies, such as Trastuzumab (commonly known as Herceptin), that bind cell surface receptors. Trastuzumab action is believed to be involved with the mitogen-activated protein kinase (MAPK) cascade [207], a pathway with which HIV-1 infection is also highly associated, for example, the MAPK-

1 gene has 22 distinct interactions with HIV-1 in our network, with 10 distinct HIV-1 elements. The systemic effects of antineoplastic agents and of HIV-1 infection may utilise many of the same intrinsic cellular pathways. AIDs-related malignancies are not an uncommon cause of death for HIV-1 infected patients [208, 209]. This suggests that research into the use of antineoplastic therapies in HIV-1 infected patients, may benefit from greater utilization of network-based approaches. Such approaches may lead to improved treatment strategies that avoid provoking greater levels of infection and virus-induced host cell perturbation. For example, to better understand crossover in specific cellular pathways, host proteins could be annotated with pathway data, points of contact with viral proteins, and drugs could then be identified and the network analysed further using JNets.

Our analysis has shown that the majority of the targets of immunosuppressive drugs are also HIV-1 interacting, an observation that we have shown to be statistically significant. In the case of HIV-1 infection, interaction with elements from the host immune response is particularly likely, as HIV infects cells that are specific to this host system, such as T-cells and macrophages. Therefore, HIV-1 requires a sufficient host immune response so that it has cells to infect but may also downregulate elements of the host immune response at the same time, presumably in order to evade other aspects of the immune system. This complexity is exemplified in our results. Figure 3.7 shows that HIV-1 elements, particularly Nef, Tat and Env, act to both agonise and antagonise proteins that are intrinsic to the host immune response. For example, there is no clear relationship that defines HIV-1 action on interleukin receptors (i.e., IL2RA, IL1RA, IL2RB). This could be due to a single, static network view being insufficient to expose variable aspects of HIV-1 infection, such as the point in the cell-cycle, the stage of HIV-1 infection and the infected cell type – it is inconceivable that these interactions all occur simultaneously in a single cell. Even so, this representation does suggest that HIV-1 possesses an intricate system for influencing the host immune response that is necessary to maintain the infection.

3.5 Evaluating JNets

In our case study, we have integrated data from DrugBank and the HHPID to perform analyses. We have demonstrated that JNets is capable of combining biological annotation and interaction network data to allow specific statistical conclusions to be drawn, from which biological inferences can be made. However, JNets is not designed to compete with heavily developed network analysis packages, such as Cytoscape [164] and Pathway Studio [187]. These applications incorporate a plethora of functions for the analysis of biological networks. A drawback of such 'heavyweight' applications is that the utilities on offer may fall beyond the scope of many users as adding greater functionality to software can complicate user interfaces and make simple tasks less accessible. Rather, JNets is designed to allow network visualization and some useful analysis to be carried out in a

simple, web-deployable tool. By this system, vendors could use JNets as a 'gateway' to allow users to interact with, and better understand, their own network data.

No part of JNets is specific to biological networks, making JNets useful across a wide range of disciplines. JNets is also configurable, to allow preset network views and subdivisions of networks to be stored. This aspect of the applet makes use of the JNets annotation filtering method, so that very different visualizations can be rendered from a single input file, available to the user at a mouse click. Much of the JNets system is based around the dynamic creation of subgroups by filtering network annotation. By offering these features, JNets is quite different from any other web deployable network visualization tool that is currently available. For example, WebInterViewer is designed to render a large number of nodes with great efficiency, although annotation can accompany the network [191]. JSquid [192] was primarily designed as an interactive viewer for the FunCoup website (http://funcoup.sbc.su.se), though it can be used for viewing independent network data. JSquid also makes use of node and edge groupings, however, unlike JNets does not perform statistical analyses on user defined features of these groups.

In summary, JNets is a platform independent, interactive network visualization and analysis tool, suitable for a wide range of network data and capable of being deployed from a website as a customized applet or run as a stand-alone application. For further information about JNets please visit http://www.bioinf.manchester.ac.uk/jnets.

3.6 Supporting material

Supplementary file S3.1 – the JNets software package. This folder contains both JNets documentation and software.

CHAPTER

FOUR

PATTERNS OF HIV-1 PROTEIN INTERACTION IDENTIFY PERTURBED HOST-CELLULAR SUBSYSTEMS

Jamie I MacPherson, Jonathan E Dickerson, John W Pinney and David L Robertson

4.1 Abstract

Human immunodeficiency virus type 1 (HIV-1) exploits a diverse array of host cell functions in order to replicate. This is mediated through a network of virus-host interactions. A variety of recent studies have catalogued this information. In particular the HIV-1, Human Protein Interaction Database (HHPID) has provided a unique depth of protein interaction detail. However, as a map of HIV-1 infection, the HHPID is problematic as it contains curation error and redundancy, in addition, it is based on a heterogeneous set of experimental methods. Based on identifying shared patterns of HIV-host interaction, we have developed a novel methodology to delimit the core set of host-cellular functions and their associated perturbation from the HHPID. Initially, using biclustering, we identify 279 significant sets of host proteins that undergo the same types of interaction. The functional cohesiveness of these protein sets was validated using a human protein-protein interaction network, gene ontology annotation and sequence similarity. Next, using a distance measure, we group host protein sets in order to identify 37 distinct higher-level subsystems. We further demonstrate biological significance of these subsystems by cross-referencing with global siRNA screens that have been used to detect host factors necessary for HIV-1 replication, and investigate the seemingly small intersect between these data sets. Our results highlight significant host-cell subsystems that are perturbed during the course of HIV-1 infection. Moreover, we characterise the patterns of interaction that contribute to these perturbations. Thus, our work disentangles the complex set of HIV-1-host protein interactions in the HHPID, reconciles these with siRNA screens and provides an accessible and interpretable map of infection.

4.2 Introduction

Acquired immunodeficiency syndrome (AIDS), caused by HIV-1, is responsible for millions of deaths every year. Therefore, research into HIV-1 biology is of critical importance and research efforts are significant and ongoing. In order to replicate, HIV-1, like all viruses, must use host-cellular machinery and induce production of viral genomic material, viral proteins and ultimately new virions. This hijack and control over host cell processes is mediated by HIV-1 proteins through a complex network of molecular events, including virus-host protein-protein interactions (PPIs) [174]. Therefore, by developing our knowledge of the virus-host interaction network, we can improve our current model of HIV-1 infection and host-cell perturbation and use this information to aid development of new antiviral treatments. One example of a successful antiviral treatment that has come from understanding HIV-host cell interaction is the drug maraviroc [77]. Maraviroc is an entry-inhibitor that binds the CCR5 co-receptor, inhibiting gp120:CD4:CCR5 complex formation and thus entry into the host cell. Targeting a host protein in this way demonstrates that the number of possible HIV-1 therapeutic drug targets is not limited to the small viral proteome and that understanding the virus-host interface can lead to the development of novel-acting therapeutic agents.

Our knowledge of HIV-1-host PPIs is extensive in relation to other pathogens [210]. A major source of HIV-1-host protein interaction data is the HIV-1, Human Protein Interaction Database (HHPID) [26, 27, 174]. This database holds over 5000 interactions involving over 1400 human proteins, curated from primary literature on small-scale protein interaction studies. In the HHPID, an impressive level of detail is recorded, including a short description of each interaction outcome, e.g., 'phosphorylates', 'binds', 'activates' etc. However, there are several problems associated with this data set:

(i) Interactions in the HHPID come from a large number of separate publications over a wide date range and are derived from a diverse array of experimental procedures, such that the quality of the data is varied and the proportion of false-positive interactions, though presumably minimised by the small-scale nature of the contributing works, is difficult to estimate.

(ii) The manual curation step introduces a potential for inconsistency and some anomalies have been identified [211].

(iii) The database contains a large amount of redundant data, where the same interaction has been reported more than once in two separate records. For example, in the HHPID there are 27 entries describing interaction between the HIV-1 Tat protein and the human CDK9 protein, including five that describe binding and five more that describe complexing, two describing activation and three describing stimulation, although from these data it is not clear whether more than one interaction actually occurs.

(iv) A second level of redundancy exists due to downstream consequences of interactions. For example, the finding that HIV-1 gp120 interaction with CD4 alters the activity of transcriptional regulators and cytokine transcription [212] is present as nine entries in

the HHPID, when this activity can be explained through a direct interaction at the cell surface, causing downstream effects in the T cell receptor signaling pathways. However, by simply taking direct interactions from this database to determine host cell perturbation, important regulatory effects may not be considered or alternatively, perhaps falsely extrapolated.

Due to these reasons, while the HHPID is a unique, detailed source of individual PPI interactions that represents a large proportion of the knowledge in the published literature, it does not immediately provide a logical and functional map of HIV-1-host interaction.

Recently, three high-throughput HIV-human protein interaction data sets have been published that are the result of individual genome-scale siRNA gene knockdown screens [144, 145, 146]. These studies each identify over 200 host-cellular factors that are necessary for HIV-1 replication, termed 'HIV-dependency factors' (HDFs) [144]. A thorough meta-analysis of HDFs has been performed by Bushman *et al.* [156]. Though the pairwise intersection of genes between the three sets of HDFs is statistically significant in all cases [156], the number of genes confirmed by more than one study is only 34 and just three genes are present in all sets. This seemingly small overlap is largely thought to be due to differences in experimental procedure, including cell-type, choice of time points analyzed and choice of filtering thresholds [146, 145, 174, 156]. Despite the apparent small overlap between HDF sets, Bushman *et al.* demonstrate that certain cellular subsystems are mutually identified, such as DNA repair and nuclear transport associated proteins. This indicates the validity of the screen results and the value that can be gained by combining these data to identify essential host-cellular functions required by HIV-1 for replication. In addition, their study shows that intersections between HHPID data and the HDF sets, while significant are quite small at 39 [144], 54 [145] and 39 [146] genes. However, while the work of Bushman *et al.* successfully consolidates information between HDF sets and validates these sets against the HHPID, the underlying differences between the HHPID and siRNA screen results have not been explored in detail. In particular, cellular subsystems prevalent in the HHPID, but not present among HDFs, have not been identified.

Our previous visualisations of the HIV-human PPI network show that there are noticeable clusters of host proteins that take part in multiple interactions with the same set of HIV-1 proteins [26, 213]. These groups possibly represent multiple interactions with biologically related proteins, e.g., from functional pathways or protein complexes. In addition, highly connected subnetworks of host proteins, where some proteins are involved in multiple HIV-human interactions, have also been identified using a combination of human-human PPI data and HIV-human interaction data [145, 156]. These subnetworks represent specific biological activities including the ubiquitin-proteasome pathway, transcription, nucleic acid binding and nuclear import – all thought to be important in facilitating the early stages of HIV-1 infection [145]. However, in all of these studies, different types of HIV-1-host interaction are not taken into consideration in the clustering method,

despite the potential for interactions to be quite dissimilar. For example, subnetwork PPIs may include direct binding interactions, indirect regulatory interactions and those with opposing actions, e.g., inhibition and activation, such that no systematic outcome is identifiable.

In this work, we explicitly utilise host-virus interactions and interaction types, as provided in the HHPID, to identify significant patterns of viral perturbation of the host cell. This permits us to gain meaningful insights into HIV-1 infection. Specifically, using a biclustering approach, we define sets of host proteins that take part in a common set, or 'profile', of HIV-1 interactions. Using a distance method to cluster these units, we identify higher-level groupings. We show that these higher-level groups of proteins map to specific biological subsystems in the host cell. By considering patterns of interaction with host cell proteins, evidence within primary literature and by assessment of support from global siRNA screens, we are able to infer the biological importance of these subsystems in terms of HIV-1 replication, host cell perturbation and regulation of the immune response. Thus, our work extracts a coherent functional map of core HIV-1-host interactions from the HHPID and consolidates findings from the major HIV-1-host PPI data sets.

4.3 Materials and methods

4.3.1 Data collection

HIV-human PPI data were obtained from the HHPID on 1st May 2009. Specifically, distinct PPIs, based upon (i) the HIV-1 protein interactant, (ii) the human protein interactant and (iii) the type of interaction, one of 68 short descriptions that characterise the PPI outcome, were obtained [27, 26]. In cases where multiple transcripts of the same gene take part in the same interaction (with respect to HIV-1 protein and interaction type), only a single instance of the transcribed gene and interaction were used throughout our analyses.

To test whether interaction types are uniformly distributed among HIV-1 proteins, interaction types for each HIV-1 protein were counted and p-values were calculated using two-tailed Fisher's exact tests and corrected for multiple tests using the Benjamini and Hochberg [193] method.

4.3.2 Bicluster identification

In order to perform biclustering, a binary matrix was created with one row per human protein and one column per HIV-1 interaction. We define an HIV-1 interaction to include both the HIV-1 interactant and the interaction type, e.g., 'capsid activates' is one such interaction. The presence of a given HIV-1 interaction, for a given protein, was represented in the matrix by a one and the absence by a zero (figure 4.2). To find sets of

human proteins that share the same set of HIV-1 interactions in this matrix, biclustering was performed using the Bimax algorithm [214]

The significance of biclusters was determined by Monte Carlo simulation. Specifically, the HIV-human PPI network was rewired at random, while the degree of each protein and interaction type frequencies were maintained. The resulting network was used to produce a new matrix for biclustering. The matrix was biclustered using Bimax and interaction types, HIV-1 proteins and the number of human proteins in each bicluster were recorded. 50 000 iterations of this process were carried out. Using these simulations, we were able to empirically calculate the probability of randomly finding a bicluster involving a given number of human proteins and the same (or larger) set of HIV-1 interactions. Biclusters were deemed significant if they had a p-value of < 0.001, after correction for multiple tests using the Benjamini and Hochberg [193] method.

4.3.3 Bicluster classification

All interaction types from the HHPID were organised into a hierarchy (see supplementary file S4.1). This hierarchy included new parent terms. For example, a parent term 'physical' was created, the child terms of which all refer to more specific forms of physical interaction. In addition, every interaction was designated a *direction*, *polarity* and *control*. *Direction* refers to whether it is the HIV-1 protein acting upon the human protein or vice versa, e.g., 'Tat inhibits p53' has a forward direction, 'Tat is inhibited by p53' has a backward direction and 'Tat interacts with p53' has a neutral direction. *Polarity* refers to the biological action of the interaction, e.g., 'Vpr activates p53' has a positive polarity, 'Vpr inhibits p53' has a negative polarity and 'Vpr interacts with p53' has a neutral polarity. *Control* refers to regulation within the interaction, additional to the polarity, e.g., 'Tat decreases phosphorylation of retinoblastoma 1' has a positive polarity but due to the verb 'decrease' has a negative control, while 'Tat increases phosphorylation of retinoblastoma 1' has a positive control. For those interaction types with no additional control, we set control as null.

This information was used to classify biclusters according to the hierarchical relationship between their interactions. We defined three types of relationship between interactions: two positive relationships, parental and sibling, and one non-relationship, independence. Positive relationships refer to the same biological event within a given interaction, described using a different and perhaps more, or less specific term. Independence, denotes that two interactions describe distinct events, both providing additional information.

For any two interactions to be part of a positive relationship, they must link the same two protein interactants, their directions must not be opposing, i.e., forward and backward, their polarities must not be opposing and the control must be the same. Parental interaction relationships are formed when one interaction is the descendant of another, e.g., 'Tat binds p53' is a descendant of the interaction 'Tat interacts with p53'. Sibling interaction relationships are formed when both interactions have the same direct parent

term, e.g., 'Tat activates Cdk2' is a sibling of 'Tat enhances Cdk2', as the parent term for both interaction types is 'protein regulation'. Interaction pairs that do not conform to parental or sibling relationships have an independent relationship. These relationship classifications give rise to five classes of bicluster. (i) *Independent*, where all interactions have independent relationships. (ii) *Parental*, where all of the interactions are descendants of one another. (iii) *Sibling*, where two or more interactions are siblings of one another in the ontology, e.g., 'Tat activates Cdk2' is a sibling of 'Tat enhances Cdk2' (iv) *Family*, where all of the interactions form a 'family' of parental and sibling relationships. (v) *Mixed*, where independent interactions and sibling or parental interactions form a bicluster.

4.3.4 Bicluster biological validation

We established a group of 692 proteins from the HIV-1 interacting set that could appear in the bicluster results. These proteins are limited to those that have more than one distinct HIV-1 interaction. This set of proteins are important to our statistical analyses and will be referred to as potential bicluster proteins (PBPs).

A human PPI network was created using protein interactions derived from multiple sources: BioGRID [83], BIND [188] and HPRD [215]. All interactions were cross-referenced using the 'gene_info' file provided by the Entrez Gene database to maintain consistent accession labeling. These data sets were obtained in July 2009. The human PPI network contained only one node per human gene, a maximum of one edge between two nodes and a total of 9000 nodes and 30478 edges.

The number of shared edges, average shortest path length and largest connected component (LCC) for the set of human protein nodes defined by each significant bicluster were calculated from this network and statistical significance was calculated by Monte Carlo simulation. In a single iteration of this simulation, a group of nodes numbering the same as the bicluster in question were selected at random using rejection sampling in order to maintain the group degree distribution. Following this, shared edge count, average shortest path length and LCC were recorded. 10000 iterations of this simulation were carried out per bicluster. The results of the simulation were used to estimate the probability that a more tightly clustered set of nodes, of given size and degree distribution would be found by random chance, p-values were corrected for multiple tests using the Benjamini and Hochberg [193] method.

To analyze similarity between proteins within biclusters, we performed local protein sequence alignments between all PBPs using the Smith-Waterman algorithm [216] with a gap open cost of 10, a gap extension cost of 0.1 and the BLOSUM62 substitution matrix. To analyze similarity in annotation between proteins within biclusters, we carried out a semantic similarity measurement [217] between all PBPs using GO annotation for all three ontologies – molecular function, biological process and cellular component. The GO data was downloaded from the Gene Ontology on the 9th December 2008. We defined

the distance between two genes using the method given in [218], using the semantic distance measurement defined in [217]. For both of these measures we compared the value distribution for protein pairs that appear in the same significant bicluster to the equivalent distribution for proteins that do not appear in the same significant bicluster using a Mann-Whitney U test. We also calculated p-values for each significant bicluster, for both of these measures, using Monte Carlo simulations. For a given bicluster of size n and a mean average alignment score, or semantic similarity, s, 100000 and 10000 simulation iterations, for the pairwise alignment and semantic similarity simulations, respectively, were performed. In each iteration we selected a non-redundant random set of n proteins from PBPs and calculated the average alignment score or semantic similarity and counted whether this value was greater than, or less than s. By this method we were able to calculate the probability of finding a set of proteins, by random chance, with greater similarity than the proteins of a bicluster, both in terms of sequence and GO annotation. We corrected the p-values for multiple tests using the Benjamini and Hochberg [193] method. In addition, we identified groups of similar human protein sequences within significant biclusters using single-linkage clustering; linking pairs of proteins that have $> 40\%$ sequence identity, determined by sequence alignments and selecting connected components.

4.3.5 Defining subsystems

The distance between any two biclusters, a and b, was calculated using the formula:

$$d(a,b) = 1 - \frac{2\times|A\cap B|}{|A|+|B|}$$

Where A and B are the set of interactions in biclusters a and b, $|A|$ and $|B|$ are the number of interactions in sets A and B and $|A \cap B|$ is the size of the intersection between sets A and B. Therefore, for two identical biclusters $d(a,b) = 0$ and for two biclusters that have no common interactions $d(a,b) = 1$. Distances between all biclusters were calculated, cubed to obtain a greater range of values and the resulting distances were used to define relationships between biclusters using neighbor joining [219]. Meaningful groups were determined, by examining bicluster proteins and interactions. These groups were characterised and named using one of the following two methods. (i) Selecting one or more over-represented GO term ($p < 0.001$), calculated using Fisher's exact tests corrected for multiple tests using the Benjamini and Hochberg [193] method. (ii) In the case where the proteins of a subgroup are all homologs or isoforms of the same product and no specific GO term pertaining to that protein product exists, a regular expression encapsulating the protein name was used to characterise the group and that group was named after the protein. The method used to define each group is specified in supplementary file S4.4.

4.3.6 Comparison with siRNA screen data

Proteins from three siRNA studies [144, 145, 146], were cross referenced against the identified host subsystems. These studies include 281 [144], 295 [145] and 290 [146] genes. The genes not expressed in T cells or macrophages, designated group 'H' in one study (reference [146]) were not included. The number of successful hits against each subsystem and the direct intersection was counted for each of the three studies and p-values for these counts were calculated using chi-square tests, using all genes annotated in the gene ontology as the population, corrected for multiple tests using the Benjamini and Hochberg [193] method. HIV-1-host PPI networks were constructed and visualised using Cytoscape [164].

4.4 Results and discussion

4.4.1 Patterns of HIV-1-host interaction

We retrieved 1434 human proteins and 3939 distinct HIV-1-human PPIs from the HHPID. In order to precisely reflect findings from HHPID source papers and to maximise our capability to discern patterns within the data, all 19 HIV-1 proteins were used in our analysis. Not surprisingly, types of HIV-human PPI are not uniformly distributed among HIV-1 proteins, due to the different molecular functions of these proteins. We found that 18 from 19 HIV-1 proteins (all except Pol) take part in one or more interaction type with a frequency greater than expected by random chance ($p < 0.001$). These over-represented interactions include 47 of the 68 interaction types given in the HHPID and 60 distinct interaction-type/HIV-1 protein combinations. To give some examples: (i) The HIV-1 protein retropepsin is a protease required in the HIV-1 reproductive cycle to cleave viral polyproteins [220]. In addition, retropepsin cleaves proteins of the host cell [221, 222, 223, 224], hence, retropepsin is responsible for all but one of 61 distinct 'cleaves' interactions. (ii) The HIV-1 accessory protein Nef can impact expression levels of multiple genes during the viral reproductive cycle including proteases, cell-surface proteins, kinases, cyclins and transcription factors [225, 226]. Hence, Nef is responsible for a greater proportion of both upregulatory and downregulatory interactions than would be expected by random chance ($p = 2.5e - 5$). (iii) HIV-1 Tat is a transcriptional regulator that does not function alone [227], rather Tat works by recruiting other regulators [228, 229] and hence takes part in a greater proportion of interactions with type 'recruits' ($p = 1.1e-7$) and 'binds' ($p = 1.1e-4$). This over-representation analysis indicates that simple patterns of interaction (linking certain HIV-1 proteins to certain interaction types) are present in the HIV-1-host interaction network.

To computationally identify more complex patterns of virus-host interaction, we investigated human proteins that take part in more than one distinct PPI with HIV-1 proteins. An outline of our method for analysis of HIV-1 interaction is given in figure 4.1. As a first

step towards identifying key host functions known to be involved in HIV infection, we use biclustering to define groups of human proteins that share a common set (or 'profile') of HIV-1 interactions, in terms of HIV-1 protein interactant and interaction type (figure 4.2). The binary interaction matrix contained 1434 rows, 1292 columns and 3939 positive values, corresponding to human proteins, all types of HIV-1 interaction and all HIV-1-human PPIs, respectively. Biclustering of this matrix yielded 1306 biclusters that include a minimum of two human proteins, each with a minimum of two distinct HIV-1 interactions. We identified 279 from 1306 biclusters that were statistically significant ($p < 0.001$) by Monte Carlo simulation. A table with details of all significant biclusters, including their constituent human proteins, HIV-1 proteins, interaction types and links to the HHPID are given in supplementary file S4.1. fsus

These biclusters define significant profiles of HIV-1 interaction and a corresponding set of human proteins, or termed differently, significant sets of human proteins that undergo similar perturbations during HIV-1 infection. Included in the significant biclusters were 246 human proteins, 18 proteins of HIV-1 (all except p6) and 1665 distinct HIV-1-human PPIs. According to the classes of bicluster, defined according to relationships between interactions, we found 122 independent, 137 mixed, 11 parental, 9 family and no sibling significant biclusters. Both independent and mixed biclusters, according to our interactions hierarchy (see supplementary file S4.1), include a minimum of two unrelated types of PPI between every HIV-human protein pair. This indicates that our study of multiple interactions is informative and potentially valuable, as in $> 90\%$ of cases, bicluster interaction profiles include two or more types of interaction that provide distinct, additional information regarding the perturbation of the human proteins.

We expected significant biclusters to be enriched for well-studied, high-confidence interactions, since they are likely to correspond to identifiable units of biological function and well established modules that have been investigated more thoroughly than smaller, insignificant biclusters or singleton interactions. This hypothesis was tested by counting publications that support the interactions, as given in the HHPID. Whilst we do not regard publication count to be an ideal measure, it is a reasonable and accessible estimate for confidence in a given PPI. We found that interactions within significant biclusters had a mean of 2.94 supporting publications, while other interactions with human proteins that could potentially be in biclusters (these take part in at least two distinct interactions with HIV-1 and are referred to as 'potential bicluster proteins' or PBPs) have a mean of 2.46 and interactions with all non-biclustered interactions had a mean of 2.29. Mann-Whitney U tests performed on the publication count distributions of biclustered interactions versus PBP interactions and biclustered interactions versus all non-biclustered interactions, demonstrated that the distributions were significantly different ($p < 0.001$, in both cases). While we do not suggest that interactions outside of these biclusters are false positives and that all interactions within these biclusters are of elevated importance, this finding does indicate that the overall patterns of interaction defined by significant biclusters that

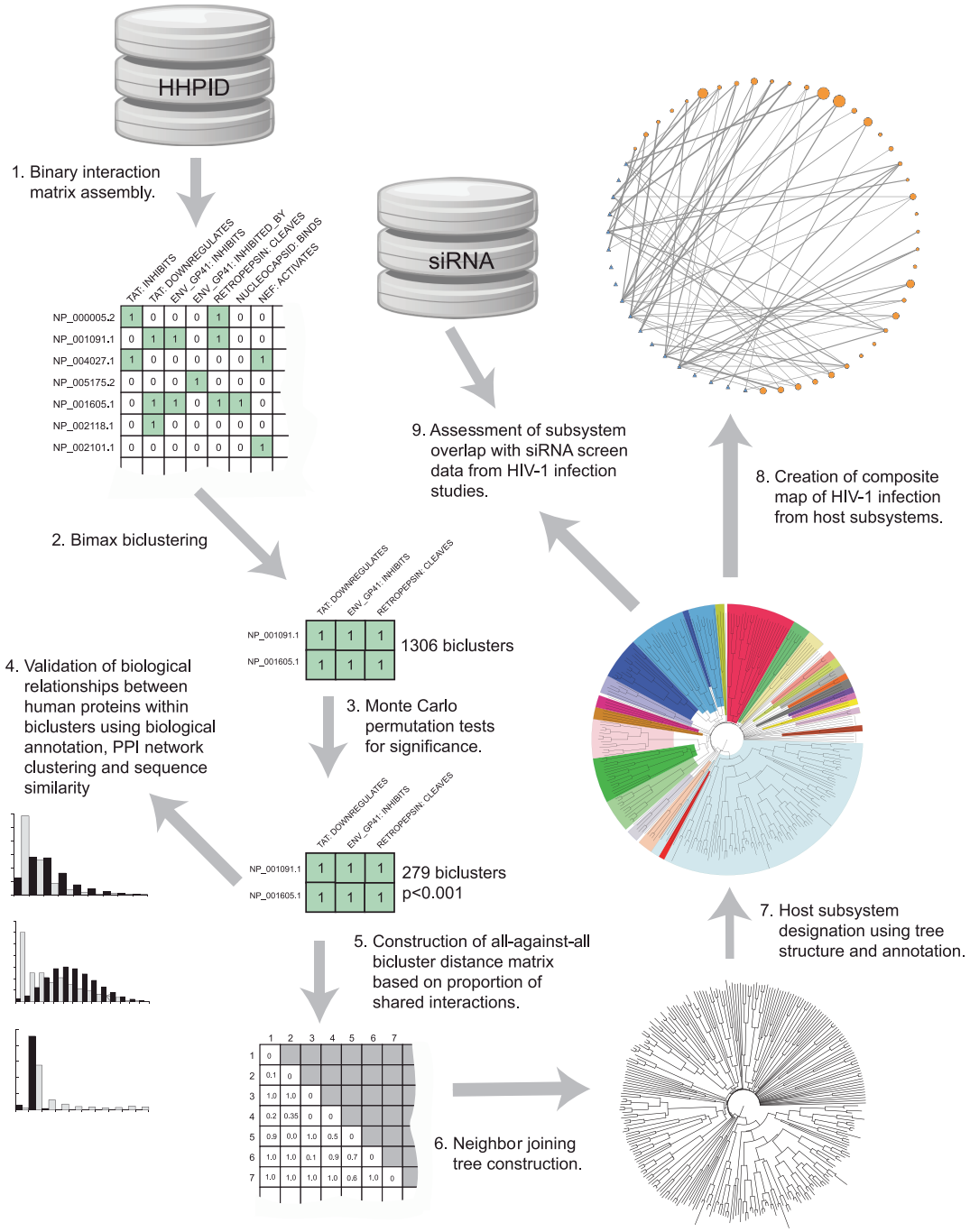

Figure 4.1: Summary of methodology. This diagram provides an outline of our method, stpdf are numbered according to the order in which they are discussed in the main text.

we discuss in this work, are likely to be biologically valid.

4.4.2 HIV-1 Interaction profiles define biologically cohesive sets of human proteins

To validate the biological significance of host protein sets and their associated interaction profiles (as defined by biclusters), we determined whether human proteins from

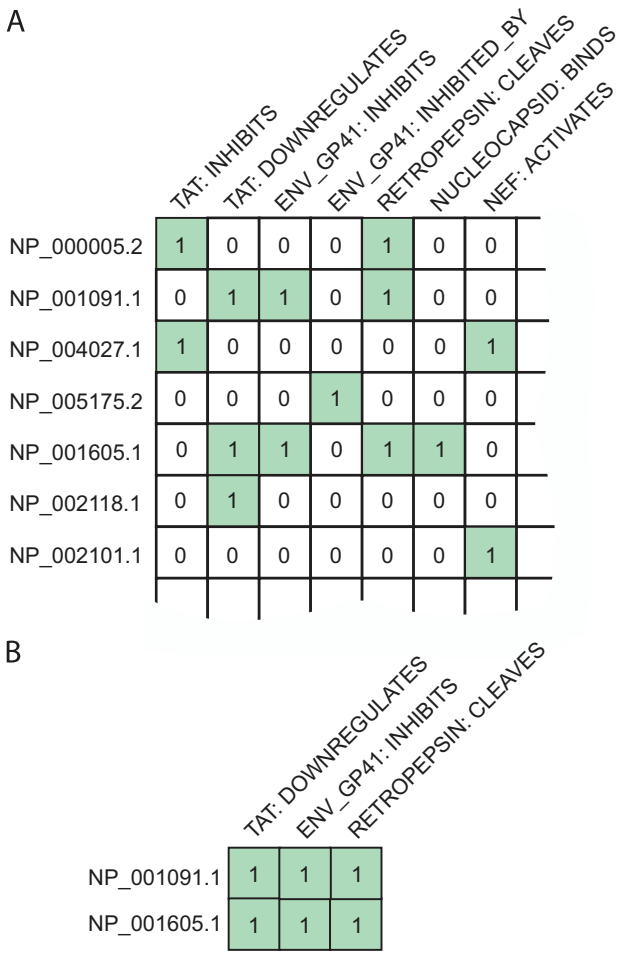

Figure 4.2: An example portion of the interactions matrix used in biclustering. (A) Shows an example portion of the interactions matrix. Here, a '1' represents the presence of a given interaction and a '0', the absence of that interaction, between a human protein interactant, shown left, and an HIV protein, the interaction having a given outcome, shown above. The entire matrix was biclustered to identify sets of host proteins that undergo the same set of HIV-1 interactions. (B) Shows an example bicluster that would be found in the portion of matrix given in (A).

within significant biclusters were more biologically similar to one another than expected by chance, assessed according to three measures: PPI network clustering to infer a greater then expected frequency of PPIs; semantic similarity in terms of shared Gene Ontology (GO) annotation [84] to infer shared biological roles; and sequence similarity to infer homologous relationships, as functional modules, such as protein complexes, are known to have a tendency to include paralogs [230]. These similarities were determined by comparing the host protein groupings to randomly selected sets sampled from 692 PBPs. Results for these measures are discussed below, followed by a summary of the three measures. In addition, detailed results, per significant bicluster, are given in supplementary file S4.3.

PPI network clustering

Integrating human proteins from significant biclusters into a human PPI network, we identified 38 biclusters where the proteins share a greater number of interactions, 24 where the proteins form a larger largest connected component (LCC) and 38 where the proteins have a smaller average shortest path length than would be expected by random chance ($p < 0.05$). A total of 66 biclusters appear in the union of these three measures and figure 4.3A gives details of their intersection.

These results show that HIV-1 has a tendency to interact in similar patterns with host proteins that share interactions with one another, indicating the presence of multiple HIV-1 interactions with host protein complexes or other closely associated host network modules. There are several prominent examples of complexed proteins that constitute all of the host proteins defined by significant interaction patterns including; class II major histocompatibility complex (MHC), general transcription factor IIH (TFIIH), casein kinase II, adaptor-related protein complex 1, protein phosphatase 2A, N-methyl-D-aspartate receptor, microtubule subunits and RNA polymerase II (RNAP II). In some cases HIV-1 interaction patterns with these complexes become significant due to the number of subunits that undergo a set of interactions. For example, one significant combination of interactions acts upon nine subunits of the RNAP II complex, hence, these proteins have more shared edges than would be expected at random. In this case, the interactions are general, pertaining to the complex rather than subunit specific, e.g., upregulation of RNAP II due to HIV-1 gp120 [231]. However, we also identify peptides from complexes that undergo subunit-specific interactions with the proteins of HIV-1. For example, one such bicluster involves HIV-1 Tat binding and regulation of specific polypeptides of TFIIH [27, 26, 174]. Yet, there are five other transcription-related host proteins within this interaction combination. In this case Tat interactions affect a functional module in the human PPI network (involving 18 interactions among the nine proteins, forming a single connected component) that corresponds to proteins of transcriptional regulation.

Semantic similarity

Biclustered proteins are more similar in terms of their GO annotation than would be expected by random chance in all ontologies; molecular function, cellular component and biological function ($p << 0.001$). Semantic distance distributions for human protein pairs from within biclusters and all other PBP pairings, for each ontology are shown in figure 4.4, graphs A to C. We identified 75 significant biclusters that include human proteins that are significantly similar in their GO annotation, for at least one ontology, from a possible 204 significant biclusters that include two or more genes with GO annotation ($p < 0.05$). Details of the intersection between results for each ontology are given in figure 4.3B.

These results show that HIV-1 interacts in a similar pattern with proteins that have similar GO annotation. We are able to observe these similarities in all GO ontologies. For example, protein kinase C (PKC) isoforms that comprise all human proteins of one biclus-

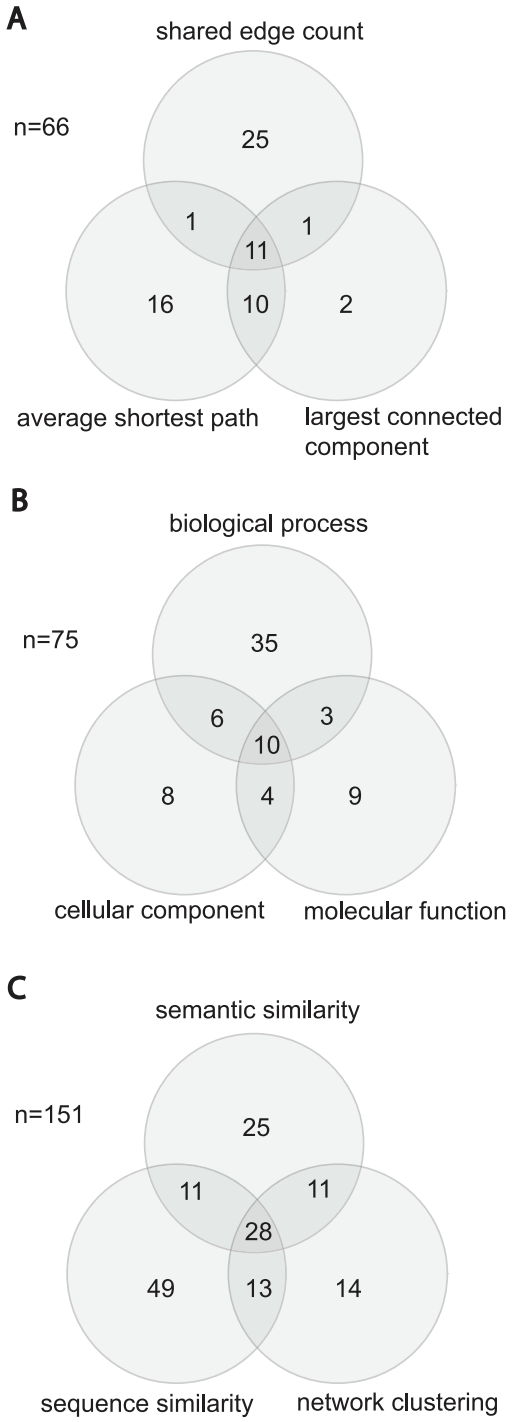

Figure 4.3: Venn diagrams showing biological cohesiveness among proteins within significant biclusters, using three measures. Counts refer to the number of biclusters that include human proteins that are significantly biologically related ($p < 0.05$) from a possible 279. (A) displays three network clustering measures: Shared edge count; average shortest path; and largest connected component. (B) displays semantic similarity in terms of the three GO ontologies. (C) displays the overlap of all three measures of biological cohesiveness; semantic similarity; network clustering and sequence similarity.

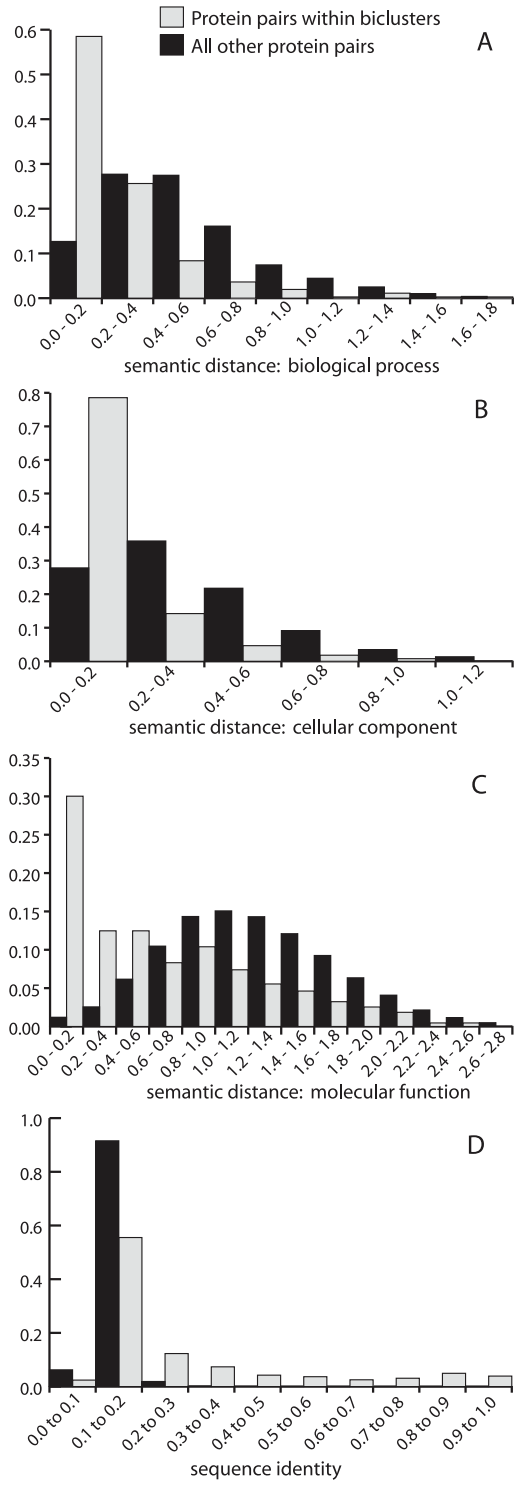

Figure 4.4: Comparison of protein pairs within significant biclusters to other protein pairs. Panels A, B and C show the semantic distance distributions for the three GO ontologies, biological process, cellular component and molecular function, respectively, for (i) human protein pairs from significant biclusters, shown in grey (ii) all other human protein pairs from PBPs, shown in black. Panel D shows the pairwise sequence similarity distributions for (i) and (ii). These charts show that human proteins from within significant biclusters are more similar in their GO annotation and sequence than other protein pairs ($p < 0.001$ in a Mann-Whitney U test, in all cases).

ter are annotated with the molecular function 'protein kinase C activity'. Some cellular component GO terms refer directly to protein complexes. Certain biclusters involving complexes are therefore linked via common annotation, such as one that corresponds to RNAP II, annotated with GO term 'RNA polymerase complex'. Interestingly, we are also able to observe HIV-1 interaction patterns that act upon specific biological processes including the immune response, protein kinase cascades, lipid modification, transcription, nuclear import and microtubule-based movement. The combinations of interaction that affect these processes can highlight the molecular methods through which HIV-1 infection perturbs cellular processes.

Sequence similarity

Human protein pairs within significant biclusters are more similar in their protein sequence than would be expected by random chance ($p << 0.001$). Distributions for sequence identity between human protein pairs from within biclusters compared to random pairings are shown in figure 4.4D. We identified 101 significant biclusters where the human proteins were more similar in their sequences than would be expected by random chance ($p < 0.05$). No biclusters were significantly less similar in their human protein sequences than would be expected.

We identify 58 biclusters for which a group of homologous proteins comprises more than half of the members of that cluster. We defined these homologous relationships by performing single linkage clustering on proteins, where proteins are linked if they share $> 40\%$ sequence identity. This cutoff was chosen as previous work has demonstrated that 40% sequence identity can accurately infer homology without the inclusion of an unacceptable proportion of false positives [218]. We found that significant biclusters with greater than expected sequence similarity among their host proteins ($p < 0.05$) were also more likely to have at least one direct physical HIV-1 PPI ($p = 7.81 \times 10^{-5}$) and the mean average proportion of direct HIV-1 PPIs among this group of biclusters was 25.5%, as opposed to 11.8% for all other significant biclusters.

These results show that paralogous groups of host proteins have a tendency to be subject to the same combinations of regulatory and physical HIV-1 interaction. Regulatory effects of HIV-1 interaction may be maintained across these groups, perhaps through stimulation of specific pathways. For example, isoforms of PKC, a kinase found to act in many, wide ranging signaling cascades [232], are the only host proteins among three particular significant biclusters. HIV-1 gp120 has been shown to upregulate multiple isozymes of PKC, possibly through classical signal transduction pathways [233], induced by binding to cell-surface receptors such as CCR5 [234]. However, the prominence of direct physical interactions among these homologous sets of proteins implies that there are conserved binding domains on members of closely related homologous groups, to which a HIV-1 protein can bind. For example, HIV-1 Vpr is designated in the HHPID to bind both importin-α 1 and 2 isoforms, these proteins are $> 40\%$ similar in a pairwise alignment,

therefore, it seems likely that Vpr would bind a particular conserved domain of these proteins. However, various members of protein families can exert distinct phenotypic responses. In the case of PKC isoforms, cellular localisation and activation input can be controlled by the specific domain structure [232]. Different members of protein families may also exert distinct phenotypic responses due to their cellular background, caused by differential expression, but also by activating downstream targets to different quantitative levels, as shown for receptor tyrosine kinases [235, 236]. Therefore, to precisely determine HIV-1 perturbation, it remains important to distinguish what protein isoforms and family members are dysregulated and in what cell type this activity occurs.

Summary for measures of biological cohesiveness

A summary of results from the three measures of biological relationship between proteins, in terms of the number of significant biclusters, is given in figure 4.3C. We find 151 from 279 biclusters are significant by one or more measure. Therefore, these measures are not mutually exclusive. In fact, in some cases, overlap may be due to a single biological phenomenon, e.g., homologous proteins that form a single complex are likely to be involved in the same biological process, in the same cellular compartment, possibly with the same molecular functions. For example, transcriptional regulators CREB binding protein (CBP), E1A binding protein (p300) and cyclin T1 are all found in one such bicluster whose interaction profile includes binding of these proteins to HIV-1 Tat and Vpr. CBP and p300 are $> 60\%$ identical in local pairwise alignment, however, rather than binding Tat individually, they form a dimer (known as PCAF) [237]. Cyclin T1 shares only a low level of sequence similarity ($< 30\%$ identity in local pairwise alignment) to the other two proteins. Therefore among these three host proteins there is a known PPI, a homologous relationship and all are transcriptional regulators involved with Tat mediated transactivation of the HIV-1 LTR [238] and hence have some common GO annotation. Furthermore, gene annotations including GO and PPIs may be attributed based on homology to genes with experimentally validated actions, for example, GO evidence code 'ISA', stands for 'inferred from sequence alignment' and is one of six codes describing computational assignment of annotation. Hence, the measures used here are linked. Some annotation is electronically inferred without any manual curation and as a result is error-prone [239], moreover, false positive annotations can be propagated electronically [240, 241]. However, we chose not to select manual annotation alone as the potential reduction in false-positives is offset by an increase in false-negatives. For example, more than half GO annotations of human genes have the evidence code 'IEA' meaning 'inferred from electronic annotation' (see http://www.geneontology.org/GO.current.annotations.shtml).

We do not identify significant biological relationships among 128 biclusters. These biclusters include significantly fewer human proteins on average ($\bar{x} = 2.32$) than the 151 biologically cohesive biclusters ($\bar{x} = 4.14$) ($p = 2.2 \times 10^{-16}$, Mann-Whitney U test). Therefore, power to detect statistically significant biological relationships (despite their

possible existence) among human proteins of these biclusters, is diminished, especially where annotation is lacking. For example two subunits of the casein kinase II complex (alpha 1 and beta) are found in one such bicluster. At the time of performing this work neither of these subunits were GO annotated, they are not more than 30% similar by local pairwise alignment and though they interact, this is insufficient to be called statistically significant. However, in some cases, no biological relationship can be discerned, even on inspection. Yet, of these 128 biclusters, 125 include fewer than three human proteins and none include more than four. This indicates that our combination of methods for detecting biologically cohesive human protein sets via biclustering and detecting biological relationships among these biclusters performs well in terms of quality, where the number of human proteins is four or greater.

4.4.3 Host functions among HIV-1-host interaction combinations

Owing to the specific biclustering method that we used for defining significant profiles of HIV-1-host interaction, multiple biclusters arise from slight differences between protein sets that are essentially similar in their interaction profile. This is partly due to differently annotated interactions, interactions that are not maintained across a group of otherwise similarly interacting proteins, or even to missing interactions that have not been experimentally proven or are missing from the HHPID, i.e., false negatives. For example, in the case of two biclusters that include homologs of Akt (also known as protein kinase B), one pertains to homologs 1 and 2, the other to homologs 1, 2 and 3. These two biclusters occur because homologs 1,2 and 3 have been shown to share similar interactions with HIV-1 gp120 and Vpr, however, while homologs 1,2 and 3 are activated by Tat, only homologs 1 and 2 are shown to be upregulated by Tat in the HHPID [27, 26, 174]. Therefore, by combining biclusters according to shared information, we can form an overview of HIV-1 interactions with a given set of host proteins.

Higher-level relationships between biclusters were identified using a distance measure based upon overlap between biclusters. Using the resulting pairwise distances a tree was constructed using the neighbor joining method [219] (see figure 4.5). This tree has been partitioned into sections, representing 37 biological subsystems within the host cell that are named according to over-represented GO terms, or after a specific protein (see materials and methods for more detail). In the tree representation we can observe subsystems that undergo a complex set of interactions during HIV-1 infection, these have a large number of terminal branches, representing many distinct but related HIV-host interaction combinations, where a single, clear pattern of interaction can not be simply defined, or does not exists, e.g., the *cytokine activity* subsystem. Conversely, the *v-akt* subsystem is relatively well defined including just two closely related HIV-host interaction combinations.

The identified subsystems and their associated patterns of interaction take place at a variety of levels within the host cell, including interactions at the cell surface and with

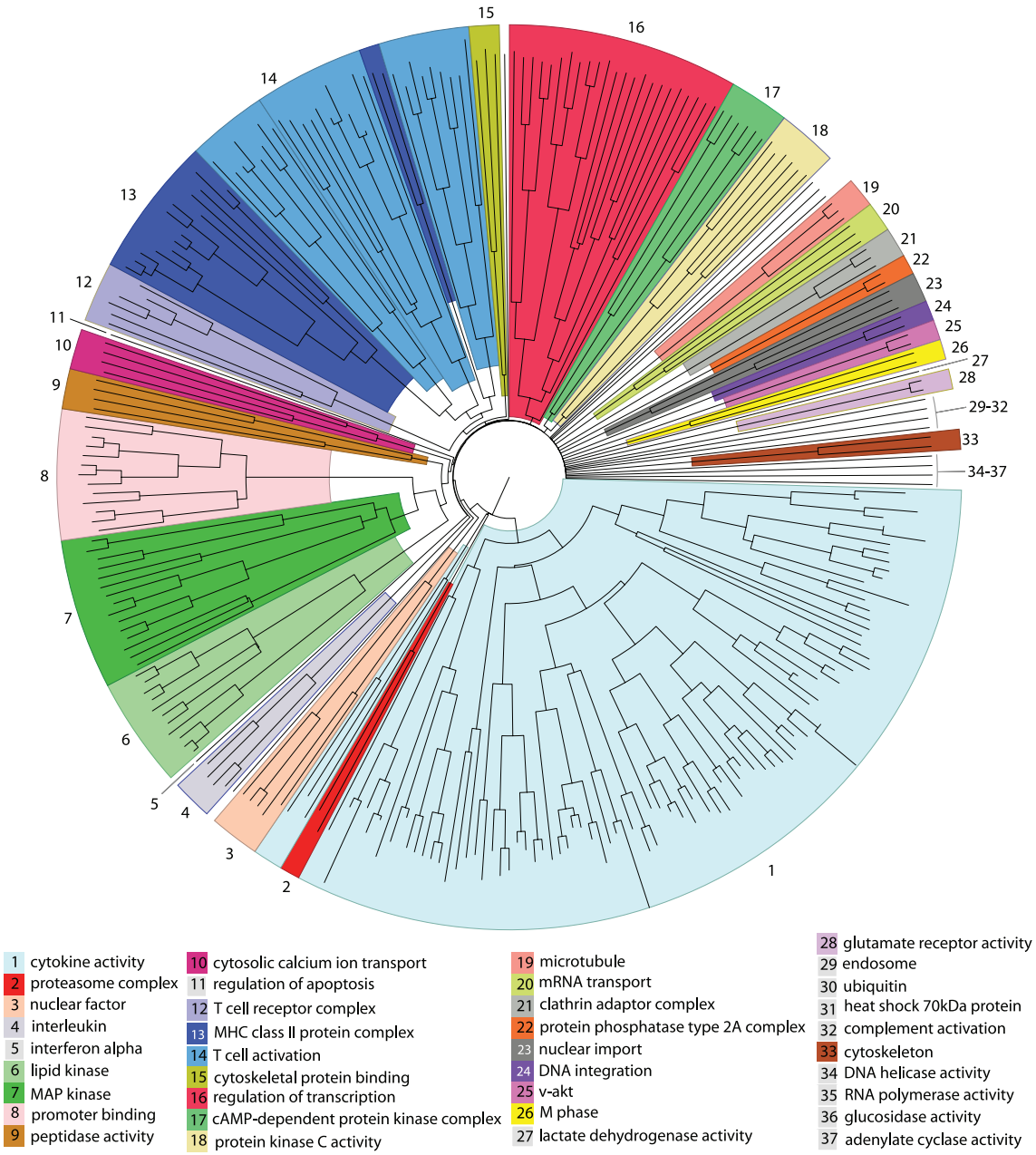

Figure 4.5: Tree showing the relationship between significant biclusters and higher-level host subsystem groupings. Individual biclusters are represented by terminal branches. Relationships are derived using a distance measure based on the proportion of shared interactions between significant biclusters and the tree was drawn using the neighbor joining method. The tree is divided into sections that show the higher-level host subsystems, largely derived using the tree structure. Subsystems of > 2 biclusters are colour coded (see key), biclusters not labelled are those that have been placed in a biologically related group not adjacent on the tree.

specific biological components such as the proteasome. Cellular processes and pathways, including intracellular signaling cascades, apoptosis pathways and stimulation of the immune response, better describe other subsystems. In addition, some subsystems can be directly mapped to specific steps in the viral reproductive cycle, including viral budding and transport of viral RNA across the nuclear membrane. Supplementary Table S3 gives

details of each subsystem, including the number of biclusters, host and virus proteins. Supplementary file S4.2 links individual biclusters and interactions to these subsystems.

Among these subsystems, there appears to be a central pathway of T cell signaling interactions that are perturbed by the proteins of HIV-1 at multiple levels in the cell. This pathway begins with inhibition of cell-surface receptor mediated signaling. For example, HIV-1 gp120 binding to CD4 prevents typical host-host cell-surface interactions, such as MHC-class II response to antigen binding [242], CD28-mediated co-signaling [243] and CD3-induced leukocyte-specific protein tyrosine kinase (Lck) and phospholipase C activation [244, 245]. In addition, HIV-1 Nef downregulates CD4, CXCR4, CCR5, CD28, CD71, CD80, CD86 and MHC class I molecules via endocytosis [246, 247, 248, 249, 250, 251].

We find continued perturbation of T cell signals at other cellular locations. For example in the *MAP kinase* subsystem we find Lck, a component of TCR signaling and an activator of other cell signal transduction proteins including the ERK family of MAP kinases [252, 253, 254, 255], is activated through gp120 binding to CD4 [256, 257, 258]. HIV-1 Nef also plays a role in the activation of the classical MAP kinase pathway via binding and activation of Lck [259, 260] and also Vav, causing downstream activation of JNK MAP kinases [261, 262]. Stimulation of these signaling cascades by proteins of HIV-1 influences a variety of cellular responses that include activation of transcription factors, for example [258, 263]. The *nuclear factor* subsystem includes nuclear factors of activated T cells (NFATs), transcriptional regulators that induce production of cytokines [264, 265]. We observe that NFATs are enhanced or activated at several levels within the host cell, by HIV-1 proteins Vpr, Tat, Nef and gp120, causing dysregulation of cytokine production [266, 267, 268, 269, 270, 30]. Altered cytokine signals will then be received by cell-surface receptors, thus completing a cycle of viral perturbation.

To summarise the interactions between cytokines and proteins of HIV-1, we produced networks of both upregulation and downregulation, taking interactions from the *cytokine activity* subsystem, including interactions that are supported by more than one publication, as given in the HHPID (see figure 4.6). These networks illustrate the complexity of cytokine dysregulation by HIV-1. From 53 distinct HIV-protein, host-protein pairings in these networks, 30 pairs involve only cytokine upregulation, 12 pairs involve only cytokine downregulation and 11 pairs involve both cytokine upregulation and downregulation, in response to the HIV-1 protein interactions. Cytokine dysregulation is likely to have major pathogenic effects on the host system. For example, an increase in plasma levels of multiple cytokines during acute HIV-1 infection, coined an 'early cytokine storm', is associated with peak viral loads and immunopathological consequences [271].

In these network visualisations there are distinguishable patterns of cytokine regulation by HIV-1, such as: The largely stimulatory effects of gp120, Tat and Nef; upregulation of TNF-alpha and Interleukins 1 and 6; the repressive action of Vpr and gp160; and downregulation of interleukin 2 and interferon-γ. However, the overall picture of

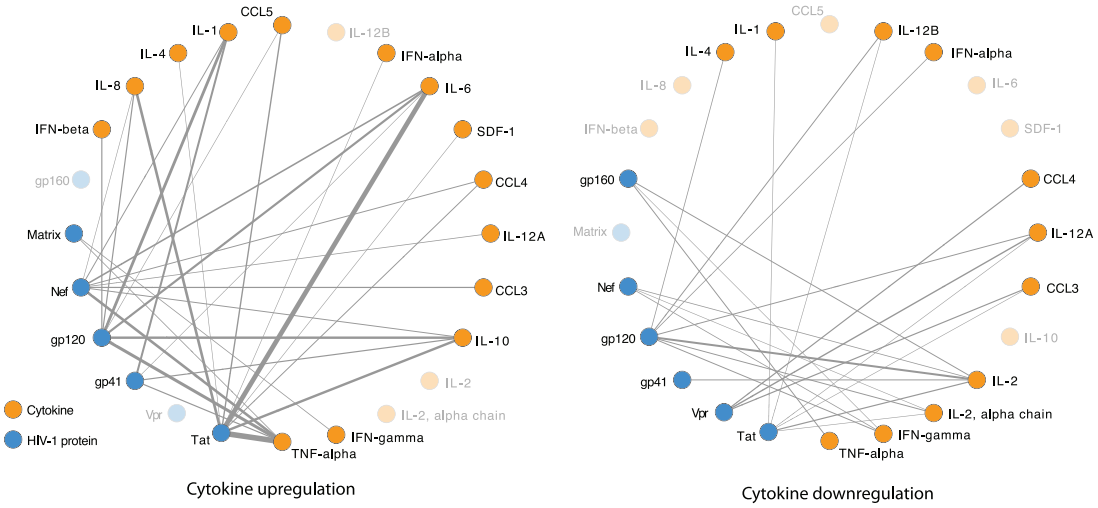

Figure 4.6: Cytokine regulation networks. These networks represent the pattern of cytokine regulation in the cytokine-activity host subsystem that were defined through identifying significant patterns of HIV-host interaction. Edges represent PPIs, edge width is proportional to the number of PPIs being represented. Here, for clarity, we only show PPIs that are reported more than once in the HHPID. These networks show that cytokine dysregulation due to HIV-1 infection is wide reaching and complex, affecting many host cytokines, both via upregulation (left) and downregulation (right).

cytokine regulation during HIV-1 infection remains unclear. Future cytokine-wide studies of HIV-1 infected cells, preferably representing multiple different stages of infection and possibly even a variety of HIV-1 strains, coupled with an accurate model of cytokine action on the host system could improve our understanding of HIV-1 pathogenesis and potential intervention targets, particularly if key HIV-host interactions are identified.

To present distilled views of the HHPID and provide an interpretable network of HIV-1-host interaction, two HIV-1-host PPI networks were constructed. Both networks include 37 nodes that represent the characterised subsystems. The first network, shown in figure 4.7, has 18 nodes that represent the proteins of HIV-1. The second network, shown in figure 4.8 has 49 nodes that represent interaction types. The edges in these networks represent HIV-1-host interactions that contribute to significant biclusters, the width of each edge is proportional to the number of distinct interactions that are represented. Due to the condensed host functions and filtering out of patterns of interaction that are not statistically significant we can observe recognisable patterns of interaction in these networks. For example: (i) The relationship between HIV-1 Tat and regulators of transcription that are stimulated, activated and recruited by HIV-1 in the process of viral transcription. (ii) The multiple sources of perturbation of T cell activation from HIV-1 Nef, Tat and the envelope proteins. (iii) The large number of regulatory interactions between proteins of HIV-1 and host cytokines.

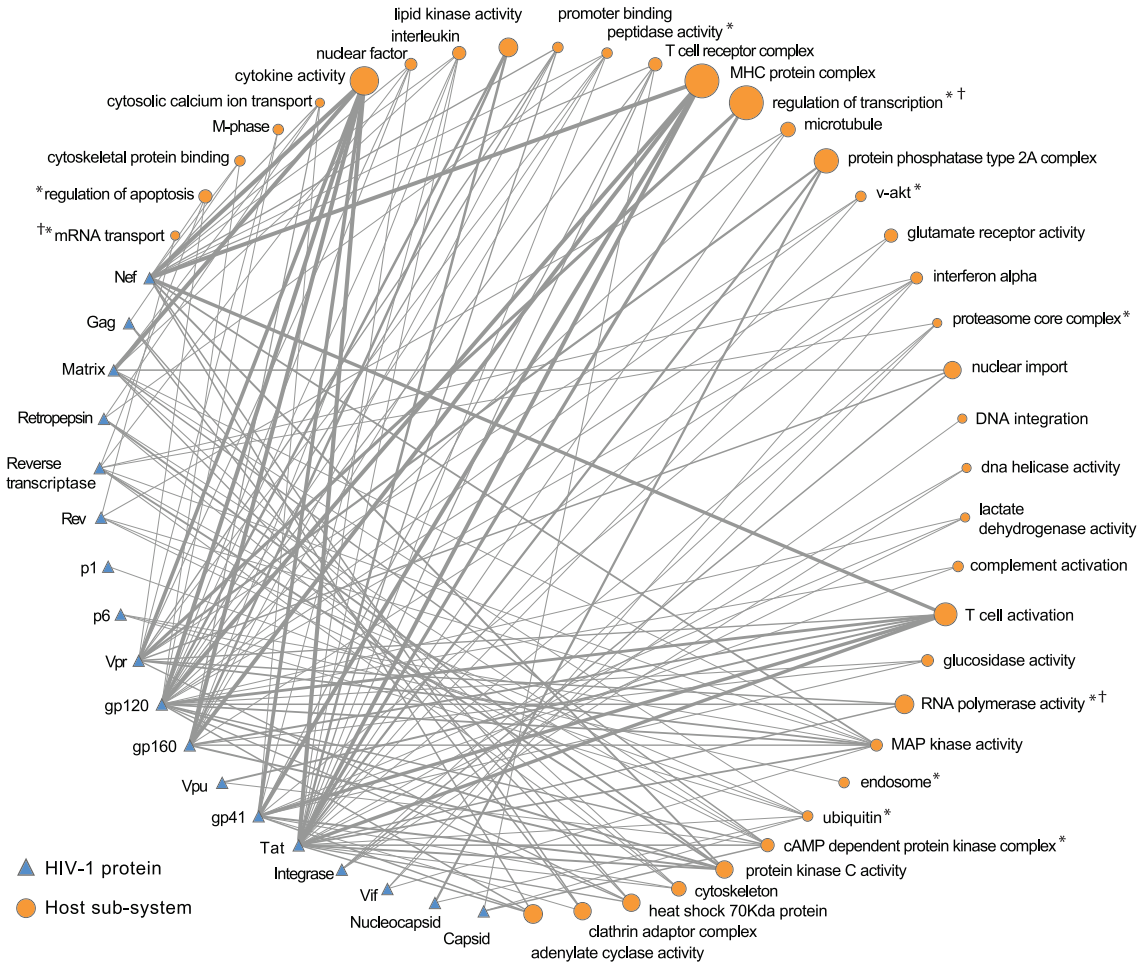

Figure 4.7: HIV-1-host interaction patterns, by HIV-1 protein. This network illustrates core patterns of HIV-host interaction. The human host is depicted as a series of cellular subsystems, represented by orange circular nodes, where the diameter of the node is proportional to the number of host proteins within that subsystem. HIV-1 is depicted by the viral proteome (blue triangles). Interactions between HIV-1 proteins and host subsystems are represented by edges, where the edge width is proportional to the number of interactions. For clarity, only those interactions that are shared by over half of the host proteins in a subsystem are shown. *Indicates a host subsystem whose subsystem annotation corresponds to a statistically significant group among HDFs ($p < 0.05$). †Indicates a statistically significant intersection between the subsystem proteins and HDF set ($p < 0.05$).

4.4.4 Support for host subsystem functions among global siRNA data sets

To assess support for the 37 host subsystems from HDFs identified by global siRNA screens [144, 145, 146], we defined subsystem annotations that consist either of defining over represented GO terms or a regular expression that encapsulates a common protein name. Subsystem annotations are given in supplementary file S4.4. We found that 10 from 37 subsystem annotations also define statistically over-represented groups among either all HDFs combined or a single HDF study ($p < 0.05$). We find that 21 from 37 subsystems

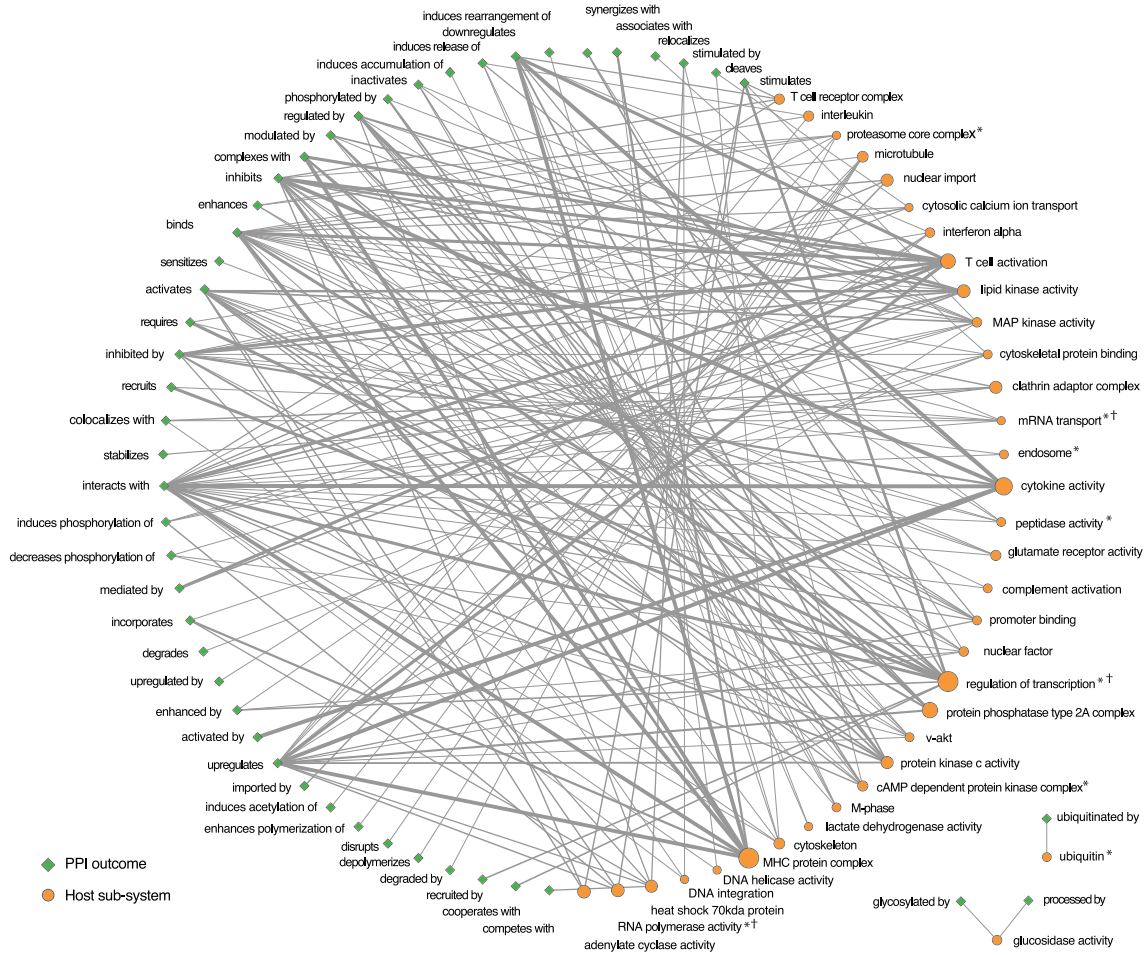

Figure 4.8: HIV-1-host interaction patterns, by interaction type. This network illustrates core patterns of HIV-host interaction. The human host is depicted as a series of cellular subsystems, represented by orange circular nodes, where the diameter of the node is proportional to the number of host proteins within that subsystem. The action that HIV-1 has on these subsystems is depicted by a series of interaction outcomes (blue diamonds). Interactions between HIV-1 and host subsystems are represented by edges where the edge width is proportional to the number of interactions. The directionality of the interaction is implicit in the description of the interaction outcome. For example, the edge linking the MHC protein complex node and the 'upregulates' node represents 'HIV-1 upregulates the MHC protein complex', whereas the edge linking the cytokine activity node and the 'activated by' node represents 'HIV-1 is activated by cytokine activity'. For clarity, only those interactions that are shared by over half of the host proteins in a subsystem are shown. *Indicates a host subsystem whose subsystem annotation corresponds to a statistically significant group among HDFs ($p < 0.05$). †Indicates a statistically significant intersection between the subsystem proteins and HDF set ($p < 0.05$).

include at least one protein that is also present among HDFs and only in three cases is the intersection statistically significant (see supplementary file S4.4 for more details).

Cellular subsystems supported by HDF sets

The 10 subsystems that are supported by HDFs are: *Proteasome core complex*; *regulation of apoptosis*; *mRNA transport*; *endosome*; *RNA polymerase activity*; *peptidase activity*; *regulation of transcription*; *ubiquitin*; *cAMP-dependent protein kinase complex*; and *v-akt*.

Subunits of the proteasome core complex are present among two of the three siRNA screens [144, 145]. A meta-analysis of these HDF sets that incorporates data from the HHPID, showed that the proteasome is an important cellular component for HIV-1 replication [156]. The role of the proteasome in HIV-1 replication remains unclear. However, the interactions that we highlight between HIV-1 Tat and the beta-8 and beta-10 subunits may be important for determining proteasome composition, towards formation of the immunoproteasome, a change that may cause increased presentation of subdominant epitopes [272, 273].

Apoptosis is widely accepted as a mechanism for T cell depletion in HIV-1 infected individuals [274]. By reviewing relevant literature, we find that several subsystems may have a role in controlling apoptosis including: *Regulation of apoptosis*; *glutamate receptor activity*; *v-akt*; *lactate dehydrogenase activity*; and *peptidase activity*. HDFs found by one siRNA screen [146] are enriched for regulators of apoptosis. However, in our results, we only identify one HDF, Cytochrome C that is GO annotated as a regulator of apoptosis in addition to Akt and components of the glutamate receptor. We speculate that prevention, rather than induction of apoptosis, is an essential part of HIV-1 infection, in order to maintain a viral reservoir in the host [275]. In this case, HDFs may not include the pro-apoptotic host proteins that we observe in these interaction patterns. In addition, proteins such as Akt and Cytochrome C have roles outside of apoptosis [276, 277], therefore, the necessity for such proteins in HIV-1 replication is not necessarily apoptosis related. However, we identify subsystems from the HHPID that can be linked to positive regulation apoptosis, such as *regulation of apoptosis* that includes the activation of pro-apoptotic caspases by multiple HIV-1 proteins. The intensity of research to elucidate key interactions responsible for T cell loss, via apoptosis, in HIV-1 infected individuals is demonstrated by the prominence of pro-apoptotic HIV-host interactions in our results. However, we suggest that a greater range of interactions between proteins of HIV-1 and host regulators of apoptosis need to be investigated, particularly involving those host factors that are present among HDFs but not identified in our results.

Interactions in the *mRNA transport* subsystem all involve HIV-1 Rev. One of the roles of Rev is to facilitate export of HIV-1 RNA from the nucleus to the cytoplasm. A nuclear export signal present in the Rev protein binds to exportin 1, while an argenine-rich domain (ARD) in Rev binds to a Rev-response-element (RRE) present in viral RNA. To undergo nuclear export, an exportin-Rev-RNA complex docks at a nuclear pore complex (NPC); this interaction is mediated by nucleoporins [36]. In the *mRNA transport* subsystem we find interactions that are specific to this process including but not limited to: Binding

[37] and recruitment [278] of exportin 1 by Rev; and direct interactions between Rev and two nucleoporin proteins [279], including Rev mediated recruitment of these host factors to the nucleus [280]. We find that there are a statistically significant proportion of host factors involved in mRNA transport in two of the three global siRNA studies (references [144, 145]). Furthermore, all host genes that make up this subsystem (total 5) are found among HDFs ($p = 0.00014$). Down-modulation of these host factors, either in small-scale experiments or by global siRNA screen, apparently inhibits the interactions described in this subsystem, thereby preventing Rev-mediated RNA nuclear export and successful viral replication.

We observe two other subsystems that appear to have a role in transport of HIV-1 material into the nucleus, *nuclear import* and *heat shock protein 70kDa*. Briefly, the *nuclear-import* subsystem involves a variety of interactions with members of the karyopherin family and heat shock 70kDa chaperone protein (Hsp70). Karyopherins bind sequence motifs called nuclear localisation signals (NLS) of proteins, causing the protein to be directed into the nucleus [281]. We observe that HIV-1 Integrase, Matrix, Tat and Rev proteins are bound or imported into the nucleus by members of the karyopherin family. In the case of Integrase and Matrix, these interactions may relate to karyopherin mediated nuclear import of HIV-1 preintegration complexes (viral ds-RNA and associated proteins known as PICs) [282], a mechanism that may also involve HIV-1 Vpr. Several isoforms of the heat shock 70kDa chaperone protein (Hsp70) promote PIC import, possibly by stimulating interaction between PIC complexes and karyopherin [283]. These two nuclear import subsystems include support from siRNA screens. One of the three studies identified karyopherin-β [145] and two studies [144, 145] identified transportin 3 (TNPO3) – a less definitively characterised member of the importin-β/karyopherin-β superfamily – as HDFs. More recently, TNPO3 has been reconfirmed by yeast-two-hybrid pull-down as a binding partner of Integrase, as an early-stage HDF in the viral reproductive cycle by siRNA screen and a clear promoter of HIV-1 PIC import [284], though subsequent work has shown that HIV-1 requirement for TNPO3 maps to interaction with Capsid rather than the Integrase protein [285]. Therefore, current experimental data indicates that TNPO3, is essential for PIC import, whereas the role for karyopherin-β in this process remains unclear. Requirement for karyopherin-β observed in [145] could be indirect, perhaps for transport of another HDF.

Budding – the release of the viral particle from the host cell plasma membrane – is an essential step in the HIV-1 reproductive cycle. We identify two subsystems that have a role in budding: *Protein localisation* and *ubiquitin*. Both of these groups include interactions involving HIV-1 p6, a region of the Gag protein that contains a late domain (L-domain). L-domains recruit host-cellular factors required by HIV-1 for budding. Our results indicate that p6 (along with other viral proteins) is ubiquitinated at the L-domain by three forms of ubiquitin (B, C and D). The p6 L-domain also interacts with subunits of the ESCRT-I complex, possible via direct interaction with AIP-1/ALIX. These interactions, though not

fully understood, have been shown to be important for HIV-1 budding [286] and are found in our results. These host factors are not identified among HDFs. However, HDFs include both ubiquitin-conjugating enzymes and ubiquitin-protein ligases. Therefore, it appears that ubiquitination plays an important role in HIV-1 replication that can be linked to viral budding.

In our results we define four subsystems where the host proteins and interactions contribute to HIV-1 provirus chromosomal integration and HIV-1 RNA transcription, namely, *regulation of transcription*, *DNA helicase*, *RNA polymerase activity* and *DNA integration*. The largest of these groups is *regulation of transcription* that includes many direct binding and co-stimulatory interactions between HIV-1 proteins Tat and Vpr and host transcriptional regulators including: Cyclin-dependent kinase 9 and cyclin T1 that form the Positive Transcription Elongation Factor b complex, general transcription factors TFIIF and TFIIH; NFκB; TATA box binding protein; cyclin-dependent kinase 9; CREB binding protein; p300; and p300/CBP-associated factor [287, 25]. Both the size of the intersection between this subsystem and HDFs and the proportion of genes annotated by GO as regulators of transcription is statistically greater than expected by random chance ($p = 0.0034$ and $p = 0.0038$, respectively). Transcriptional regulators we identify that are also among HDFs include cyclin t1, NFκB, p300, TFIIF and TFIIH, as well as subunits of the RNA-polymerase II complex, as found in the *RNA polymerase activity* subsystem. Therefore, these host factors appear to form an essential functional module, with a clear pattern of interaction required for HIV-1 replication.

The *DNA integration* subsystem includes interactions between HIV-1 Integrase and three host proteins: LEDGF, a transcriptional activator; hSNF5, a subunit of the SWI/SNF ATP-dependent chromatin-remodeling complex; and embryonic ectoderm development (EED) protein. Integrase is involved in binding interactions with both LEDGF and hSNF5. LEDGF binds to Integrase and tethers it to host chromatin, an interaction identified as essential to HIV-1 infectivity [288, 289]. However, LEDGF is not found among HDFs, perhaps because this host factor is only required at a very low level, thus could elude identification by siRNA knockdown screening [289]. This highlights the possibility that more host proteins shown to be essential for HIV-1 replication in specific, small-scale experiments may not be found among HDFs.

By cross referencing host proteins involved in significant patterns of interaction from the HHPID we have found support among siRNA screen data for host subsystems that can be linked to viral transcription, viral budding, PIC integration, transcription of viral RNA, changes to proteasome composition, export of viral RNA from the nucleus and regulation of apoptosis. However, of these, all but regulation of apoptosis and changes to proteasome composition might be considered an essential molecular mechanism for HIV-1 replication. Moreover, from our results it is unclear whether pro-apoptotic interactions are essential.

Lack of support for T cell signaling and immune-related subsystems among HDFs

The remaining subsystems are not well supported by data from siRNA screens, in particular, those pertaining to cytokine dysregulation caused by HIV-1 infection. We do not find that HDFs are enriched for components of the TCR or for proteins involved in T cell activation. However, we do find that CD4 and CXCR4 have both been identified by our interaction patterns and by two siRNA screens, probably as these receptors were essential for virus entry in the two studies using HeLa cell lines [144, 146], whereas in the third study, CD4 and CXCR4 were not identified, presumably because an engineered mechanism for viral entry was employed [145].

There is little support for proteins involved in MAP kinase or PI3K-mediated intracellular signaling among global siRNA data sets. We find no HDFs that are GO annotated as having MAP kinase activity and just one HDF with lipid kinase activity (phosphatidylinositol-4-phosphate 5-kinase type-1-γ), though we find two HDFs with PKC activity (PKC-η and serine/threonine-protein kinase N2). These findings indicate that knock-downs of single proteins from these cascades are generally insufficient to significantly inhibit HIV-1 replication. However, we surmise these central cascades are able to maintain signal transduction through multiple routes – the KEGG representation of the MAP kinase cascade indicates this possibility [182, 290, 291]. Furthermore, HIV-1 interaction with these cascades is largely regulatory, rather than the result of direct interactions. Therefore, HIV-1 may not require any one specific protein from a central signal transduction cascade, such as a particular MAP kinase, for HIV-1 replication. Yet transduction of virally induced signals through the host cell is almost certainly an important mechanism in the proliferation of HIV-1.

We do not find that the subsystem annotations of any of *cytokine activity*, *interleukin*, *interferon-*γ and *nuclear factor* subsystems represent statistically significant sets among HDFs. However, among HDFs, we do find five genes that are designated by GO as having cytokine activity including IL-1, chemokine-like factor, two additional interleukins (IL-18 and IL-22) and Interferon-related developmental regulator 2. These results indicate some cytokines and chemokines are likely to enhance HIV-1 replication. However, in our results cytokines form a far larger and more prominent set of host proteins and interactions. We suggest that this disparity is because while *in vivo* cytokines play a key part in modulation of viral replication, by providing a pool of cells for infection [292], immune system activation via cytokine release may not be essential for viral replication within any given cell. Indeed, *in vitro*, HIV-1 regulation of cytokines is likely to be of diminished importance as there is no functional acquired or innate immune system for the virus to interact with – either for the purpose of evasion or hyper-stimulation. Furthermore, small scale *in vitro* studies that have been explicitly designed to test the significance of HIV-1 protein interactions with cytokines and *in vitro* siRNA screens that test for HIV-1 dependence on host factors on a global scale, are unlikely to reach the same conclusion, regarding the relevance of cytokines to HIV-1 infection. The diminished importance of cytokines

among HDFs is also indicated by the lack of support for the NFATs that promote cytokine transcription.

Innate cellular immune responses, such as APOBEC activity, the interferon system and TRIM22-induced interferon activation will be important *in vitro* [293, 294, 295, 296]. Though as these innate immune factors exert a negative effect on HIV-1 replication, they are unlikely to be highlighted among HDF sets.

4.4.5 Conclusion

By capturing the published knowledge of HIV-host interactions, the HHPID represents a hugely valuable resource for HIV-1 research. However, redundancy and heterogeneity of the PPI data, in terms of experimental methods, age of findings and quality of data, make the HHPID a difficult data set from which to draw conclusions about the overall system of HIV-1 infection, such as the identification of specific host functions and processes that are essential for HIV-1 replication. Using the strategy presented here, we identify significant patterns of HIV-host interaction – sets of host proteins that take part in similar, enriched combinations of interactions during the course of HIV-1 infection. We have confirmed that these host protein sets, linked by their HIV-1 interaction profiles, are biologically related, tending to include proteins with common biological processes, proteins that share a high number of interactions with one another, subunits of the same complex and paralogs. In addition, we find that by identifying significant interaction patterns, we select for higher-confidence, well-studied interactions, based on the number of supporting journal articles. Hence, the identified higher-level groups, based on shared interactions, represent significant cellular subsystems used by HIV-1. Notably, our method incorporates the biological action of each PPI. Therefore, unlike other studies that identify cellular subsystems important to HIV-1 [145, 156, 174], the subsystems presented here, respect specific activity-related patterns of viral perturbation.

By assessing these subsystems using scientific literature and support from three global siRNA screen HDF sets, we have been able to describe systems of interaction that are invoked by HIV-1 to hijack host functions in order to successfully replicate including virus entry, mechanisms for viral gene transcription, export of viral RNA from the nucleus, viral budding and control of the proteasome. In addition, we also highlight mechanisms through which HIV-1 infection perturbs host processes at multiple cellular levels through a cycle of interactions that is not necessarily essential for viral replication, yet appear detrimental and potentially lethal to the human host by damaging the host immune response through dysregulation of cell surface receptor mediated signaling, signal transduction pathways, host gene expression, cytokine release and cell death.

Our approach permits a detailed study of the overlap between significant patterns of HIV-host interaction in the HHPID and HDFs. The modest overlap may be attributed to the fundamental difference in the methods of construction between the source data sets. The siRNA screens do not explicitly identify host cell proteins that undergo direct phys-

ical interactions with the proteins of HIV-1 or whose expression is altered during HIV-1 infection, as with many of the host cell proteins given in the HHPID. Rather, these screens are designed to identify host-cellular proteins that are required by the virus for replication. Therefore, HDF sets will not necessarily capture host proteins that are misregulated during HIV-1 infection, i.e., may perturb normal cellular responses, or host proteins that are potentially detrimental to HIV-1 infection, such as APOBEC3G [293]. In addition, each study has its own intrinsic bias. Particularly, the HHPID will be subject to study bias [297], where aspects of perceived medical importance, such as T cell depletion, receive greater attention. Whereas methods employed in each siRNA screen will be better tailored to picking certain host proteins over others. For example, one siRNA screen was specifically designed to discover host factors involved in the early stages of HIV-1 replication [145], another used a viral strain that expresses a truncated Vpr protein and does not express Nef or Vpu [144]. In addition, the stage in the viral reproductive cycle is also likely to be an important factor in determining the activation of PPI modules in the host cell [298], therefore, not all studies may capture the same results. Hence, the lack of overlap between these small and global-scale data sets is not unexpected.

The direct intersection between any one HDF set and the HHPID probably represents a small set of high-confidence HIV-1 interacting host proteins important to HIV-1 replication, however, analysis of this intersection alone is unlikely to provide a thorough insight into host defense mechanisms, perturbations caused by HIV-1 infection, or proteins that are essential to virus replication. We suggest that future experimental work could expand the core knowledge presented here. In particular, we suggest that proteins and pathways that are indicated by siRNA screen to be essential for HIV-1 replication, though otherwise poorly understood, are studied in greater detail to continue to bridge the knowledge gap between high and low throughput data sources. A successful example of this approach is conformation of TNPO3 as an essential protein for HIV-1 PIC import [284, 285] after initial identification as an HDF [144, 145].

The HHPID data set has been used previously to validate HDF sets. Specifically HHPID interactions and host factors have been used in conjunction with HDFs to aid identification of well-connected subnetworks, corresponding to certain host cell functions prevalent among HDFs [145, 156]. Several of these subnetworks represent functions identified in our results including the proteasome and transcriptional regulation. However, we are not aware of any other work in which core host cell functions, represented in HHPID data, have been assessed in terms of their presence among HDFs.

In this study, we have used a computational approach to disentangle a complex set of interactions to provide an accessible map of core HIV-1-host interaction patterns for virologists. Our methodology can be generalised and take PPI data from any source. Hence, our work will contribute to defining core host subsystems for other pathogens, particularly as a reference against which results from increasingly prevalent high-throughput data sources might be compared. In addition, aiding prediction of currently undiscovered

host-virus PPIs using interaction profiles may be possible. This could be done by taking the interaction profile of a given human protein, i.e., a 'subject profile', and comparing it with interaction profiles from other human proteins, i.e., a set of 'query profiles' to look for common interactions that are missing for the subject profile that are common to many other similar profiles (the distance measure in our work would be a method to quantify this commonality). However, this would form just part of such a prediction process and other established biological phenomena that impact upon PPI activity, such as interaction interfaces and cellular localisation, would also have to be considered to make successful predictions. Notably, our results and the potential predictive power to which we refer, are reliant upon an accessible and structured description of biological action for each PPI, as supplied in the HHPID. We have, thus, demonstrated that the inclusion of concise annotation in large-scale data can enhance resolution and allow greater depth of computational analysis.

4.5 Supporting material

Supplementary file S4.1 – hierarchy of protein interaction types. A hierarchy that incorporates all of the interaction types found in the NCBI HIV-1, host protein interaction database (HHPID) with the addition of parent terms for these types. HHPID interaction types have a unique id, and polarity, direction and control attributes. These attributes are explained in detail in the methods section in the main text of this article. Interaction types found in the HHPID are present as instance elements, parent terms are designated as interactionType elements.

Supplementary file S4.2 – table of significant biclusters and their HIV-host interactions. A table of significant biclusters. Each row represents a single HIV-host interaction within a significant bicluster. The biclusters are divided into higher-level groups, known as sub-systems, based on shared interactions and labeled according to the biological role of the included host proteins. From right to left, the columns show: name of the sub-system; bicluster id; p-value for the bicluster; corrected p-value, calculated using the Benjamini and Hochberg FDR correction method; entrez gene id corresponding to the human protein interactant; name of the human protein interactant; entrez gene id corresponding to the HIV-1 protein interactant; a string identifier for the interaction type, consisting of a short HIV-1 protein name and a description of the interaction outcome, separated by an underscore; the relationship of that interaction to other interactions within the same bicluster; id for the corresponding interaction in the HHPID; the HHPID description of the interaction.

Supplementary file S4.3 – table of biological cohesiveness measures for significant biclusters. A table of biological cohesiveness measures. Each row represents a significant bicluster. Sequence similarity, semantic similarity and network clustering are measures pertaining to the proteins of a given bicluster. From right to left, the columns

show: bicluster id; p-value for sequence similarity; corrected p-value for sequence similarity, calculated using the Benjamini and Hochberg FDR correction method; number of proteins found in the largest protein cluster, within that bicluster, determined by single linkage clustering, using a linkage cut-off of 40% sequence similarity; as in the latter but using a cut-off of 80% sequence similarity; p-value for human PPI network shared edge count; corrected p-value for human PPI network shared edge count, calculated using the Benjamini and Hochberg FDR correction method; p-value for human PPI network largest connected component; corrected p-value for human PPI network largest connected component, calculated using the Benjamini and Hochberg FDR correction method; p-value for human PPI network average shortest path length; corrected p-value for human PPI network average shortest path length, calculated using the Benjamini and Hochberg FDR correction method; p-value for semantic similarity using the GO biological process ontology; corrected p-value for semantic similarity using the GO biological process ontology, calculated using the Benjamini and Hochberg FDR correction method; p-value for semantic similarity using the GO cellular component ontology; corrected p-value for semantic similarity using the GO cellular component ontology, calculated using the Benjamini and Hochberg FDR correction method; p-value for semantic similarity using the GO molecular function ontology; corrected p-value for semantic similarity using the GO molecular function ontology, calculated using the Benjamini and Hochberg FDR correction method.

Supplementary file S4.4 – table of host subsystem details. A table of host subsystem details. Each row represents a host subsystem. From right to left the columns show: the name of the subsystem; the number of biclusters included in the subsystem; the number of human genes in the subsystem; the intersection between the subsystem and the Brass *et al.* [144] siRNA screen; p-value for the latter; corrected p-value for the latter, calculated using the Benjamini and Hochberg FDR correction method; the intersection between the subsystem and the Konig *et al.* [145] siRNA screen; p-value for the latter; corrected p-value for the latter, calculated using the Benjamini and Hochberg FDR correction method; the intersection between the subsystem and the Zhou *et al.* [146] siRNA screen; p-value for the latter; corrected p-value for the latter, calculated using the Benjamini and Hochberg FDR correction method; the type of subsystem annotation used to identify the subsystem; the details of the subsystem annotation; the number of human genes from the sub-system that fit the subsystem annotation; the p-value for the subsystem annotation; a corrected p-value for the subsystem annotation, calculated using the Benjamini and Hochberg FDR correction method; number of genes from the Brass *et al.* siRNA screen that fit the subsystem annotation; p-value for the latter; corrected p-value for the latter, calculated using the Benjamini and Hochberg FDR correction method; number of genes from the Konig *et al.* siRNA screen that fit the subsystem annotation; p-value for the latter; corrected p-value for the latter, calculated using the Benjamini and Hochberg FDR correction method; number of genes from the Zhou *et al.* siRNA screen that fit the subsystem annotation; p-value for the latter; corrected p-value for the latter,

calculated using the Benjamini and Hochberg FDR correction method.

CHAPTER

FIVE

AN INTEGRATED TRANSCRIPTOMIC AND META-ANALYSIS OF HEPATOMA CELLS USED FOR HCV CELL CULTURE

5.1 Abstract

Hepatitis C virus (HCV) is a global problem. To better understand HCV infection researchers employ *in vitro* HCV cell-culture (HCVcc) systems that use Huh-7 derived hepatoma cells that are particularly permissive to HCV infection. A variety of hyper-permissive cells have been subcloned for this purpose. In addition, subclones of Huh-7 which have evolved resistance to HCV are available. However, the mechanisms of susceptibility or resistance to infection among these cells have not been fully determined. In order to elucidate mechanisms by which hepatoma cells are susceptible or resistant to HCV infection we performed genome-wide expression analyses of six Huh-7 derived cell cultures that have different levels of permissiveness to infection. A great number of genes, representing a wide spectrum of functions are differentially expressed between cells. To focus our investigation, we identify host proteins from HCV replicase complexes, perform gene expression analysis of three HCV infected cells and conduct a detailed analysis of differentially expressed host factors by integrating a variety of data sources. Our results demonstrate that changes relating to susceptibility to HCV infection in hepatoma cells are linked to the innate immune response, secreted signal peptides and host factors that have a role in virus entry and replication. This work identifies both known and novel host factors that may influence HCV infection. Our findings build upon current knowledge of the complex interplay between HCV and the host cell, which could aid development of new antiviral strategies.

5.2 Introduction

Hepatitis C virus (HCV) is prevalent in approximately 3% of the human population, though some countries, e.g., Eygpt, have a much greater prevalence [68]. The acute

phase of infection is often asymptomatic whereas chronic infection is a major cause of liver cirrhosis, hepatocellular carcinoma and liver transplantation. There is, however, no HCV vaccine, as high level of virion production, combined with the error-prone HCV RNA polymerase, causes frequent mutation of the viral genome resulting in production of immune escape mutants. Treatment of chronic HCV infection is currently based on interferon-α that evokes a general antiviral response and ribavirin, a nucleoside analogue. In combination, these antiviral agents do not reliably eradicate HCV in infected patients [71] and in addition, treatment is often interrupted due to the side effects that these drugs cause [72]. Therefore, development of improved anti-HCV drugs would be of great benefit.

Drugs that bind specific host proteins essential to the virus life cycle pose an attractive approach in viral disease therapy, as these targets have less potential for mutation and associated emergence of resistance than viral protein targets. Thus, anti-HCV drugs that bind specific host proteins are currently in development. For example alisporivir, a Cyclophilin A inhibitor, has recently entered phase II trials [79]. Cyclophilin A is essential for efficient HCV replication, probably due to direct physical interaction with NS5A and mediation of the viral polymerase [154, 299]. Also, inhibitors to microRNA mir-122, a molecule that regulates production of infectious virus particles, are also being investigated [80, 81]. By developing a greater understanding of the complex interplay between HCV and host cells, novel drug targets might be identified.

Significant advances in *in vitro* model systems to study HCV-host interaction have been made in the recent past [59]. Model systems greatly accelerated HCV research and led to production of a variety of genome-scale data sets including a host-virus interaction network [126], infection-induced changes in gene expression [300, 113] and host factors required for viral replication [154]. In addition to large data sets, numerous small-scale studies that use *in vitro* model systems have captured important details of key viral processes, such as virus cell entry [301]. However, we are still some considerable way from fully understanding the HCV life cycle and the role for each implicated host factor.

The initial breakthrough in HCV model systems allowed study of genomic viral RNA replication *in vitro* using replicons and permissive Huh-7 hepatoma cell lines [302, 303]. More recent HCV cell-culture (HCVcc) systems have permitted study of the entire virus life cycle and rapid cell-to-cell transmission, using a specific combination of the JFH-1 HCV strain and a particularly permissive hepatoma cell line (Huh-7.5.1) [304, 59]. The Huh-7.5.1 cells have a deactivating mutation in retinoic acid-inducible gene I (RIG-I), a protein that would normally bind to HCV RNA and initiate an interferon based antiviral response in the cell [305, 62]. Also, further subcloning of Huh-7.5.1 has led to the production of a more permissive cell (designated Huh-7.5.1c2 [306]), though the underlying mechanism of increased permissiveness in this subclone is not understood.

In addition to HCV susceptible cells, infection resistant Huh-7 derived cells have also been produced. One HCVcc study by Zhong *et al.* [307], a prerequisite to this study,

detected coevolution of JFH-1 HCV virus and Huh-7 and Huh-7.5.1 derived host cells. In particular, evolution of an increasingly aggressive virus was associated with emergence of several resistant cells. Follow-up analysis revealed that reduced cell surface expression of the CD81 viral coreceptor [308, 10] was partly responsible for resistance in a subset of these cells and that additional defects must be present that perturb the viral life-cycle. Therefore, mechanisms of both HCV resistance and susceptibility for Huh-7 derived cells are yet to be determined. This knowledge will be valuable for understanding specific host-cell dependencies in the viral life cycle and developing novel antiviral strategies.

In order to elucidate mechanisms by which hepatoma cells are susceptible or resistant to HCV infection we performed genome wide expression analysis of six Huh-7 derived cell cultures that have different levels of permissiveness to infection (Figure 5.1). To focus our investigation, we identified host proteins from HCV replicase complexes that were present in small vesicles located in the membranous web – a specific membrane alteration that is the site of HCV replication [309] – and also performed gene expression analysis of three permissive HCV infected cells. We found that a high number of genes, representing a wide spectrum of functions, including factors known to be involved in viral entry were differentially expressed between cells with different permissiveness to infection. Following this we conducted an in-depth analysis of differentially expressed host factors by integrating multiple data sources. Using this approach, we demonstrate that changes relating to susceptibility to HCV infection can be specifically linked to the innate immune response, secreted signal peptides, known host factors that influence virus entry and replication and putative, novel HCV infection-related host factors. In addition, our study also helps to characterise Huh-7 derived cells which may aid interpretation of results from subsequent studies that use HCVcc.

5.3 Materials and Methods

5.3.1 HCV-resistant cells R1.09, R1.10 and R2.1

R1 and R2 cells were obtained from stocks held at the The Scripps Research Institute, La Jolla, California, from a previous study [307]. Cryopreserved cells were thawed and put in culture. Initially, both cells displayed very limited viability after two independent thawing attempts of the original cryopreserved stock. Nevertheless, we were able to rescue both cell lines by slowly expanding the surviving colonies, subsequently labelled R1.1 and R2.1. To verify the resistance of these cell lines to HCV Con1 (genotype 1b) subgenomic (SG)-replicon replication, R1.1, R2.1 and the parental Huh-7.5.1 cells were transfected with the corresponding replicon RNA encoding a neomycin resistance gene and the formation of G418 resistant cell clones was monitored. R1.1 and R2.1 cells are partially resistant to HCV replication (supplementary Figure S5.2A). However, a higher percentage of cell clones in the R1.1 and R2.1 cell lines seem to support HCV replication than was previously observed in R1 and R2 cells before cryopreservation [307]. To

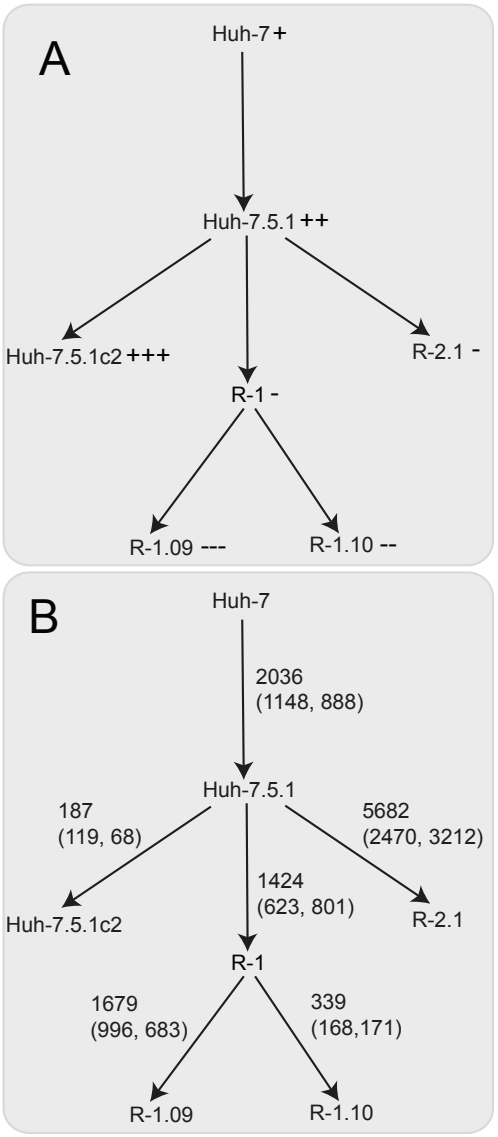

Figure 5.1: Tree of hepatoma cell cultures. Cell cultures are joined by arrows, going from the parent to the descendent, that indicate a subcloning event (or in the case of Huh-7 to Huh-7.5.1 a series of subcloning events). (A) The relative susceptibility of these cells to HCV infection where "+" represents susceptibility and "-" represents resistance and more symbols represent greater susceptibility or resistance. (B) Differential expressed genes between subclones. Differentially expressed genes were assigned to this tree either directly from expression comparison between cells or indirectly using a parsimony method. On each arrow, the first number indicates the total number of differentially expressed genes that have been attributed to the subcloning event. Below in brackets are the number of these genes that are (i) downregulated following subcloning, (ii) upregulated following subclononing.

verify that this result was reproducible for other replicon RNA preparations, we repeated the transfection experiment using Con1 full-length replicon RNA to select HCV replication resistant cell clones within the R1.1 and R2.1 cell populations (supplementary Figure S5.2B). R1.1 and R2.1 cells were subcloned by limiting dilution on feeder cells. Individual subclones were tested for resistance to subgenomic Con1 replicon replication and subclones R1.09, R1.10 and original R2.1 cells were selected for further analysis (supplementary Figure S5.2C). R1.09, R1.10, R2.1 and Huh-7.5.1 were harvested for microarray analysis.

5.3.2 HCV susceptible cells, Huh-7, Huh-7.5.1 and Huh-7.5.1c2

Three Huh-7 derived cell cultures that can support HCV infection [310, 311] were used in this study – Huh-7, Huh-7.5.1 and Huh-7.5.1c2. Cells were obtained from stocks held at the The Scripps Research Institute, La Jolla, California. All cells have been used in previous studies (Huh-7 [311], Huh-7.5.1 [304] and Huh-7.5.1c2 [306]).

The HCV susceptible cell types, Huh-7, Huh-7.5.1 and Huh-7.5.1c2, were infected with wild type JFH-1 (genotype 2a) virus at an moi=0.05. Infected cells were harvested for microarray analysis when virtually 100% of the cells were infected as determined by staining of cells for viral E2 protein at 3 (Huh-7.5.1c2), 4 (Huh-7.5.1) and 7 (Huh-7) days post inoculation (supplementary Figure S5.2D). Uninfected controls were harvested at the same time point as infected cells. The infectivity of supernatant produced from infected cells was measured at 2, 4 and 8 days post infection (supplementary Figure S5.2E). These results show that Huh-7 has the lowest susceptibility and Huh-7.5.1c2 the greatest susceptibility to HCV infection. In addition, uninfected Huh-7, Huh-7.5.1 and Huh-7.5.1c2 cells were harvested at 20 hours post infection in order to perform direct comparisons of their gene expression profiles by microarray analysis.

5.3.3 Preparation of microarrays

Cell cultures were centrifuged at 1000rpm for 5 minutes and the cell pellet was resuspent in 350μl lysis buffer (Qiagen). Each lysate was homogenised with a Qiashedder column (Qiagen) and the RNA was extracted using the RNeasy Mini Kit following the manufactures instructions. On-column DNA digestion was carried out by means of the RNase-Free DNase Set (Qiagen) and the integrity of the RNA was confirmed via Agilent (RIN of 9.7-10). 100ng of RNA from each sample was used to prepare cDNA with the Affymetrix GeneChip 3Õ IVT Express Kit and hybridised to Affymetrix U133 Plus 2 microarrays following the manufacturer's instructions.

The washing and staining procedure was performed in the Affymetrix Fluidics Station 450. The probe array was exposed to 10 washes in 6xSSPE-T at 250C followed by 4 washes in 0.5$\times$SSPE-T at 500C. The biotinylated cRNA was stained with a streptavidin-phycoerythrin conjugate, final concentration 2 mg/ml (Molecular Probes, Eugene, OR) in

6xSSPE-T for 30 min at 250C followed by 10 washes in 6×SSPE-T at 250C An antibody amplification step followed using normal goat IgG as blocking reagent, final concentration 0.1 mg/ml (Sigma) and biotinylated anti-streptavidin antibody (goat), final concentration 3 mg/ml (Vector Laboratories). This was followed by a staining step with a streptavidin-phycoerythrin conjugate, final concentration 2 mg/ml (Molecular Probes, Eugene, OR) in 6×SSPE-T for 30 min at 250C and 10 washes in 6×SSPE-T at 250C. The probe arrays were scanned at 560 nm using a confocal laser-scanning microscope (Affymetrix Scanner 3000 7G). CEL files were generated and used for further analysis. All microarray procedures were done at AROS, Denmark.

5.3.4 Computational analysis of microarray probe set intensity data

All analysis of microarray intensity data was carried out using R statistical software [312] including Bioconductor [313]. GeneChip® probe sets definitions were assigned using Entrez gene version 12.1 of a custom chip description file (CDF) from Psychiatry/MBNI Microarray Lab [92, 93, 94, 95]. This CDF included 17726 probe sets that correspond to an NCBI gene.

Initial quality checks of each chip were carried out using Bioconductor core tools and package affyQCReport [314]. Quality checks and visual inspection of array intensities showed that the quality of the array data was acceptable. Robust Multichip Average (RMA) expression values [100, 99, 95] were computed using the Bioconductor affy package. Genes that were called “not present" on all 39 GeneChip array data sets using Microarray Suite version 5.0 (MAS5) presence calls were removed [104], leaving a total of 13760 genes (supplementary table S5.10). This set was used throughout as a background for statistical tests and will be referred to as the microarray gene set.

Exploration of RMA expression values across all microarrays chips was performed by PCA using singular value decomposition (SVD). PCA was carried out using the Bioconductor package pcaMethods [315] and PCA results were plotted using the R package scatterplot3d. PCA results (supplementary Figure S5.2) show that biological replicates cluster together. The greatest variation is seen between replicates from experiment 1 that have been in culture for a longer time period with no JFH-1 infection. However, replicates that have been infected with JFH-1 cluster very closely, indicating a clear gene expression response to infection. Replicates from experiment 2 are all very tightly clustered and Huh-7.5.1 replicates from experiment 2 generally cluster with other uninfected Huh-7.5.1 derived cells. The PCA analysis results showed no major outliers or unexpected results and the array quality was shown to be acceptable. Therefore, the microarray data appeared sufficiently reliable to conduct analysis to identify differentially expressed genes.

Probability, false discovery rate (FDR) [193] and fold-change values for differential expression of genes between cells, using three biological replicates from each, were calculated in a pairwise manner using the limma method [105]. Genes were defined to be significantly differentially expressed if they achieved an FDR of < 0.01 and a minimum

fold-change of 1.5. Hierarchical clustering of genes across cells was performed in R using Pearson's r correlation (genes) and Spearman's rank correlation (cells) and results were visualised using the gplots package [316].

The design of our experiment permitted HCV infected cells Huh-7, Huh-7.5.1 and Huh-7.5.1c2 to be compared to uninfected controls harvested after 20 hours in culture and also uninfected controls harvested at the same time point as infected cells. We define those genes that are differentially expressed due to infection as genes that are significantly differentially expressed in the same direction (up- or downregulated) over both comparisons, as this will filter out genes whose expression fluctuates due to additional time spent in culture. We take the p-value, fdr and fold-change values from the comparison with the most conservative comparison.

MIAME compliant raw and processed gene expression data is freely available for download as a GEOarchive [87], accession number GSE29889.

5.3.5 Purification of crude replicase complexes (CRCs) for proteomic analysis by mass spectrometry

CRC preparations were produced using a protocol described by Quinkert *et al.* [317]. Protein samples from purified, proteinase K-treated CRCs were fractionated by one-dimensional SDS-PAGE and the gel was segmented according to molecular weight. Proteins contained in gel segments were digested with proteinase (trypsin/chymoptrypsin) prior to analysis by liquid chromatograph-mass spectrometry (WFZ Fungene, Greifswald). Using this method, 236 host-encoded proteins were identified as components of HCV replication complexes.

5.3.6 Collection of external gene sets

Four sets of genes that relate to replication of HCV were collected from external sources:

(i) Genes that encode products that interact with proteins of HCV were retrieved from two studies [126, 318]. We identified 465 such genes that correspond to an NCBI protein.

(ii) A non-redundant list of genes that have been identified by siRNA screen to play a significant role in HCV replication were obtained from five separate studies [150, 151, 152, 153, 154]. This list contains 399 genes including 363 that were shown to be necessary for propagation of HCV (proviral) in one or more study and 37 that have been shown to be detrimental to HCV propogation (antiviral) from the study by Brass *et al.*. Brass *et al.* identify 203 genes (193 proviral, 10 antiviral) that act during an early stage of infection and 59 genes (44 proviral, 15 antiviral) that act during a late stage in the viral life cycle.

(iii) Human host genes that are differentially regulated due to chronic infection by HCV were mapped from an *in vivo* study comparing chronically infected chimpanzees to uninfected controls [319].

(iv) 27 Cellular receptors and lipoproteins that are thought to have a role, either via positive or negative, in regulating HCV virion cell entry or particle release were manually curated from a recent review article [301]. These genes, by gene symbol are CD81, CD209, CLEC4M, CLDN1, CLDN6, CLDN9, SCARB1, LDLR, VLDLR, ASGR1, ASGR2, OCLN, APOC1, APOC2, APOE, APOB, MTTP, ISGF8, SAAL1, SAA4, EIF2A, EIF2AK2, IFNA2, IFNA5, IFNA8, IFNA16 and APOC3.

5.3.7 Assignment of differentially expressed genes to a cell tree and calculation of expression profile scores

A tree showing the lineage of relevant cells is shown in Figure 5.1. Changes in gene expression were assigned to branches of this tree directly from comparison of the ancestor and descendent cell, except for those branches linking R1 to other cells, which were imputed using a simple parsimony method. Genes identified as differentially expressed in comparisons Huh-7.5.1 versus. R1.09, Huh-7.5.1 versus R1.10 or R1.09 versus R1.10 were assigned to: (i) The branch linking Huh-7.5.1 and R1 if the gene is called differentially expressed and undergoes the same direction of regulation (up/down) in both comparisons involving Huh-7.5.1. (ii) The branch linking R1 to a descendent cell if the gene is only called differentially expressed with a given direction in just one of the two comparisons involving Huh-7.5.1. (iii) The branch linking R1 to a descendent cell if the gene is called differentially expressed in the comparison between R1.09 and R1.10; in this case, the descendent and direction of regulation is chosen from the comparison involving Huh-7.5.1 and a descendent where the largest fold change is observed.

To test whether genes appear more regularly on multiple branches of this tree than would be expected by random chance a permutation test was used. In a single permutation of this test we assign genes to branches of a model tree by randomly selecting them from the microarray gene set, selecting the same number of genes per branch as observed in the real tree. We then derive the frequency distribution for genes appearing in branches of the model tree. The frequency distribution for the model data is compared to the frequency distribution from the real data by Mann-Whitney U test to generate a p-value, testing the hypothesis that the real data values will be greater than that of the model data.

Using the tree and associated sets of differentially expressed genes, expression profiles consisting of an integer score per gene were derived. Where a gene changes regulation on a branch linking a more HCV susceptible parent cell to a more HCV resistant descendant cell, the profile scores -1 if the gene is upregulated (antiviral) and $+1$ if the gene is downregulated (proviral). Where a gene changes regulation on a branch linking a more resistant parent to a more susceptible descendant, the profile scores -1 if the gene is upregulated and $+1$ if the gene is downregulated. The overall gene profile score is calculated as the sum of these values over all tree branches. Genes that are present in the microarray gene set but not among the differentially expressed genes on this tree were assigned a score of zero.

A permutation test was used to test the hypothesis that a given gene set comprises genes with greater expression profile scores than would be expected by random chance. A distribution of scores from a set of subject genes were compared against the distribution of scores from all remaining genes from the microarray universe in a one-tailed Mann Whitney U test to identify whether the subject set has significantly greater scores than expected. The test statistic (U) for this test was recorded. We then repeated this test using a subject gene set of randomly selected genes, recording U for every permutation. 1000 permutations were carried out. The p-value was determined to be the proportion of times a more significant U value was generated by random permutation than for a real subject gene set.

A permutation test was used to test if genes from a given set have greater absolute expression profile scores given the number of times that the genes are differentially expressed, than would be expected by random chance. Using this measure we can ascertain whether the subject genes have a genuine tendency to be proviral or antiviral, or if they are simply hyper-variable in their gene expression. For each gene in the set we calculated a normalised score (s-norm), being the expression profile score divided by differential expression count. If the gene set has a significant tendency for the genes to be proviral or antiviral rather than hyper-variable, we would expect the s-norm to be greater than those for randomly selected genes. Hence, we test whether the genes of the subject set have a greater mean s-norm than a gene set of the same size, selected at random from the microarray gene universe. When randomly sampling genes, we maintain the distribution of differential expression counts observed in the subject gene set using rejection sampling. The p-value for the test was determined to be the total number of times that the random set has a greater mean average s-norm than the subject set, divided by the number of permutations. 1000 permutations were performed.

5.3.8 Analysis of the biological function of genes

Gene sets were subjected to functional enrichment using the Database for Annotation, Visualisation and Integrated Discovery (DAVID) version 6.7 functional annotation clustering and functional annotation chart tools [137, 108]. In both cases a custom background population consisting of the microarray gene set was used. All remaining DAVID 6.7 tools settings were left as the default. The set of differentially expressed genes from comparison between Huh-7.5.1 and R2 cells was limited to 3000 by selecting the set with lowest p-values for differential expression, as this corresponds to the maximum gene set size that can be analysed using DAVID. Annotation clusters were deemed to be significant if the enrichment score was > 2, this corresponds to a geometric mean from all term enrichment p-values of 0.01.

Functional clustering networks were produced using results from DAVID functional annotation clustering. Significant annotation clusters were represented as nodes, where the node diameter is proportional to the enrichment score. Edges signifying shared anno-

tation were created between nodes where the annotation clusters being represented share at least one quarter of annotating terms, where the edge diameter is proportional to the fraction of shared annotation terms. Networks were visualised using Cytoscape [164].

5.3.9 Construction of HCV protein network neighbourhoods

Human protein interaction data was retrieved from multiple sources compiled by the National Centre for Biotechnology Information (NCBI) and available as a download (ftp://anonymous@ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz). NCBI interactions data was downloaded on 10th August 2010. Physical protein-protein binding interactions were taken per gene, not including homo-dimer interactions (i.e., self edges in the network). All retrieved interactions, were used to compile a global interaction network where each interaction is treated uniformly, consisting of 48467 interactions between 10360 genes. HCV-human PPI data was retrieved from two HCV-human interaction studies [126, 318], on a per HCV protein-human gene basis and consisted of 533 interactions including 465 human genes and 11 HCV proteins. HCV protein network neighbourhoods were constructed that included specific differentially expressed genes and data from additional data sets. First a node corresponding to a HCV protein was created, next additional nodes corresponding to HCV-interacting host factors were added and finally we added nodes corresponding to differentially expressed genes that share an interaction with factors already present in the network. Finally, nodes that correspond to those non-differentially expressed host genes that do not share an interaction with a differentially expressed gene were pruned. The result is a network where the maximum path length from the HCV protein is two and the maximum path length for a non-differentially expressed host gene is one. Networks were visualised using Cytoscape [164].

5.4 Results/Discussion

5.4.1 HCV infection causes significant changes to gene expression

Expression analysis of three hepatoma cell cultures – Huh-7, Huh-7.5.1 and Huh-7.5.1c2 – that are susceptible to infection was performed. By comparing infected and uninfected cells we identified genes that are differentially expressed (false discovery rate corrected p-value < 0.01), in all cell lines: 1743 genes in Huh-7 (1525 upregulated, 218 downregulated); 7025 in Huh-7.5.1 (3503 upregulated, 3522 downregulated); and 3891 in Huh-7.5.1c2 (1485 upregulated, 2406 downregulated, supplementary table S5.4). This represents 7475 distinct genes – 54% of all genes analysed. 3835 of these genes (51%) were differentially expressed in more than one cell and 1181 (16%) were differentially expressed with the same response to infection (up- or downregulation) in all three comparisons. A hierarchical clustering plot (Figure 5.2) shows a clear pattern of gene expression that corresponds to infection by HCV. Intersections between these gene sets are shown in

figure 5.3.

Previously, Woodhouse *et al.* [320] performed whole genome expression analysis of Huh-7.5 cells infected with JFH1 HCV, harvested at the peak of infection and identified 1351 differentially expressed genes. Though we identify more differentially expressed genes, our results overlap significantly with those of Woodhouse *et al.* for all three cells ($p < 0.01$, Fisher's exact tests) and the response to infection of genes identified in both studies is well conserved at 90-99% for each cell line.

By performing functional annotation clustering on the subset of 1181 genes found to be differentially expressed with the same response to infection over all infected versus control cell comparisons (Figure 5.3, section G), we identifiy a core set of host cellular functions that are affected by HCV infection (supplementary table S5.5). Interestingly, the two most enriched clusters comprise genes involved in transcription. Zinc-finger domain containing proteins are highly over-represented in this set (249 genes are annotated with the SwissProt keyword "zinc-finger" [321], false-discovery-rate corrected p-value of 8.6×10^{-29}). The change in expression of such a large number of zinc-finger domain encoding genes remains unexplained, particularly as many of these factors are not known to be associated with viral infection. However, of these 249 proteins, 143 are also annotated with the SwissProt keyword "transcription regulation". Given the scale of change in gene expression between infected and uninfected cells, extensive change in the expression of transcriptional regulators is fitting.

In addition, other cellular processes including microtubule organisation, ubiquitin and ubiquitin-like (ubl) conjugation pathway and DNA repair (particularly DEAD and DEAH box helicases) are enriched. HCV requires a functional microtubule network for entry into Huh-7.5 cells and early post-entry steps of infection through interaction with tubulin proteins [322]. We identify 11 tubulin isoforms that are either up- or down-regulated during HCV infection, indicating that HCV infection may exert control over the microtubule network at the level of transcription.

DEAD box helicases, RIG-I and IFIH1, are interferon stimulated genes (ISGs) that act to detect RNA viruses and initiate further interferon production [62]. Another DEAD box helicase, DDX3X, encodes a factor that is required for successful HCV replication [323, 151, 154]. However, DDX3X can also cause immune activation [324] and the role for this protein in HCV infection is unclear. RIG-I is transcriptionally upregulated in Huh-7 and Huh-7.5.1 cells but not Huh-7.5.1c2 following infection, and IFIH1 is transcriptionally upregulated in Huh-7.5.1 cells but not Huh-7 or Huh-7.5.1c2 following infection. These results indicate a potential weakness in the innate immunity of Huh-7.5.1c2 at the level of gene expression.

Virally triggered RIG-I mediated antiviral signaling evokes the production of type I interferon [325]. However, in our results we do not observe increase in transcription of type I interferon in either Huh-7 cells that have functional RIG-I or Huh-7.5 derived cells whose RIG-I gene has a known deactivating mutation [305]. This result suggests

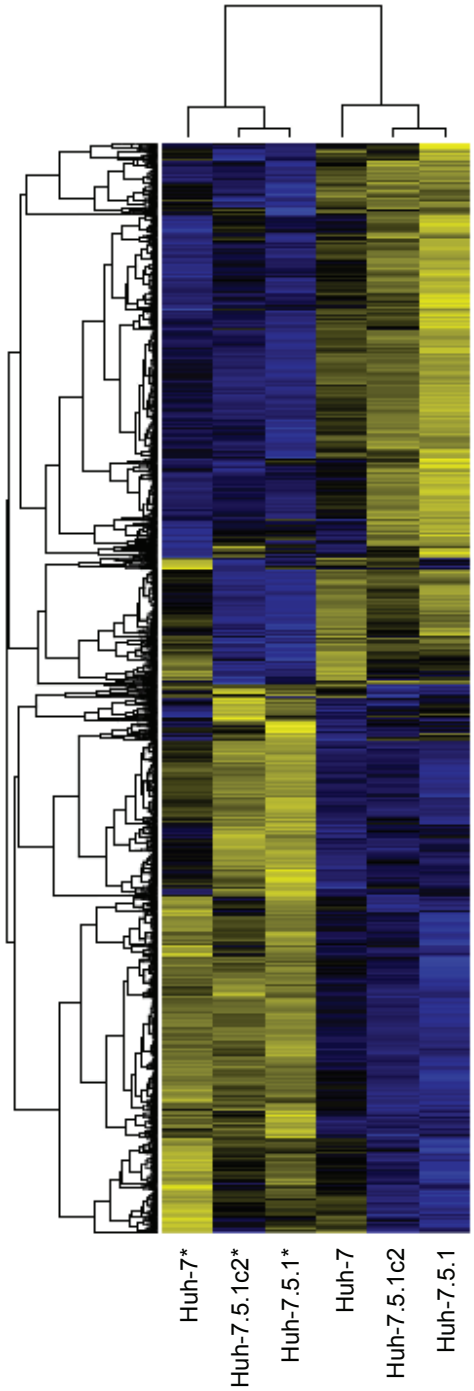

Figure 5.2: Hierarchical clustering plot displaying differentially expressed genes from infected and control cells. Genes are represented by horizontal bands and cells by columns. Infected cells are denoted with an asterisk (*). Bands are coloured blue if the gene is downregulated and yellow if they are upregulated compared with the mean expression level for that gene. Greater colour intensity signifies greater fold change. Infected cells cluster with one another and gene clustering shows a clear pattern that corresponds to HCV infection induced expression.

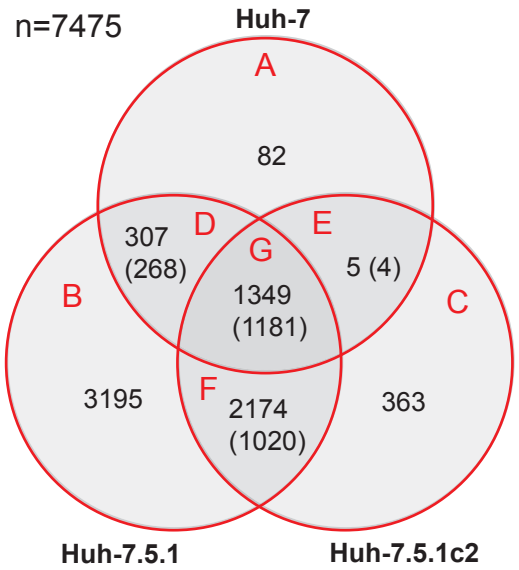

Figure 5.3: Venn diagram showing the overlap in genes differentially expressed due to HCV infection among susceptible cells. The absolute numbers of significantly differentially expressed genes are given for Huh-7, Huh-7.5.1 and Huh-7.5.1c2 cells. The numbers in brackets refer to those genes that share the same direction of regulation (up- or downregulated) following infection, across multiple comparisons.

that HCV successfully attenuates interferon production. The virus can achieve interferon attenuation through several mechanisms including NS3/NS4A protease activity that disrupts both RIG-I and toll-receptor signaling [325]. The regulation of ISGs following infection of hepatoma cells was investigated. A list of ISGs was obtained from the ISG database [114] and a total of 455 genes were present in the ISG data and also present in our microarray gene set. We find that 62, 198 and 160 ISGs are differentially expressed in Huh-7, Huh-7.5.1 and Huh-7.5.1c2 cells, respectively, following HCV infection. However, these values do not represent statistically significant enrichment of ISGs, consistent with our observation regarding lack of significant transcriptional upregulation of interferon.

Ubiquitin conjugation has been identified as an important cellular function for both viral and bacterial pathogens [326]. Firstly, deubiquitylating enzymes (DUBs), such as ubiquitin specific peptidases (USPs) can modulate host cell innate immunity [326]. Ubiquitin-like protein ISG15 is an ISG that is expressed following infection and target proteins become “ISGylated” following conjugation of ISG15. Ubiquitin specific peptidase USP18 is involved in deISGylation to attenuate innate immunity [327, 326]. Another

USP, USP7, is targeted by viral proteins. USP7 interacts with both the herpes-simplex virus protein ICP0 and Epstein-Barr nuclear antigen I and may have a role in regulation viral replication [328, 329]. We find that 27 USPs (though not including those specific USPs mentioned) are differentially expressed in one or more infected versus uninfected comparison and given all comparisons there are 59 instances of differential expression of these genes from which 56 instances identify USP as transcriptionally upregulated in the infected cell.

Another DUB that has a role in pathogenic infection is CYLD. CYLD expression is induced in cells infected with *Haemophilus influenza* and the absence of this gene confers hypersensitivity to this bacterial pathogen [326]. In our results, CYLD is also upregulated in all HCV infected cells. Therefore it seems likely that DUB upregulation is a significant marker for HCV infection in these cells. Interestingly, HCV NS5A has been shown by a yeast-two-hybrid assay to interact with USP19 [126], a DUB known to positively regulate cell proliferation [330]. The functional role of this protein interaction is not known and further experimental validation and investigation could provide valuable insight into HCV infection.

Functional annotation clustering was repeated on gene sets from sections A, B, C and F of the Venn diagram in Figure 5.3 (see supplementary table S5.5). The intersection of genes regulated with the same response from Huh-7.5.1 and Huh-7.5.1c2 but not Huh-7 infection studies (Figure 3, section F, 1020 genes) is enriched for functions that can be directly attributed to heightened HCV infection, e.g., transforming growth factor β signaling [331] and a generally heightened metabolic state, e.g., positive regulators of transcription. Interestingly, one function whose enrichment is found in section C of the Venn diagram (corresponding to the set of genes differentially expressed following infection of Huh-7.5.1c2 cells), but not sections A or B (corresponding to infection of both Huh-7 and Huh-7.5.1), is apoptosis. More specifically, the most overrepresented annotation terms in this cluster refer to the negative regulation of cell death and these genes are predominantly upregulated in infected Huh-7.5.1c2. Specifically, there are 21 genes annotated with the GO term *negative regulation of cell death* and 16 of these are transcriptionally upregulated in infected Huh-7.5.1c2 cells. For example, NFKB and BCL2 genes are established as anti-apoptotic proliferative factors in human cancers and both are upregulated in infected Huh-7.5.1c2. Apoptosis is an important defense mechanism against infection that is initiated by the innate immune response [332] and this result indicates that Huh-7.5.1c2 could be a more permissive host for HCV than either Huh-7 or Huh-7.5.1 by being less prone to apoptosis.

5.4.2 Subclones of Huh-7 derived cells have significantly altered gene expression

We performed gene expression analysis on six cell cultures that display a range of susceptibilities to HCV infection: HCV susceptible Huh-7, Huh-7.5.1 and Huh-7.5.1c2 and

HCV resistant subclones of Huh-7.5.1, R1.09, R1.10 and R2.1. Differentially expressed genes were detected in all comparisons with a false discovery rate corrected p-value of < 0.01 and minimum fold-change of 1.5. To identify differences in gene expression between these cells, 'parent-child' cell comparisons were made (see table 5.1 for a summary and supplementary S5.4 for full details).

Table 5.1: The number of differentially expressed genes identified in pairwise comparisons between hepatoma cell subclones. The genes were found to be differentially expressed with a false discovery rate corrected p-value of < 0.01 and fold-change > 1.5.

Original cell	Subclone cell	Total DE genes	Downreg in subclone	Upreg in subclone
Huh-7	Huh-7.5.1	2036	1148	888
Huh-7.5.1	Huh-7.5.1c2	187	119	68
Huh-7.5.1	R1.09	2830	1534	1296
Huh-7.5.1	R1.10	1714	771	943
Huh-7.5.1	R2.1	5682	2470	3212

The pattern of gene regulation highlighted in heatmaps (Figure 5.4) correlates with the subcloning of these cells, where the 'child' subclone retains a significant proportion of the gene expression profile of the 'parent'. For example, many of the same genes are found to be differentially expressed with the same direction of regulation in the comparisons: (i) Huh-7.5.1 versus Huh-7 and Huh-7.5.1c2 versus Huh-7 (1479 genes in common) and (ii) R1.09 versus Huh-7.5.1 and R1.10 versus Huh-7.5.1 (1424 genes in common). Therefore, we represent cells and differential expression on a hierarchical tree structure (Figure 5.1). This shows all six cells and the total numbers of differentially expressed genes, both upregulated and downregulated. A full list of differentially expressed genes for each branch is given in supplementary table S5.6. A total of 7503 genes are differentially expressed following subcloning events. This represents a substantial proportion of both all genes on the microarray (42%) and the subset of those that are expressed in these cells (54%). This indicates that multiple changes in expression could contribute to the susceptibility to infection found among the hepatoma cells.

Genes that are differentially expressed following subcloning events were found to be enriched for specific biological annotation using functional annotation clustering. A brief summary of the most significant annotation clusters identified among each set of genes is given in table 5.2 and a full list of results is given in supplementary table S5.7. From table 5.2 it is clear that some areas of biological annotation are significantly enriched among more than one set of genes. For example, an annotation cluster corresponding to secreted glycoproteins and signal peptides appears in five out of the six sets and three out of five also include an annotation cluster that corresponds to proteins of the acute inflammatory response.

To assess (i) overlap in biological function between each set of differentially expressed genes, (ii) overlap in biological function these gene sets may have with other genes that

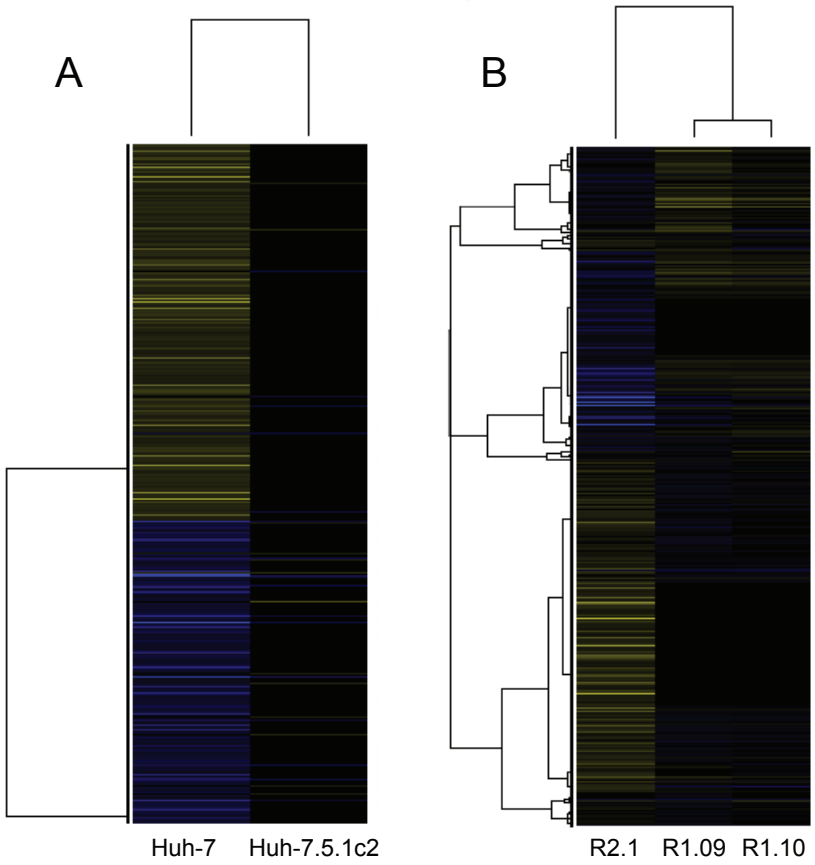

Figure 5.4: Hierarchical clustering plots showing the expression levels of differentially expressed genes between hepatoma cells. Genes are represented by horizontal bands and cells by columns. Bands are coloured blue if the gene is downregulated and yellow if they are upregulated relative to their expression in Huh-7.5.1. Greater colour intensity relates to a greater fold change. Black bands represent genes whose expression is a similar level to Huh-7.5.1. (A) Comparison of gene expression levels between susceptible cells. Here, the Huh-7.5.1c2 cell line is clearly more similar in gene expression to Huh-7.5.1 than Huh-7. (B) Comparison of gene expression levels between resistant cells and Huh-7.5.1. The R2.1 cell line is more divergent from Huh-7.5.1 than either R.109 or R1.10 in terms of gene expression. R1.09 and R1.10 show similar patterns of gene expression.

Table 5.2: Functional enrichment among significantly differentially expressed genes between original cells and subclones. Shown are the number of functional annotation clusters that achieve an enrichment score (ES) of > 2 and the number of differentially expressed genes that these clusters include. The right-most three columns give details of the top scoring annotation clusters (a maximum of 5 are shown)

Original cell	Subclone	No. clusters $ES > 2$	No. genes $ES > 2$	Top 5 clusters ES	No. genes	Annotations
Huh-7	Huh-7.5.1	15	866	5.81	565	Extracellular and secreted; signal peptide; glycoprotein; disulfide bond.
				4.93	119	Response to hormone stimulus and organic substance.
				3.92	47	Response to steroid hormone and glucocorticoid stimulus.
				3.52	48	Response to extracellular stimulus, nutrients, retinoic acid and vitamin A.
				3.37	37	Complement and coagulation cascades; acute inflammatory and defense response.
Huh-7.5.1	Huh-7.5.1c2	2	108	7.09	83	Glycoprotein; signal peptide; secreted; disulfide bond.
				4.93	74	Response to hormone stimulus and organic substance.
Huh-7.5.1	R1.1	18	704	10.55	416	Extracellular and secreted; signal peptide; glycoprotein; disulfide bond.
				5.03	88	Response to wounding; acute inflammatory and defense response.
				4.42	113	Mitosis; organelle fission; cell-cycle; M-phase.
				4.23	54	Enzyme inhibitor; endopeptidase and protease inhibitor; SERPIN family; reactive bond.
				3.60	50	Proteinaceous extracellular matrix; basement membrane.
R1.1	R1.09	13	582	11.32	108	DNA replication and DNA metabolic process.
				7.07	151	Mitosis; organelle fission; cell division; chromosome segregation;cell-cycle; M-phase.
				6.41	96	Chromosomal part; centromeric region; chromatin.
				5.46	140	DNA repair; response to DNA damage; stress response.
				4.85	45	Condensed chromosome; kinetochore; centromeric region .
R1.1	R1.10	2	108	4.12	55	Extracellular space.
				4.09	108	Secreted; signal peptide; glycoprotein; disulfide bond.
Huh-7.5.1	R2	25	1325	6.34	165	Sequence-specific DNA binding; Homeobox DNA binding domain.
				4.97	95	Embryonic morphogenesis; appendage development.
				4.93	40	Acute inflammatory response; acute phase.
				4.74	153	Acute inflammatory, wounding and defense response.
				4.59	73	Extracellular and secreted; signal peptide; glycoprotein; disulfide bond.

relate to HCV infection and (iii) to identify potential functions that contribute to susceptibility to HCV infection, we created a functional clustering network using all significant annotation clusters described in supplementary table S5.7. This functional clustering network comprises 36 subnetworks, shown in supplementary file S5.1. These subnetworks correspond to areas of shared, enriched biological function between the gene sets. Figure 5.5 shows 12 of these subnetworks that include at least three nodes, at least one node corresponding to an enriched function from a subcloning event and at least one node corresponding to an enriched function from an additional data set linked directly to HCV infection. These visualisations highlight the possibility that changes in expression of genes of some of these particular functions may contribute to susceptibility to HCV infection in more than one subcloning event. For example, Figure 5.5B shows that genes encoding protein products that interact with proteins of HCV and also genes that are differentially expressed between several independent subcloning events (corresponding to both increase and decrease in susceptibility to HCV infection), are all enriched for extracellular and secreted disulphide-bond containing proteins and signal peptides.

The subcloning of hepatoma cells has caused extensive changes to transcriptional activity, both in terms of the absolute number of differentially regulated genes and of biolog-

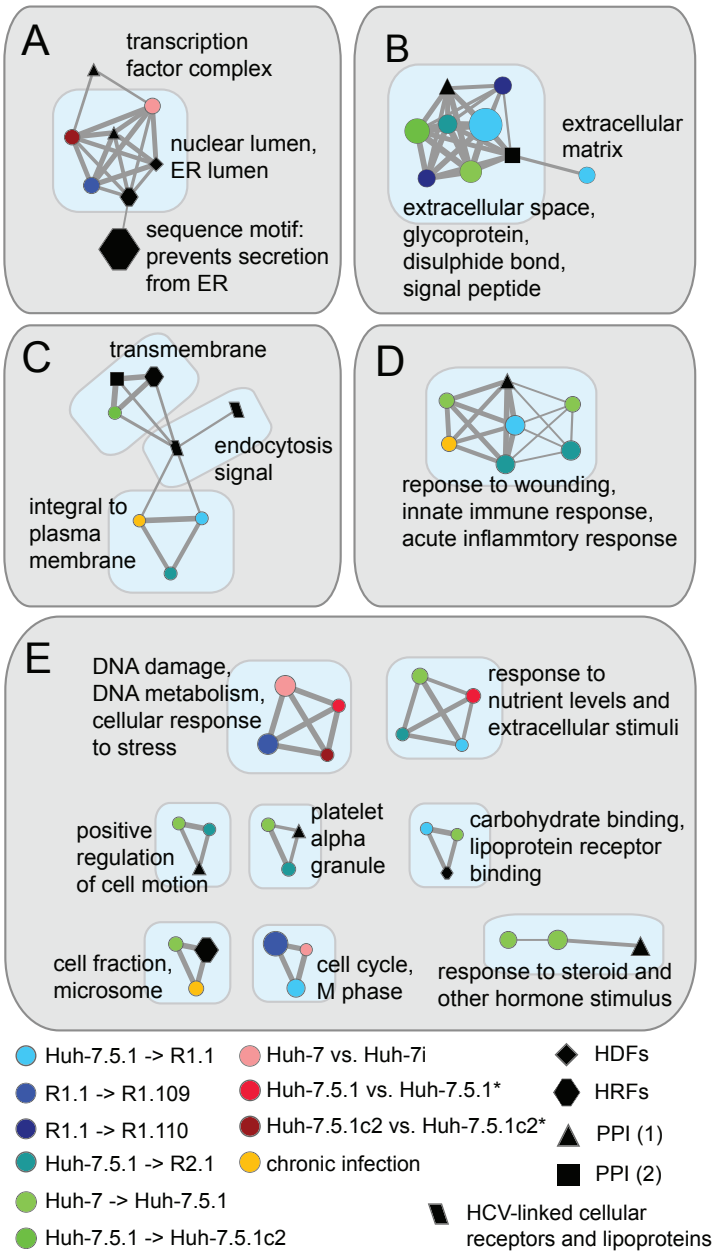

Figure 5.5: Functional annotation cluster networks from differentially expressed genes and other HCV-related data sources. These networks highlight areas of shared enriched function between gene sets that we identify as differentially expressed between hepatoma cells and also gene sets that relate to HCV infection. Nodes represent annotation clusters from the data source denoted by the node colour. Edges represent shared annotation terms between clusters. Only nodes that share at least $1/4$ of annotating terms are connected by an edge. Node diameter is proportional to the level of enrichment of the biological function in the gene set. Edge width is proportional to the proportion of annotating terms shared between two clusters. Subnetworks A-D are those with > 6 nodes, subnetworks shown in E have between 3 and 6 nodes. Annotation clusters from two PPI data sources are shown: PPI (1) from reference [126] and PPI (2) from reference [318].

ical functions affected. In the case of Huh-7.5 cells, RIG-I mutation is known to increase susceptibility to HCV infection and complementing these cells with wild-type RIG-I induces greater resistance [305]. However, differential expression of over 2000 genes from a variety of functions between Huh-7 and Huh-7.5.1 is not necessarily due to a single mutation of RIG-I, indeed this seems unlikely. Therefore, it is impossible to say whether Huh-7.5.1 derived cells are more susceptible than Huh-7 due to RIG-I alone, as change in regulation of other genes may also play a role. Huh-7.5 derived cells are commonly used for HCVcc but the extent to which subcloning-induced cellular alteration distances these cells from hepatocytes that are being modeled warrants greater consideration given the scale of change we report.

A previous study by Inoue *et al.* [333] made comparison of two Huh-7 subclones that had varying HCV replication efficiency. Inoue *et al.* identify 17 genes that have an increased level of expression and 19 genes that have a decreased level of expression in the more efficient of the cells. Though the present study and that of Inoue *et al.* have a shared aim, there is very little concurrence of results. This could be because Inoue *et al.* observed a different mechanism causing change in susceptibility to infection but it could also be due to the relatively small size of their result set. Regardless, our study has greater power as six Huh-7 derived subclones rather than two were analysed and expression of approximately 17 thousand genes, as opposed to approximately 8500 were assessed.

5.4.3 Host factors linked to HCV are differentially expressed in subclones of Huh-7

HCV dependency factors

The 7503 genes differentially expressed following subcloning events are enriched for genes shown by siRNA gene knockdown to be necessary for HCV replication [150, 151, 152, 153, 154], termed HCV dependency factors (HDFs). A total of 292 genes that are expressed among the six cells are among HDFs and 176 of these genes are differentially expressed ($p = 0.050$, Fisher's exact test). This result indicates that differences in expression are likely to impact susceptibility of the cells to infection.

Cellular receptors and lipoproteins

Cellular receptors and lipoproteins involved in HCV entry are differentially expressed in comparisons between both resistant and susceptible cells. Five genes – CLDN1, CD81, LDLR, ASGR1 and APOE – that promote virus entry are differentially expressed in a comparison between susceptible cells (Figure 5.6). Of these five genes, all except APOE are downregulated in the Huh-7.5.1 cell, relative to Huh-7. APOE is transcriptionally upregulated in Huh-7.5.1 relative to Huh-7 (a fold-change of 1.84) and ASGR1 is upregulated in Huh-7.5.1c2 relative to Huh-7.5.1 (a fold-change of 1.57). Therefore, only APOE and ASGR1 are regulated in a manner that fits with the observed susceptibility to

infection and neither undergo a substantial fold-change, thus, it does not seem likely that enhanced viral entry is a cause of the relative permissiveness to infection in Huh-7.5.1 and Huh-7.5.1c2 cells.

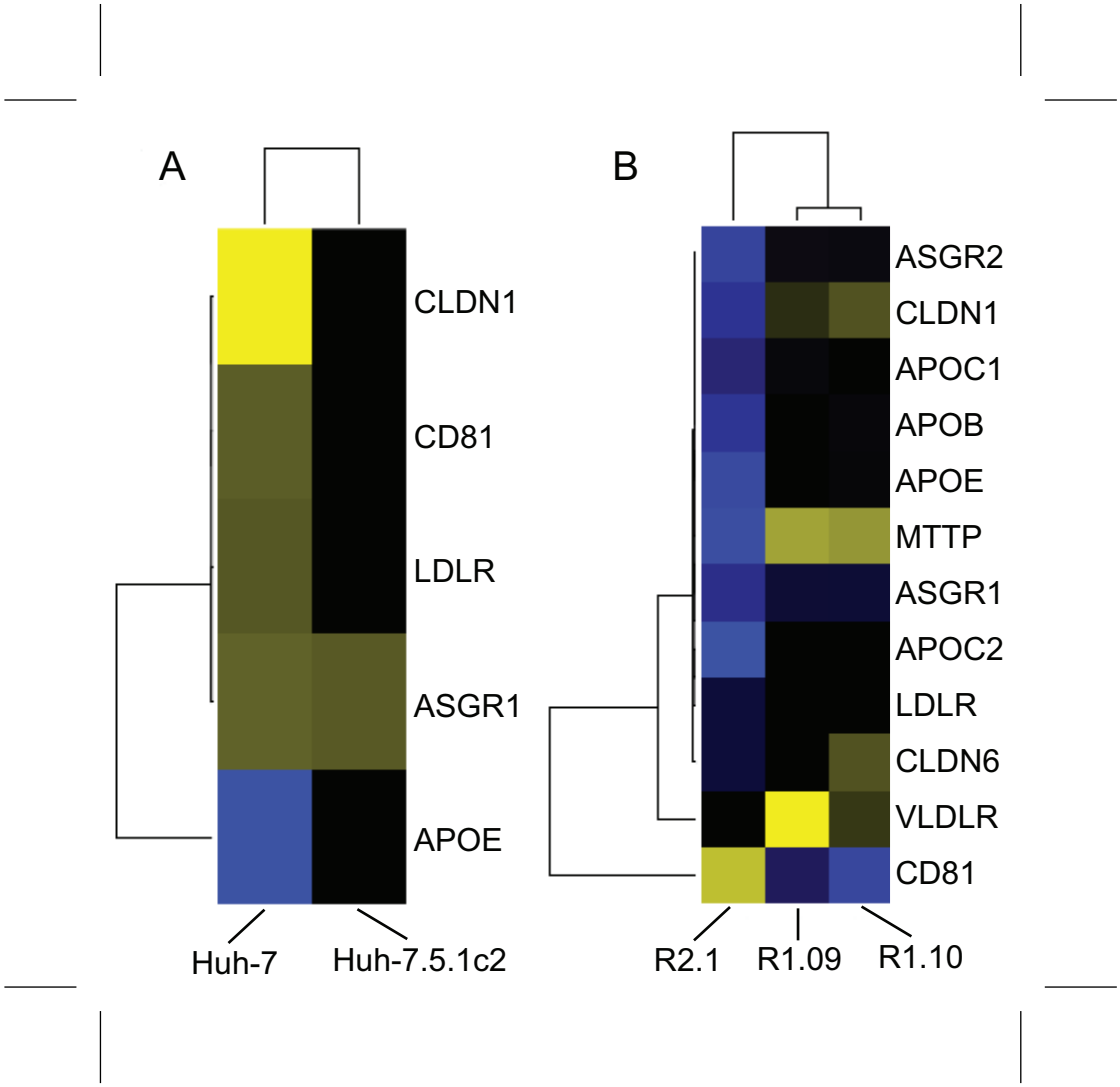

Figure 5.6: Hierarchical clustering plots showing the expression levels of differentially expressed HCV-linked cellular receptors and lipoproteins. Genes are represented by horizontal bands and cells by columns. Bands are coloured blue if the gene is downregulated and yellow if they are upregulated, relative to their expression in Huh-7.5.1. Greater colour intensity relates to a greater fold change. Black bands represent genes whose expression is a similar level to Huh-7.5.1. (A) Comparison of gene expression levels between susceptible cells. The majority of entry factors that undergo a significant change in expression are found at a higher level in Huh-7 than either of the Huh-7.5.1 derived cells. (B) Comparison of gene expression levels between resistant cells and Huh-7.5.1. R1.09 and R1.10 cells have have a lower level of expression of CD81 than Huh-7.5.1. Though R2.1 cells have a relatively high level of CD81 expression relative to Huh-7.5.1, other entry factors are expressed at lower levels.

Ten factors that influence HCV entry are differentially expressed in comparison between resistant cells and the Huh-7.5.1 parent (DD81, ASGR1, ASGR2, CLDN1, CLDN6, VLDLR, LDLR, APOC1, APOC2 and APOE). CD81, an important coreceptor in HCV

cell entry [301], is downregulated in both R1.09 and R1.10 relative to Huh-7.5.1. CD81 expression is significantly greater in R2.1 than Huh-7.5.1. However, eight of the nine remaining differentially expressed entry factors, all except for VLDLR, are downregulated in R2.1 relative to Huh-7.5.1, including four by more than 32-fold (APOC2, APOE, ASGR1 and CLDN1). These results are consistent with the findings of Zhong *et al.* who identify that low ectopic CD81 contributes to the resistance of the original R1 (but not R2) cell, from which R1.09 and R1.10 are descended. However, additional mechanisms of resistance must exist, as transduction of R1 cells to express CD81 did not fully restore the susceptibility to infection observed in Huh-7.5.1 [307]. From our expression analysis it does not appear that R1.09 and R1.10 cells lack other cell entry factors. Therefore, the mechanism of resistance to infection, additional to CD81-mediated entry in R1 derived subclones, is unlikely to be due to viral entry. Zhong *et al.* attribute infection resistance of R2 to processes other than CD81-mediated entry. However, we identify that many entry factors aside from CD81 are downregulated in R2.1. These factors include CLDN1, a component of tight junctions, the silencing of which prevents HCV entry into Huh-7.5 cells [334]. CLDN1 is transcriptionally downregualted in R2.1 compared to Huh-7.5.1, with a fold-change of approximately 36-fold, indicating impeded viral entry could contribute to R2.1 resistance to infection. However, like R1 and Huh-7.5 derived cells, evidence suggests that processes other than viral entry affect the permissiveness of R2.1 to infection by HCV; particularly as Zhong *et al.* show that HCV replicon replication was defective in R2 cells and over five thousand genes are differentially expressed following subcloning of R2.1 cell from Huh-7.5.1. Indeed, MTTP and APOB are associated with HCV particle formation and particle secretion [301]. Like the virus entry factors previously mentioned, both MTP and APOB are downregulated in R2.1 relative to Huh-7.5.1 by more than 32-fold. Therefore, it appears that R2.1 cells may also lack the ability to support aspects of the HCV replication cycle that take place post-entry.

Proteins associated with the HCV replicase complex

Host proteins from crude HCV replicase complexes were identified by mass spectrometry. These host proteins correspond to 236 host genes that we term host replication factors (HRFs, supplementary table S5.8). A total of 212 HRFs were expressed among the cells that underwent microarray analysis and 145 of these are differentially expressed. This is significantly more than would be expected by random chance ($p = 3.8 \times 10^{-5}$, Fisher's exact test). This result suggests that the ability of these cells to support replication of viral RNA is unlikely to be consistent between these cells. Interestingly, among these 145 host genes are 12 (APOA1, APOE, CALR, CANX, FTH1, GNB2, HSPA5, OS9, PFN1, PPIB, SSR4, and TUBB2C) that encode a product known to interact with one or more HCV protein [126]. For example, Chang *et al.* show that APOE is required for production of infectious HCV, probably for virion assembly rather than viral RNA replication [335] and CANX is involved in the folding of HCV glycoproteins [336].

Also among the 145 are two HDFs, DDOST and PPIA [154]. DDOST encodes a subunit (dolichyl-diphosphooligosaccharide-protein glycosyltransferase) of the oligosaccharyltransferase complex. DDOST is required during the late stages of HCV replication, possibly to perform an essential glycosylation step on HCV envelope proteins, E1 and E2 [154]. PPIA encodes a cyclophilin A the protein target of anti-HCV drug alisporivir [79]. Both DDOST and PPIA are downregulated in the R2.1 subclone compared to Huh-7.5.1 with fold change 2.16 and 1.82, respectively. APOE encodes apolipoprotein E, a constituent of lipoproteins. Surprisingly, the level of APOE gene expression in R2.1 cells is lower than in Huh-7.5.1 by over 100-fold. Taken together, these results suggest that the regulation of expression of DDOST, PPIA and particularly APOE might be sufficiently altered in R2.1 such that the cell are unable to form a competent HCV replication complex.

5.4.4 Gene expression profiles highlight host factors and biological functions that are linked to HCV infection susceptibility

A significantly greater proportion of expressed genes appear on multiple branches of the tree of cell subclones (Figure 5.1) than would be expected by random chance (no Mann Whitney U test p-value exceeded 0.001 in 1000 permutations). This result indicates that the likelihood of undergoing a significant change in expression following subcloning is not equal for each gene. This may be due to a number of reasons including simple hyper- or hypo-variability of certain genes, or, more interestingly, some genes being differentially expressed multiple times during subcloning due to an effect they have on HCV infection susceptibility, their change in expression having been selected by subcloning.

To distinguish factors that may alter HCV susceptibility and to identify specific biological functions and proteins that may contribute to HCV infection susceptibility, we define a gene expression profile score that accounts for significant change in expression following independent subcloning events (see Materials and Methods). A negative score represents an antiviral expression pattern a positive score represents a proviral expression pattern. The frequencies of attained scores are given in table 5.3 and a full list of scores per gene is given in supplementary table S5.9. To demonstrate the significance of our measure, we tested whether other gene sets that are linked to HCV virus propagation have greater scores than would be expected. We find that HCV-linked cellular receptors and lipoproteins (including many factors involved with cell entry), genes that encode proteins that interact with HCV proteins and HRFs have a greater mean score than expected by random chance. The test result was not significant for HDFs (see table 5.4). These results indicate that our score is significant and is a useful measure for aiding identification of host cell factors that affect susceptibility to HCV infection. Furthermore, these results indicate that factors involved in virus entry into the cell, replication and those that have a direct association with proteins of HCV are likely to be important.

Though our score does penalise expression profiles that exhibit both antiviral and

Table 5.3: The frequency of gene scores. Here, we give the profile scores, the frequency of the score among 7503 differentially expressed genes and the corresponding proportion. A negative score represents an expression profile that indicates a possible antiviral activity, whereas a positive score indicates a possible proviral activity.

Score	Frequency total	Proportion
-5	1	0.013%
-4	11	0.15%
-3	77	1.03%
-2	592	7.89%
-1	3039	40.50%
0	1042	13.89%
1	2054	27.38%
2	554	7.38%
3	113	1.51%
4	19	0.053%
5	1	0.013%

Table 5.4: Mean profiles scores. Details of the profile scores of four gene sets that we predict may have a higher score than would be expected by random chance. These gene sets are: HCV-linked cellular receptors and lipoproteins, the majority of which facilitate virus entry (but also and particle formation and release, here labeled "receptors"), host factors that we isolate from vesicles that harbour the HCV replication complex (HRFs), host factors that are required for HCV replication determined by siRNA screen (HDFs) and HCV interacting proteins (HCV interacting). For each gene set we show the mean profile score and the significance of the score enrichment, determined by Mann Whitney U test permutation. For all gene sets other than HDFs, the profile scores are on average greater than we would expect by random chance, given a p-value cutoff of < 0.05.

Gene set	No. of genes	Mean profile score	Permuted p-value
Receptors	14	0.5	0.0070
HRFs	173	0.283	< 0.001
HDFs	230	-0.0174	0.173
HCV interacting	465	0.0452	0.0030

proviral tendencies, we would still expect biological functions that comprise genes that are hyper-variable in their expression in the hepatoma cell culture system to have a greater range of scores than biological functions whose genes tend to be expressed at a constant level. Therefore, we devised a test to ascertain whether the enriched functions that we highlight in Figure 5.5 are simply hyper-variable or are consistent with a profile that corresponds to antiviral or proviral action (see Materials and Methods for details of this test).

We find that two areas of function are linked to the observed differences in susceptibility to HCV infection in more than one of the six cell cultures: (i) secreted signal peptides and glycoproteins (343 genes, Figure 5.5B) and (ii) the acute and innate inflammatory re-

sponses (71 genes, Figure 5.5D) were significant ($p = 0.049$ and $p = 0.013$ respectively). The innate immune response and particularly interferon-stimulated pathways play an important role in cellular defense against viral infection [337]. "Secreted signal peptides and glycoproteins" is a description relevant to proteins from a broad spectra of activities and there are several important factors among these that are differentially expressed which directly relate to HCV infection. These include TGF-β [338], low-density lipoprotein receptors and their associated proteins that have previously been discussed and TNF and serpin peptidase inhibitors [301, 339] (discussed in the next section).

We also define a set of 'high-scorers' – genes with an absolute score ≥ 3. There are 222 high-scorers, representing approximately the top 3% of differentially expressed genes. Genes that scored > 3 or < 3 are listed in table 5. Among high-scorers are two HRFs, neutral cholesterol ester hydrolase 1 (NCEH1) and visinin-like 1 (VSNL1). NCEH1 catalyses hydrolysis of intracellular cholesterol ester, to produce free cholesterol. Free cholesterol may then be re-esterified or efflux to an extracellular cholesterol acceptor [340]. We identify NCEH1 as differentially expressed comparisons corresponding to: subcloning of Huh-7.5.1 from Huh-7, R1 from Huh-7.5.1, R1.09 from R1 and R2.1 from Huh-7.5.1. NCEH1 follows an antiviral expression profile without deviation and scores -4. Visnins are calcium sensor proteins that modulate multiple intracellular targets [341]. In contrast to NCEH, VSNL1 has a unanimously proviral expression profile of +3, as it is differentially expressed in three comparisons corresponding to: subcloning of R1 from Huh-7.5.1, R1.09 from R1, and R2.1 from Huh-7.5.1. Both of these genes are also differentially expressed in comparisons between infected and uninfected cells; NCEH1 is upregulated in both infected Huh-7.5.1 and Huh-7.5.1c2 cells compared to the uninfected cells, whereas VSNL1 is downregulated in Huh-7.5.1c2 cells following infection. Also among high-scorers are three proviral HDFs: PROX1, GCAT and ATP10D. In agreement with their HDF status, these genes have unanimously proviral expression profiles, each scoring +3. These high-scoring genes that appear in multiple HCV-related data sources may have a significant role in HCV infection.

5.4.5 Investigation of HCV protein neighbourhoods reveals plausible mechanisms for change to infection susceptibility

Investigation of the network neighbourhoods of HCV proteins could identify plausible mechanisms for change in HCV infection susceptibility (Figure 5.7). In order to focus our search we only investigated differentially expressed genes from high-scorers or functional clustering networks corresponding to (i) secreted signal peptides and glycoproteins and (ii) the acute inflammatory response that we find to be significantly pro- and antiviral in their expression. In addition, we only evaluated interactions between these differentially expressed genes and HCV proteins, HDFs, HRFs and HCV-linked cellular receptors and lipoproteins.

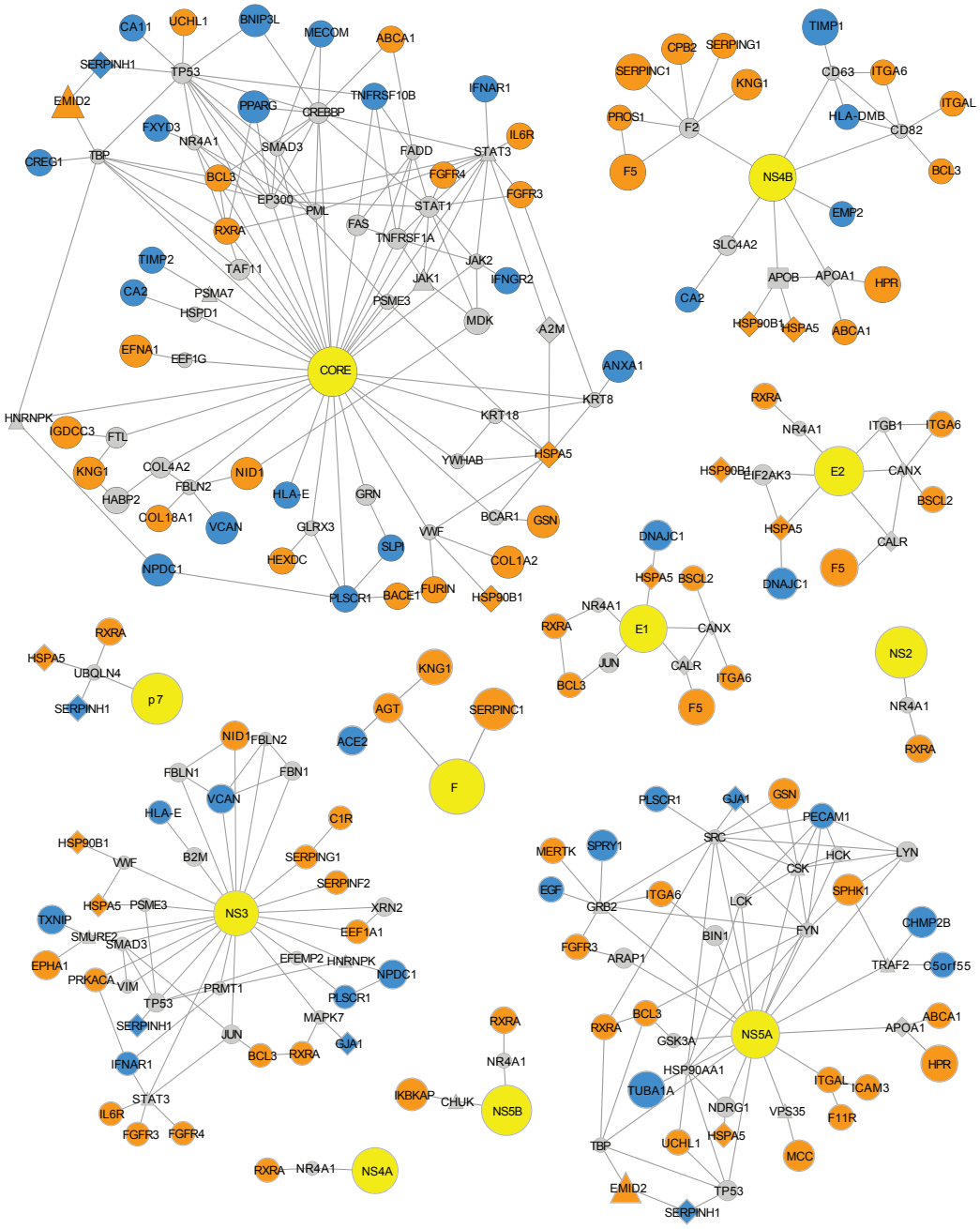

Figure 5.7: Protein interaction neighbourhoods of HCV proteins. HCV proteins are denoted by yellow nodes. Host proteins encoded by genes from either the high-scorer set or from significant antiviral and proviral biological functions are denoted by orange or blue nodes, indicating proviral or antiviral expression profiles, respectively. Other HCV interacting host proteins are denoted by grey nodes. Edges represent interactions between these proteins.

Control over STAT3 protein activation

Stat3 (signal transducer and activator of transcription 3) is a transducer for a variety of signals including response to cytokines and growth factors. Upon activation by phosphorylation, Stat3 proteins dimerize and are translocated to the nucleus where they act as activators of transcription [342]. Stat3 is among the host cell interactants of HCV NS3 and core proteins [126] and is activated by core through a direct interaction, causing proliferation and possibly promoting tumorigenesis [343]. In addition, Stat3 has been shown in two independent siRNA screens to be essential for HCV replication [151, 154]. In contrast, other work has shown that Stat3 activation following interferon or IL6 treatment can prevent HCV subgenomic replicon replication by inducing an antiviral response [344]. Therefore, despite a clear importance, the effect of Stat3 signaling in HCV infected cells has yet to be fully understood.

In both the NS3 and CORE protein networks, we identify three host proteins, encoded by genes IL6R (interleukin 6 receptor), FGFR3 (fibroblast growth factor receptor 3) and IFNAR1 (interferon-α receptor 1), that act as activators of Stat3 activity. IL6R is an activator of Stat3, in response to interleukin 6 [342]. IL6 is included in gene sets taken from functional clustering networks (Figures 5.5B and 5.5D) and has a profile score of +2. Specifically, IL6R is upregulated in Huh-7.5.1 in comparison with Huh-7 (fold change of 2.6) and downregulated in R2.1 in comparison with Huh-7.5.1 (fold change of 6.1). This change in regulation, combined with substantial fold changes purports a proviral action. However, this is not consistent with the findings of Zhu *et al.* [344], who show IL6 mediated signaling to be antiviral. FGFR3 is also included in the gene set taken from the functional clustering network (Figure 5.5B), has a score of +2 and is downregulated in R2.1 (fold change of 1.6) and R1.10 (projected fold change of 1.8) following subcloning. Conversely, IFNAR1 has an negative score of -2, as it is downregulated in both the Huh-7.5.1 and R2.1 cells following subcloning, both with a fold change of 1.6. This activity is consistent with the findings of Zhu *et al.* on the basis that the IFN-induced antiviral activity can be mediated by this receptor.

Modulation of TNF-mediated signals and NF-kB activation

HCV modulates the host innate immune response using multiple strategies [345]. One of these strategies involves regulation of TNF-induced NF-kB, a transcriptional regulator and an important controller of inflammation and immune activation. Several HCV proteins are known to regulate NF-kB including NS5A, NS5B, core and F [346, 347, 348, 64]. NF-kB activation is mediated through engagement of the TNF receptor. Upon stimulation, components of a signaling complex are recruited to the receptor. Signaling complex formation requires adaptor proteins including TRAF2 (TNF receptor associated factor family 2) [349]. NS5A appears to negatively regulate TNF-α-mediated activation of NF-kB through a direct interaction with TRAF2 [346]. However, TRAF2 has been shown by siRNA screen to be necessary for HCV replication [150], therefore it is unlikely that

HCV infection simply requires suppression of TRAF2 activity. The sphingosine kinase 1 (SPHK1) and its product, an anti-apoptotic lipid mediator, sphingosine-1-phosphate, have recently been confirmed as important factors in TRAF2-mediated NF-kB activation [350]. The SPHK1 gene is expressed at a greater level in Huh-7.5.1 cells when compared to Huh-7 and resistant cells and is among the high-scorers with a score of +3, indicating that this gene may play an important proviral role in HCV infection, perhaps via a link to TRAF2 and an involvement in NF-kB induction, prevention of apoptosis and regulation of the innate immune response. We also identify two other genes among the subset of differentially expressed genes that we investigated that interact with TRAF2: CHMP2B (chromatin modifying protein 2B) and putative gene C5orf55. Both C5orf55 and CHMP2B have negative scores of -2 and -3, respectively, indicating a possible antiviral link. This provides additional evidence that TRAF2-related processes may have an effect during HCV infection.

Phospholipid scramblase 1 as an enhancer of interferon signaling

Interferons are important regulators of the innate immune response to viral infection [337]. Indeed, interferon-α is used as a treatment to reduce viral load in HCV infected patients [71]. PLSCR1 (phospholipid scramblase 1) is an interferon-stimulated gene that contributes to the interferon-mediated antiviral response. Though the underlying mechanism for antiviral activity of PLSCR1 remains to be fully understood, evidence indicates that this action is dually mediated at the cell membrane, where PLSCR1 can alter the distribution of phospholipids and in the nucleus, where this protein binds to DNA, possibly to potentiate transcription [351]. PLSCR1 appears in the gene set taken from the functional clustering network (Figure 5.5D) and has a score of -2, as a result of being upregulated in all resistant cells (with fold changes of between 1.6 and 2.2), relative to Huh-7.5.1. In addition, PLSCR1 interacts with HCV core, indicating a mechanism through which HCV may act to control PLSCR1 signaling [126]. Therefore, PLSCR1 is a candidate for increased resistance to infection of R1.09, R1.10 and R2.1 cells.

Cholesterol efflux

A link between cholesterol efflux and HCV infection has been made previously. The scavenger receptor SR-BI mediates cellular uptake of cholesterol and the flux of cholesterol between HDL and the cell [352]. SR-BI is also an important HCV virus entry factor, possibly promoting viral entry through regulation of plasma membrane organisation, being a provider of cholesterol and interaction with other entry factors [301]. Virion-associated cholesterol is also a requirement of HCV infectivity [353]. We have previously mentioned the antiviral expression profile of NCEH1 and it's inclusion among high-scorers and HRFs. Another host protein with a related role is ATP-binding cassette protein (ABCA1). This protein is a cholesterol efflux pump for removal of cellular lipids [354]. ABCA1 is among the gene set gene set taken from the functional cluster-

ing network (Figure 5.5B) and unlike NCEH1 it has a proviral profile score of +2. The ABCA1 encoded protein effluxes cholesterol to apolipoprotein A-I, a major constituent of HDL and these two proteins interact directly. Apolipoprotein A-I is present among HRFs and also interacts directly with HCV NS5A, probably as part of HCV-associated lipid metabolism dysregulation [355]. This evidence suggests that enzymes with the ability to alter balance of cholesterol efflux may also impact HCV infection. Therefore, ABCA1 and NCEH1 may have an effect on susceptibility to HCV infection in hepatoma cells.

Serpins as mediators of HCV NS3 protein activity

HCV NS3 protein is a serine protease that contains a helicase domain and a serine protease domain. NS3 is responsible for cleavage of viral polyproteins and disruption of host innate immune response [356]. NS3 protease action is inhibited by serpin C1 through a direct physical interaction [339] and serpin-mediated inhibition of NS3 has been proposed as a possible anti-HCV therapy [339, 357]. However, we find that SERPINC1, the gene that encodes serpin C1 and a second serpin encoding gene SERPINA6 are among the high scorers with proviral expression profiles that score $+4$ and $+3$ respectively and these instances of differential expression also include substantial fold-changes. Serpin C1 has also been found to interact with the HCV F protein [358], NS3 also interacts with other serpins G1 and F2, and serpins C1 and G1 are both found in the NS4B PPI network neighbourhood (Figure 5.7). The latter serpin genes all have proviral expression patterns. Furthermore we also identified that serpins encoded by genes SERPINH1 and SERPINA1 are part of the HCV replication complex. These results raise the question of whether serpins play an additional proviral role in mediation of NS3 (and possibly F and NS4B) protein activity HCV life cycle, possibly as part of the HCV replication complex.

5.5 Conclusion

In this study we performed multiple genome scale expression studies of Huh-7 derived hepatoma cells with the aim of identifying genes and biological functions that have a significant role in HCV infection. This permitted a detailed account of changes to gene expression caused by HCV infection, determined key differences between commonly used HCVcc cells and implicated novel host factors in determining cellular permissiveness to infection.

Firstly, by comparing uninfected and infected hepatoma cells we identified a set of host cellular functions that are regulated during HCV infection including proteins associated with microtubule organisation, ubl conjugation, zinc-finger domain-containing transcription factors and proteins with helicase activity (supplementary table S5.5). Though proteins involved in microtubule organisation ubiquitination and DEAD-box helicases have previously been identified as differentially expressed following HCV infection of hepatoma cells *in vitro* [113], the sensitivity of our study has highlighted the breadth of

regulation of genes with these functions. In addition, we identify transcriptional upregulation of ubiquitin specific peptidases as a particular mark of HCV infection in Huh-7 derived cells. Furthermore our results indicate that transcriptional upregulation of anti-apoptotic and proliferation stimulating factors may be a cause of increased permissiveness to HCV infection in Huh-7.5.1c2 cells.

Secondly, we examined the expression profiles of six hepatoma cells that have been subcloned from Huh-7, including three cell types that are resistant to HCV infection and genes differentially expressed between subcloned cells and their parent cells were identified. We were able to confirm that cells derived from the R1 subclone have significantly reduced levels of CD81 owing to a mechanism that acts at the level of gene expression. Additionally, we identified 236 host factors that are associated with HCV replication complex in the membranous web of infected cells (HRFs, supplementary table S5.8). This is the largest set of HRFs that has been identified to date. From HRFs we implicate change in expression of APOE, DDOST and PPIA in the resistance to infection of the R2.1 cell. We also identify a subset of HRFs that interact with proteins of HCV, including APOE, CALN that are known to be involved in production of HCV [335, 336]. We scored genes according to their expression profile and used these scores to identify antiviral and proviral candidate genes. Table 5 lists the top scoring genes that include both novel candidate host factors and factors linked to HCV replication, such as tubulin-α [322] and two HRFs, NCEH1 and VSNL1 – NCEH1 is a potentially antiviral factor and VSNL1 is a potentially proviral factor.

Our analysis of HCV infected cells also highlighted the ability of HCV to attenuate interferon upregulation, even in the Huh-7 cell that, unlike Huh-7.5 derived cells, is not reported to have a defective RIG-I signaling pathway. However, when we performed a meta-analysis of gene expression in the six uninfected Huh-7 derived cells, five of which are subcloned from Huh-7.5, we identify that the acute and innate inflammatory responses, as well secreted signal peptides and glycoproteins, are likely to be linked to differences in susceptibility to infection between these subclones. Furthermore we can predict that these mechanisms of susceptibility or resistance to infection are independent of RIG-I signaling. Following these observations network neighbourhoods of HCV proteins were explored and hypotheses for changes to susceptibility to infection were postulated that involve novel HCV-related factors including ABCA1, SPHK1 and CHMP2B, in addition to supporting previously implicated factors such as PLSCR1 [126] and STAT3 [126, 343].

Interestingly, we find that secreted proteins (particularly glycoproteins) are linked to HCV infection: Genes with this annotation are over-represented among differentially expressed genes from multiple parent-child subclone comparisons and we identify these as having more significant proviral and antiviral expression profiles than would be expected than by random chance. Among these secreted proteins are factors involved coagulation, such as complement components, serpins and coagulation factors and these factors have largely proviral expression profiles (see table 5.5 for some examples). Indeed, members

of complement and coagulation pathways have previously been identified as potentially important cellular cofactors of NS4B through yeast two-hybrid and functional network analysis [318]. Also, among HRFs are factors that have a role in folding and secretion of coagulation factors, such as CALR and CANX [359] and CANX is also involved in production of HCV glycoproteins [336]. Hence, changes in HCV infection susceptibility could relate to the ability of the cells to produce viral glycoproteins. For example, HSPA5 encodes a heat-shock protein that is involved in protein folding and assembly in the endoplasmic reticulum [360]. HSPA5 is downregulated in all HCV resistant cell types compared to Huh-7.5.1 with approximate fold-changes between 1.5 and 2, though with highly significant probability (fdr $< 1 \times 10^{-7}$ in each case). Other host factors with chaperone and protein folding activity that achieve high profile scores and appear in HCV protein network neighbourhoods include heat-shock proteins DNAJC1 and HSP90B1 [361, 362]. Another heat-shock protein, Hsp90, has been shown previously to form a complex that includes HCV NS5A and has an important role in HCV RNA replication [363]. Investigation of the ability of these chaperones to influence virus protein production will potentially identify additional mechanisms important to HCV infection.

Overall, our study builds upon current knowledge of infection and our results may contribute to the development of new antiviral treatments to counter the global HCV problem, particularly where we identify potentially proviral proteins that could act as drug targets. Further study, such as genomic sequencing of Huh-7 derived cells would provide greater insight into the extent that these cells mutate in order to effect the extensive differences in gene expression that we observe following subcloning. Changes to susceptibility to infection could then be attributed to gene-specific mutations, in the same way that Huh-7.5 susceptibility has been linked to mutation of RIG-I. Genome sequencing may also highlight other previously undetected mutations in Huh-7.5.1 as well as other Huh-7 derived cells to further our understanding of HCVcc systems and their suitability for modeling infection.

5.6 Supporting material

Supplementary file S5.1. Functional annotation cluster networks from differentially expressed genes and other HCV-related data sources. These networks highlight areas of shared enriched function between gene sets that we identify as differentially expressed between hepatoma cells and also gene sets that relate to HCV infection. Nodes represent annotation clusters from the data source denoted by the node colour. Edges represent shared annotation terms between clusters. Only nodes that share at least $1/4$ of annotating terms are connected by an edge. Node diameter is proportional to the level of enrichment of the biological function in the gene set. Edge width is proportional to the proportion of annotating terms shared between two clusters.

Supplementary file S5.2. (A) Huh-7.5.1, R1.1 and R2.1 G418 resistant colonies

Table 5.5: Genes with top-scoring antiviral and proviral expression profiles. Genes with a profile score of < 3 (antiviral profile) are listed above and > 3 (proviral profile) are listed below. Also given is the change in regulation of the gene in Huh-7, Huh-7.5.1 and Huh-7.5.1c2 cells (in that order) from uninfected versus infected comparisons, where "n" represents no significant differential expression, "+" represents upregulation in the infected cell and "–" represents downregulation in the infected cell.

Gene name	Profile score	infected vs. uninfected
trophoblast glycoprotein	-5	n n n
GalNAc-T7	-4	+ + +
sperm associated antigen 1	-4	+ + +
tubulin, alpha 1a	-4	n + +
neutral cholesterol ester hydrolase 1	-4	n + +
interleukin 17D	-4	n n –
proline-serine-threonine phosphatase interacting protein 2	-4	n + n
discoidin domain receptor tyrosine kinase 1	-4	n n +
lectin, galactoside-binding, soluble, 1	-4	n n +
ependymin related protein 1	-4	n n n
MHC class I polypeptide-related sequence B	-4	n n n
TIMP metallopeptidase inhibitor 1	-4	n n n
complement component 3	5	n – n
peptidoglycan recognition protein 2	4	– – –
potassium channel, subfamily T, member 2	4	– – –
coagulation factor XII (Hageman factor)	4	n – –
annexin A9	4	n – –
orosomucoid 2	4	n – –
complexin 1	4	n – –
hydroxysteroid (11-beta) dehydrogenase 2	4	n – –
transmembrane protein 86B	4	n – –
solute carrier family 7, member 10	4	n – –
haptoglobin	4	n – –
serpin peptidase inhibitor, clade C (antithrombin), member 1	4	n – –
haptoglobin-related protein	4	n – –
left-right determination factor 1	4	n – n
reelin	4	+ n n
argininosuccinate synthetase 1	4	n – n
ATP-binding cassette, sub-family B (MDR/TAP), member 4	4	n n n
KIAA1462	4	n n n
orosomucoid 1	4	n n n
coagulation factor V (proaccelerin, labile factor)	4	n n n

transfected with HCV genotype 1b (Con1) subgenomic replicon encoding a neomycin resistance gene. (B) Huh-7.5.1, R1.1 and R2.1 G418 resistant colonies transfected with Con1 full-length replicon RNA encoding a neomycin resistance gene. (C) Huh-7.5.1, R1.09 and R1.10 G418 resistant colonies transfected with HCV genotype 1b (Con1) subgenomic replicon encoding a neomycin resistance gene. (D) Staining of HCV infected cells for viral E2 protein at 3 (Huh-7.5.1c2), 4 (Huh-7.5.1) and 7 (Huh-7) days post inoculation.(E) Infectivity of supernatant produced from infected cells at 2, 4 and 8 days post infection.

Supplementary file S5.3. PCA analysis was carried out on RMA expression values of each array. Principal components 1, 2 and are plotted for each array.

Supplementary file S5.4. Results of differential expression analysis including Entrez gene ID, cell comparison in which the gene is differentially expressed, log fold change,

p-value and corrected p-value.

Supplementary file S5.5. Output from DAVID 6.7 functional annotation clustering on subsets of genes that are differentially expressed following HCV infection. Only annotation clusters that have an enrichment score > 2 are shown. Sheets A-G correspond to the gene sets that are illustrated in Figure 5.3 in the main text.

Supplementary file S5.6. Differentially expressed genes assigned to branches of the tree of cells. Shown are the Entrez gene IDs, the tree branch to which the differentially expressed gene is ascribed and the cells in which the gene is up- and downregulated.

Supplementary file S5.7. Output from DAVID 6.7 functional annotation clustering. The first six sheets show results for gene sets that are differentially expressed on a specific branch of the tree of cells. The following three sheets show results from genes that are differentially expressed in HCV susceptible cells following infection. The remaining six sheets show results for other HCV-related data sets: HCV-linked cell receptors and lipoproteins, HRFs, HDFs, two sets of HCV-protein interacting factors (from studies [126] and [318], respectively) and genes differentially expressed during chronic HCV infection [319].

Only annotation clusters that have an enrichment score > 2 are shown. Each sheet shows results from a different branch and sheets are named accordingly.

Supplementary file S5.8. List of HCV replication factors (HRFs). Shown is the Enrez gene ID, gene name and protein accession.

Supplementary file S5.9. Expression profile scores for genes that were differentially expressed (DE) and assigned to a tree branch (supplementary table S5.6). Shown are the Entrez gene ID, number of times the gene is found to be differentially expressed, the overall pattern of gene regulation, the score (s) and normalised score (s-norm).

Supplementary file S5.10. The microarray gene set following removal of genes that are not expressed in any cell type (see Materials and Methods for details). Shown are the array annotation ID, Entrex gene ID, gene symbol and gene name.

CHAPTER

SIX

DIFFERENTIAL GENE REGULATION NETWORKS IN PATHOGENIC AND NATURAL HOST SIV INFECTIONS

6.1 Abstract

SIV infection of rhesus macaques (RMs) and HIV-1 infection of humans are similarly pathogenic diseases, both causing immune dysfunction. Conversely, SIV infection of "natural" hosts, such as African green monkeys (AGMs), are effectively non-pathogenic. Studies have shown that innate immune regulation during SIV infection is different between natural host and pathogenic infections. Furthermore, gene-regulation is important in determining the outcome of infection. In this work we perform expression profiling of genes from CD4+ cells from two types of tissue from SIV infected RMs and AGMs, and infer gene-regulatory relationships in order to further explore pathogenesis of SIV infection. We define expression profiles for genes differentially regulated during SIV infection. We show that the profiles include antiviral responses and cell-cycle dysregulation in AGMs and RMs. Using a mutual information measure we identify gene-regulatory relationships and employ a bayesian approach to infer 32 regulatory interactions, some of which are supported by multiple samples. Using a network model incorporating host-virus interactions, we infer 19 cellular genes, including interferon regulator IRF7, that may be significant effectors of SIV-induced gene regulation. Our results highlight several cellular host factors that appear to be be important in regulation of the immune responses during SIV infection of AGMs and RMs. Using gene-regulatory networks, we show that genes IFIT1 and ISG15 could be important for negative feedback during SIV infection and we highlight the importance of IRF7 as an effector of immune activation. Our work complements previous SIV studies by identifying gene regulatory networks related to resistance and provides potential avenues for treatment of HIV-1 infection.

6.2 Background

Type I human immunodeficiency virus (HIV-1) infection causes immune deterioration through the infection and depletion of CD4+ T cells. Paradoxically, disease progression is also associated with a greater level of T cell activation and chronic immune activation in infected patients. Increased immune activation is considered the driving force of CD4 T cell depletion ultimately leading to acquired immunodeficiency syndrome (AIDS), characterised by the inability to mount sufficient immune defense against opportunistic infections [42, 364]. Although deterioration of the host immune response occurs over a long time period and the onset of AIDS takes years if the infection is left untreated [365], important events in innate immune responses and also dysregulation of immune activity occur within the incubation and acute stages of infection, i.e., within the first few weeks after contracting the virus [42].

Simian immunodeficiency virus (SIV) infection of natural host species, such as the sooty mangabey and African green monkey (AGM), is usually not pathogenic [366]. However, SIV infection of "non-natural" host species, such as rhesus macaques (RMs) is pathogenic and leads to immune system dysfunction and a condition similar to human AIDS [367]. Hence, SIV infection of RMs is used as a model system, proximal to HIV-1 infection of humans, that is applied to better understand aspects of pathogenicity and host-virus interaction [367, 366]. Similarly, natural hosts might provide better understanding of AIDS resistance in human long-term non-progressors and elite controllers, allowing the identification of potential targets for interference.

Three recent studies, including one of our own, compared temporal changes in gene expression between the SIV infection of natural hosts (AGMs and SMs) and pathogenic macaque hosts (RMs and Asian pigtailed macaques) [48, 189, 49]. These studies highlight that the innate immune response to infection becomes activated in both species, despite their different disease outcomes. This antiviral defense response includes robust production of type-I interferon (IFN) and strong upregulation of many IFN-stimulated genes (ISGs). However, the pattern of gene expression linked to a type I IFN response is quite different between primate species. Interestingly, in a previous study [189] we identified that induction of ISG expression is at least as rapid and as strong in AGMs as in RMs. However, whilst ISG expression in AGMs is quickly attenuated (by the end of the acute infection period), in RMs, ISG expression is more gradually, progressively upregulated during acute infection. A similar pattern of expression was also observed by Lederer *et al* [48].

Both our study [189] and Lederer *et al.* [48] conclude that a regulatory mechanism that attenuates the innate immune response shortly after its activation, is probably active in SIV infected AGMs but not RMs. Furthermore, it is assumed that this negative control mechanism could play an important role in preventing SIV infection from becoming severely pathogenic by preventing chronic immune activation. The importance of rapid antiviral response to SIV infection, as observed in AGMs, remains unclear. An early and

strong innate immune response has been shown to have a limiting effect on SIV pathogenesis in mucosal tissue [368]. However, a study that compared another natural SIV host, the Sooty Mangabey (SM), with non-pathogenic infection, to RMs, did not find that the innate immune response to infection was activated more quickly in SMs than RMs [49].

Gene-regulatory relationships (and subsequently gene-regultory networks) can be inferred from gene expression profiles using a measure called mutual information (MI) [107, 116]. In this context, MI quantifies the mutual dependence that the expression of two genes have upon one another, thus, a large MI value between two genes indicates a high likelihood for the existence of a gene-regultory relationship. Though gene-regulatory relationships identified using MI are not necessarily direct (they may preferentially identify, indirect "short-cuts" in regulatory pathways), MI methods compare favorably with other gene-regulation inference methods, such as bayesian networks in terms of their recall and accuracy [116]. In this work we retrieve gene expression data from SIV infected AGMs and RMs, from both lymph node (LN) and peripheral blood (PB) CD4+ cells originating from our previous study [189] and use these data to further explore SIV-induced gene regulatory patterns. Initially, we classified differentially expressed genes based on their expression profiles. Following this, we use an MI method [107] to infer gene-regulatory relationships between differentially expressed genes in order to provide additional insight into pathogenic versus non-pathogenic SIV infection. We use gene ontology annotation [173] criteria and employ a bayesian approach to identify interactions with improved confidence and also use a modeling approach that incorporates virus-host interaction data [27] to infer significant cellular effectors of SIV-driven changes to gene expression. Our work contributes to the understanding of pathogenicity and control over immune activation during SIV infection.

6.3 Methods

6.3.1 Gene expression in SIV infected primates

Pre-processed gene expression data, from both PB and lymph-node (LN) CD4+ T cells was obtained from our study of SIV infection among African green monkeys (AGMs) and rhesus macaques (RMs) [189]. Gene expression data included expression levels from individual animals (typically six at each time-point) and also combined data giving mean $\log_2$ expression levels for probes compared with baseline expression and statistical significance for differential expression. Mappings between simian and human genes were also present in the expression data. The raw expression data can be downloaded from the MACE database (http://mace.ihes.fr) using accession numbers 3070984318 (AGM), and 2932572286 (RM).

6.3.2 Clustering of gene expression profiles

Probe expression profiles, consisting of mean $\log_2$ gene expression values for all mutually available time-points post-infection (1, 14, 28 and 65 days) from AGMs and RMs, for both PB and LN CD4+ cells were selected for differentially expressed probes ($P < 0.1$, at one or more time-point). Expression profiles, regardless of their source were pooled and clustered using Mfuzz soft clustering with a "fuzzification" parameter of 1.25 [369] and a stringent within-error (α) of > 0.6, hence the choice of a permissive P value cutoff for differential expression. Probe IDs were assigned to a single cluster that they fit with the largest value for α. A full list of probe ID to expression profile ID mappings are given in supplement S6.6.

6.3.3 Functional enrichment analysis

Functional enrichment analysis of clustered genes was performed using DAVID 6.7 [137, 108], taking the Benjamini and Hochberg [370] corrected P values as a measure of significance. In addition enrichment for interferon stimulated genes (ISGs) was calculated separately for each primate species-cell type combination, for each expression profile. ISGs were retrieved from a database [114]. Clusters were tested for enrichment of ISGs by Fisher's exact test if they contained one or more ISG, using the number of genes expressed for the given species-cell type source as a background population. Obtained P values were adjusted for performing multiple tests [370].

6.3.4 Inference of regulatory interactions using mutual information (MI)

MI values were calculated between probes using MRNet, implemented in the minet software package [107, 119]. Pearsons's correlation was used as the entropy estimator for mutual information calculation. Using this method MI matrices were produced for both AGMs and RMs, for PB and LN T cells, for each gene that was present in expression data from all six samples and differentially expressed ($P < 0.1$) at one or more time point, for the given species-cell combination . In each case, data from the maximum number of available time-points were used from one up to 115 days post-infection. These criteria allowed between 77 and 91 genes to be analysed using 24-42 samples, dependent on the primate species and cell type.

SynTReN [371] was used to estimate the performance of MRNet for the given number of genes and samples, using the default settings and a gene regulatory network from *Saccharomyces cerevisiae* comprising 795 unique gene-gene interactions. Simulated results show that TP/FP ratios for MI thresholds between 0.3 and 0.6 are approximately 0.5 (supplementary file S6.2).

Gene pairs that are involved in regulatory interactions (TP) are enriched for pairs that have high semantic similarity using the term overlap (TO) measure [173] and GO

annotation [84] data and GO annotation[84]. For example, 31 gene pairs have TO > 10 which constitutes a statistically significant enrichment (Fisher's exact test $P = 4.6 \times 10^{-8}$) given that there are 1711 gene pairs that meet this criteria from 131328 possible unique gene pairs from 513 GO annotated genes from the 795 regulatory interactions. Thus, improved (posterior) probabilities of identifying true positive regulatory interactions by selecting gene pairs that have greater semantic similarity than a given cutoff could be calculated using Bayes' theorum:

$$P(A|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1)+P(A_2)P(B|A_2)}$$

Where: $P(A_1)$ is the probability that a regulatory interaction between two genes exists above a given MI threshold. $P(A_2)$; is the probability that a regulatory interaction does not exist above a given MI threshold, i.e., $1 - P(A_1)$. $P(B)$ is the probability that a gene pair has TO of greater than a given threshold. $P(B|A_1)$ is the probability that the TO value is greater than a certain threshold given a correct regulatory interaction. $P(B|A_2)$ is the probability that the TO value is greater than the threshold given given the lack of a regultory interaction. $P(A|B)$ is the probability that a regulatory interaction exists between two genes given that the TO value is greater than a given threshold that we call P_{TP}: the probability that an interaction is a true positive given a given TO cutoff. $P(A_1)$, $P(A_2)$, $P(B)$, $P(B|A_1)$ and $P(B|A_2)$ can be ascertained using counts in the observed data from the yeast model and P_{TP} can be calculated.

Regulatory interactions networks were visualised using Cytoscape [185], incorporating HIV-1 regulatory interaction data from the HIV-1 human protein interaction database [27, 26].

6.3.5 Computation of statistics between expression profile gene sets

An all-against-all comparison of clusters was performed and two statistics were calculated: (i) The number of probe IDs common to two clusters (i.e., from different microarray samples) was identified. Where the intersection was larger than the expected proportion under a null model, a P value was calculated using Fisher's exact test. (ii) A statistic to determine whether the MI values between probes from two clusters are greater than expected by random chance. In this test the mean MI value from all probes of one cluster against all probes from another cluster was recorded. Secondly, a set of MI values, equal in size to the real data, was randomly generated where at least one of the probes of each pair was a member of the non-random set and the mean MI was recorded. Mean MI values for random sets were generated 1000 times and these values were compared to the real mean MI using a one-sample, one sided T-test with the alternative hypothesis that the real mean MI is greater than the randomly generated MIs. In both (i) and (ii), P values were adjusted for performing multiple tests [370].

6.3.6 Identification of regulatory relationships that differ between primate species

Pairs of GO categories were selected along with the corresponding genes from those categories for which MI values were calculated for all four species-cell type combinations. The MI values for all gene pairs generated by comparing gene lists from two GO categories were retrieved. From these MI values, mean MI was calculated for AGM and RMs and the difference in mean MI was determined. This procedure was repeated using randomly shuffled MI values and the probability of obtaining a greater difference in mean MI for any pair of GO functions was obtained by permutation. 100 000 permutations were performed and the obtained p-values were adjusted for performing multiple tests [370].

6.3.7 Detection of significantly sized virally activated cellular regulatory subnetworks

Regulatory networks were obtained from MI data by selecting edges with $MI > 0.3$. Edges were assigned a probability, P_{TP} for being a TP by utilising a Bayesian method that takes GO annotation overlap into account (as described previously). Cellular genes that are dysregulated during HIV-1 infection were selected as seed nodes. Dysregulated genes were obtained from the HIV-1 host protein interaction database [27], selecting only those cellular genes that are regulated, upregulated or downregulated by one or more HIV-1 protein (excluding Vpu that is not conserved in SIV). Subnetworks consisting of connected components of cellular genes containing each seed node were obtained and statistical significance for subnetwork sizes were obtained using a permutation test. In each permutation, a rewired (randomised) network was produced by repeatedly swapping one of the two incident nodes between two randomly selected edges and a mean average of 1000 swaps per edge was performed in each rewire. Subnetworks for each seed node were also obtained from the randomised network. To take into account false positive interactions, edges were removed with a likelihood $1 - P_{TP}$, from both randomised and unperturbed subnetworks. The size of the connected component containing the seed node was calculated for both randomised and original subnetworks. Following 100 permutations, subnetwork sizes for each seed node were compared for unperturbed and perturbed networks by Mann Whitney U test. P values were adjusted for performing multiple tests [370].

6.4 Results

6.4.1 Clustered expression profiles

Gene expression profiles for differentially expressed genes from SIV infected African green monkeys (AGMs) and rhesus macaques (RMs) from both PB and LN CD4+ cells

were clustered using a fuzzy clustering algorithm [369]. Clustering was performed on a gene set comprising 268 AGM-LN, 210 AGM-PB, 484 RM-LN and 338 RM-PB genes that were differentially expressed during acute SIV infection ($P < 0.1$). To best represent a variety of expression profiles we generated a set, limited to 15 to avoid sparseness, of unique clusters (Figure 6.1). The clusters include 241 AGM-LN genes (90%), 193 AGM-PB genes (92%), 464 RM-LN genes (96%) and 323 RM-PB genes (96%). The expression patterns of the remaining genes did not cluster sufficiently closely in order to fit into one of the 15 sets of clusters. However, the clusters represent a large majority of the expression profiles of differentially expressed genes from all primate species and both cell types.

The genes from the primate species and cell-type combinations are not evenly distributed between clusters (χ^2 test, $P << 0.001$, Figure 6.2). Indeed, some expression profiles best identify a certain species (e.g., profiles 1, 7 and 11), others are characteristic of a single source (e.g., profiles 3, 8, 10 and 12) and finally, some represent more general expression trends across all four sources (e.g., profiles 4 and 14).

There are twelve statistically significant intersections ($P < 0.05$) between the expression profile gene sets (Figure 6.3A). Six of these significant intersections are between gene sets of different sources that are present in the same expression profile cluster, i.e., the equivalent genes undergo a similar profile of expression in multiple cell types and primate species (these relationships are denoted by self-edges). The remaining six significant intersections occur between different expression profiles. Interestingly, from these six remaining intersections, three occur between gene sets of different primate species but the same cell type, i.e., the intersection represents a gene set that has a consistent expression profile within a given primate species but a varied expression profile between the two primate species. These three gene sets could be involved in key mechanisms through which RMs and AGMs differ in their response to SIV infection.

6.4.2 Functional enrichment of clustered genes

A total of eight expression profile clusters comprise genes that are enriched for one or more functional annotation term with a significance of $P < 0.05$ (profiles 1, 2, 6, 9, 11, 12, 13 and 14). Table 6.1 gives the most significant annotation in each case and supplement S6.1 gives details of all significantly enriched terms. We find the most enriched term from all expression profiles encompasses genes from two or more sources (Table 6.1). In addition, all of these enriched terms are attributed to genes from different cell type sources and the most enriched annotations from expression profiles 1 and 9 include genes from both primate species. Therefore, in the majority of cases we show that expression profiles cluster functionally related genes from multiple sources.

Two major functional themes are identified by the functional annotation enrichment analysis: (i) response to viral infection and immune system functions (profiles 2, 11, 12 and 13); and (ii) control of cell cycle progression (profiles 1, 6 and 14). Furthermore, profile 9, which is enriched for genes prominent in hematopoietic cell development also

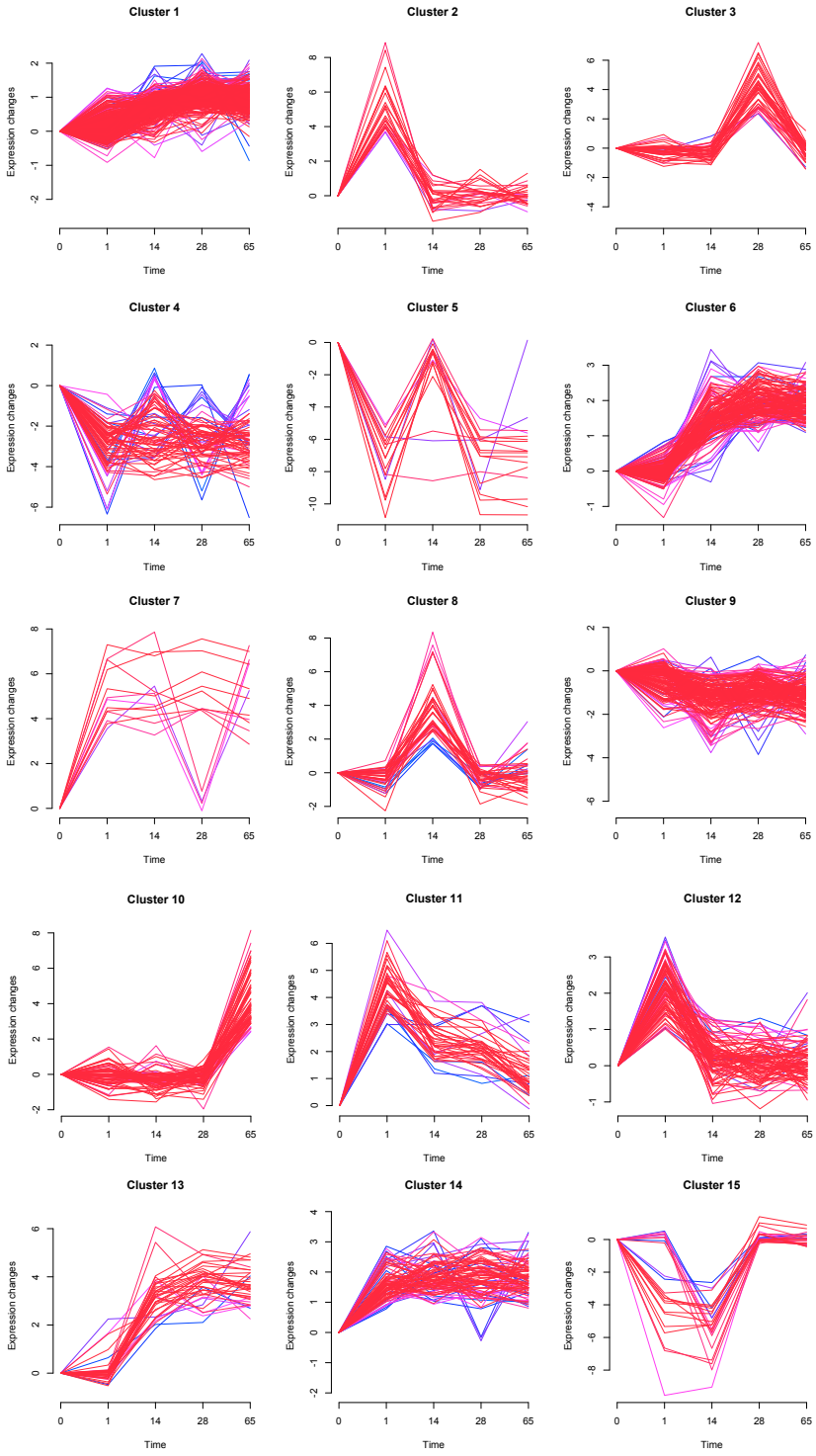

Figure 6.1: Clustered expression profiles. Each plot represents a set of microarray probes clustered according to their expression profile following SIV infection. Each line represents the expression of a gene from either SIV infected AGMs or RMs from either blood PBMC (PB) or lymph node (LN) CD4+ cells. Line colour indicates the goodness of fit for genes to the given expression profile where red is the best fit, blue the least good fit. Intermediate colours represent an intermediate fit.

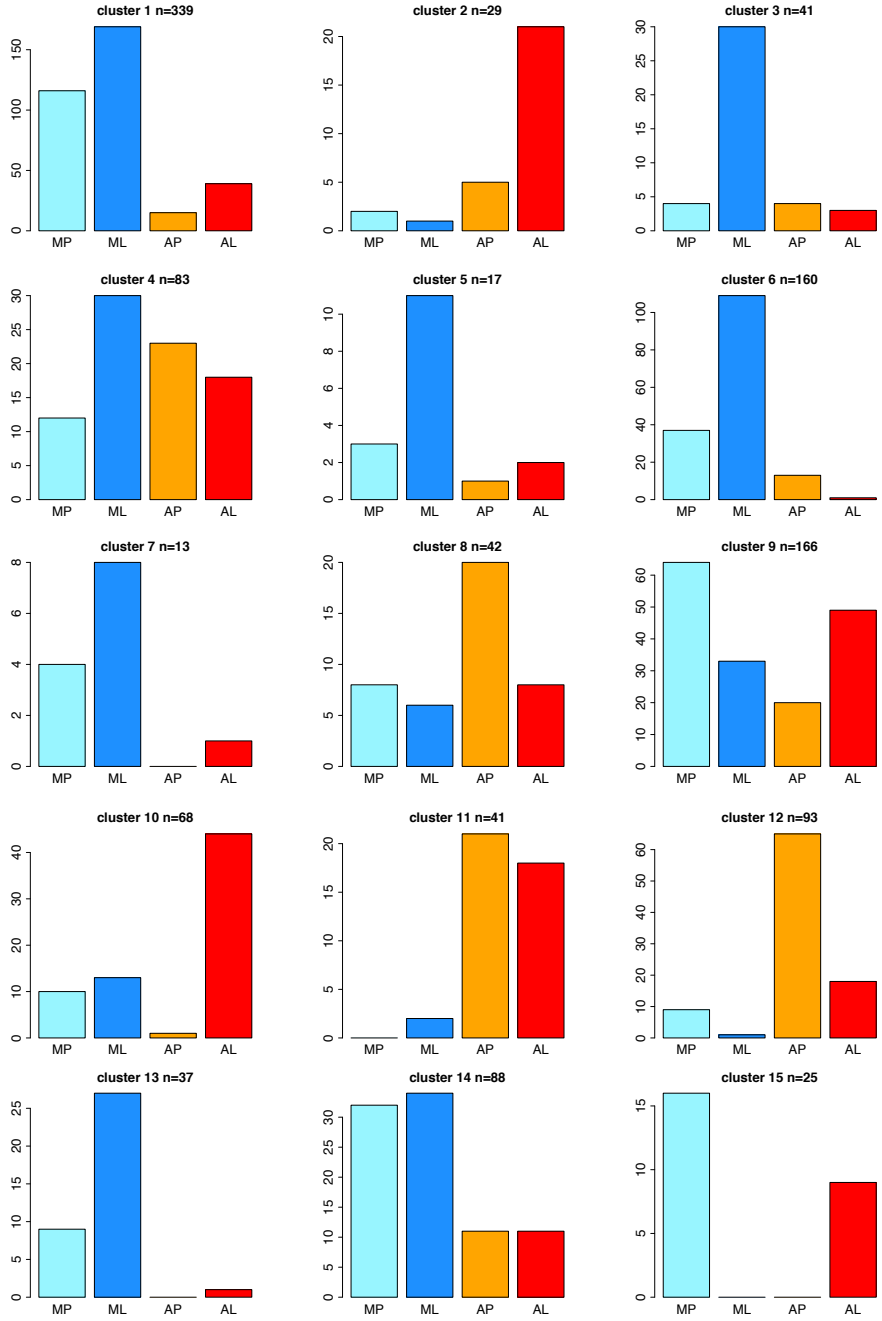

Figure 6.2: Composition of clustered expression profiles. Each plot shows the distribution of gene sources (primate species and cell type) from a cluster, where AL = AGM-LN, AP = AGM-PB, ML = RM-LN, MP = RM-PB. The y-axis values denote the number of clustered genes from each source that are present in a cluster.

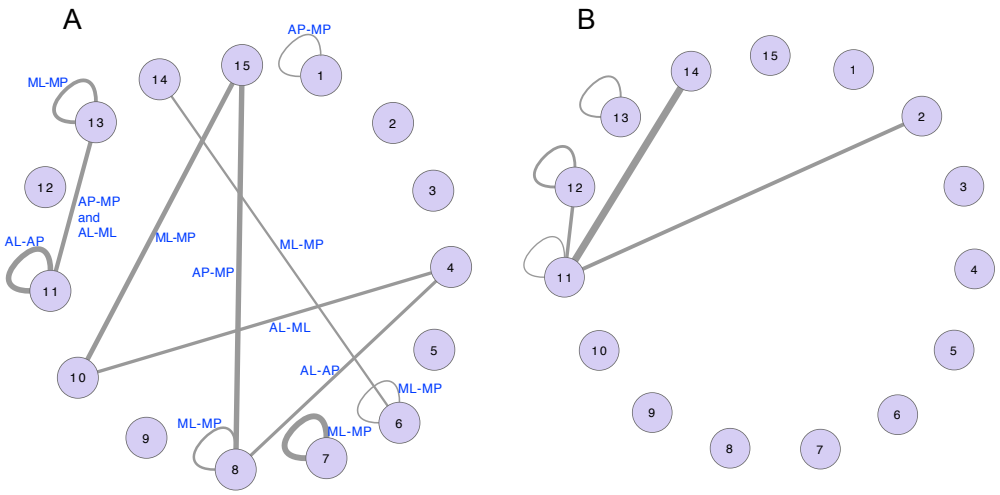

Figure 6.3: Relationships between clustered expression profiles. Nodes represent clusters 1-15 (as numbered) and edges represent relationships between clusters. (A) Clusters that have statistically significant intersection in their gene sets (Fisher's exact tests, where corrected $P < 0.05$) stemming from different sources, where AL = AGM-LN, AP = AGM-PB, ML = RM-LN, MP = RM-PB, are linked. The first source in each label refers to the genes from the cluster with the lowest numerical cluster ID, and the second source refers to the cluster with the highest numerical cluster ID, i.e., the edge joining clusters 4 and 10 indicates that the genes from cluster 4 from the AGM-LN source have significant intersection with those genes from cluster 10 from the RM-LN source. In addition, edge-width is proportional to the Matthews Correlation Coefficient calculated from the gene set intersections. (B) Similarly, clusters that have larger MI values than expected by random chance between the genes of connected clusters are depicted.

Table 6.1: Most significantly enriched annotations from clustered genes

Profile	Annotation	Fold change	Annotated	Corrected P	Sources
1	cell cycle process	3.7	11.7%	5.0×10^{-8}	AL, ML, MP
2	cytokine activity	19.6	21.7%	6.2×10^{-3}	AL, AP
6	M phase	11.0	20.1%	1.1×10^{-17}	ML, MP
9	hematopoietic cell lineage	9.3	5.3%	1.3×10^{-3}	AL, AP, ML, MP
11	response to virus	39.2	22.2%	3.6×10^{-5}	AL, AP
12	response to virus	20.1	12.5%	1.4×10^{-7}	AL, AP
13	response to virus	25.9	17.2%	7.6×10^{-3}	ML, MP
14	cell division	14.0	18.8%	5.7×10^{-11}	ML, MP

includes genes encoding cytoskeleton components, immune activation, cell death and response to wounding, hence bridging both of the former categories.

We identified several expression profiles that do not appear to form cohesive functional groupings of genes. Most of these profiles are quite "simple", involving up- or down-regulation of the genes at a single time-point, i.e., profiles 3, 8, 10, and 15. "Simplicity" referring here to the low number of observed regulatory events required to generate such an expression pattern over time. Such categories are thought to contain a higher number of biologically non-relevant genes, as noisy expression or expression-detection

is unlikely to generate more complex patterns. Therefore, identification of biologically relevant genes within these profiles is more challenging than in those where the profile of expression is statistically significantly different from baseline at more than one measured time-point.

Expression profile 6, that illustrates a fairly consistent and steep increase in expression, comprises 28 genes involved in M phase of the cell cycle but also three genes linked both to control of cell-cycle and apoptosis (BIRC5, BUB1 and CDC2). None of these cell-cycle linked genes from AGMs are present in expression profile 6. A similar set of observations can be made for expression profile 14, as it involves 16 cell-division and M-phase linked host genes, all of which are only differentially expressed in RMs and all undergo upregulation throughout acute SIV infection. Only expression cluster 1 encompasses cell-cycle linked AGM and RM genes (14 and 17 genes, respectively). Indeed these genes are upregulated during infection but not to the same extent as those genes in expression profile 6 or 14. Therefore, it appears that dysregulation of transition to M phase of the cell-cycle is disrupted to a far greater extent in RM than AGM infection.

As a secondary and more specific investigation, we tested whether the genes in each cluster were enriched for interferon stimulated genes (ISGs) [114], as the response of ISGs is of particular importance during acute SIV infection [189]. This test was performed separately for all combinations. We found that eight expression profiles (1, 2, 6, 9, 11, 12, 13 and 14), were statistically significantly enriched for ISGs ($P < 0.05$) and all of these except for profiles 1, 2 and 9 are enriched for ISGs for each individual source. Therefore, all expression profiles for which we previously detected functional enrichment (Table 6.1) are also enriched for ISGs, including profiles 1, 6 and 14 that are enriched for genes of the cell cycle. Note that indeed, some ISGs are involved in the control and progression of the cell cycle [114]. From 23 ISGs present in expression profile 1, sixteen in profile 6 and eleven in profile 14, there are four genes (BARD1, CENPE, PSMB10 and RGS2), two genes (CENPE and MCM4) and zero genes, respectively, that are annotated with the GO term "cell-cycle". All three of these ISG subsets are enriched for proteins involved in antiviral defense (Fisher's exact test, $P < 0.001$). Therefore, the expression patterns of cell-cycle regulators and interferon-stimulated virus response genes during SIV infection are not mutually exclusive. Furthermore, genes that are both ISGs and cell-cycle mediators could play an important role in effecting interferon-stimulated control over cell-cycle activity during SIV infection.

Previously, we postulated that strength, rather than rapidity of innate immune response is a determining factor in SIV pathogenesis [189], as induction of innate immune responses in SMs is not observed to be earlier than in RMs [49]. However, on further inspection, we believe that this assessment should be treated with some caution, this is because the maximum fold-change values post-infection for significantly differentially expressed ISGs are not significantly greater in AGMs than in RMs ($P = 0.1626$, Mann-Whitney U test). However, the time-points post-infection at which ISGs achieve their

maximum fold-change are significantly greater in RMs than AGMs ($P = 1.421 \times 10^{-8}$, Mann-Whitney U test).

6.4.3 Inferring regulatory relationships using mutual information

Mutual information (MI) between expression profiles can be used to infer gene-regulatory relationships [119]. There are five instances where the MI values between the gene sets of expression profiles are statistically greater than expected by random chance (represented by edges in Figure 6.3B). Therefore, it is probable that gene regulatory relationships exist between the expression profiles linked by these edges. Three of these links are internal to the expression profile, denoted in Figure 6.3B by self-edges. The remaining two links are between independent profiles.

True positive (TP) and false positive (FP) rates, among regulatory interactions inferred from MI data, given the array sample sizes, were estimated for the nearest available characterised model, *Saccharomyces cerevisiae*, available for a gene-regulation simulation tool [371]. Simulated results in *S.cerevisiae* show that TP:FP ratios for MI thresholds between 0.3 and 0.6 are approximately $1 : 2$, i.e., the probability of a selected interaction being a TP is approximately $1/3$ (supplement S6.2). However, gene pairs that are involved in a *bona fide* regulatory interaction are enriched for gene pairs that are semantically similar, estimated using the term overlap (TO) method [173]. Application of Bayes' theorum indicates that when a gene pair with a given TO value is selected, interactions of greater confidence can be obtained (Table 6.2). For example, by selecting potential interactions between genes that have a TO of 10 or more, the probability of selecting a *bona fide* gene-regulatory interaction from a set with prior probability of $1/3$ is 0.61. Though the precise probability of these events are not of great interest, as, owing to their origin in a yeast model they can only roughly approximate equivalent values for the simian host-virus system, it is clear that regulatory interactions of greater confidence are likely to be identified using additional functional annotation criteria.

MI networks, using a minimum MI value of 0.3, were visualised for each cell type, primate species, and all combinations thereof (supplement S6.3). By selecting only gene pairs with biological process GO term overlap of > 10, 32 regulatory interactions are inferred (Figure 6.5). Three host-host gene pairs are present in more than one species-cell type source (IFIT1-ISG15, AURKB-UBE2C and UBE2C-BIRC5). Interestingly, several of these genes are present in different expression profiles, dependent on the data source. IFIT1 is a negative regulator of the innate immune response to viral infection and it can downmodulate interferon stimulated genes (ISGs), including ISG15 [372]. These interactions, detected in more than one data set, are likely to represent *bona fide* gene regulatory events during SIV infection. Furthermore, these interactions have a specific expression profile dependent on the cell type and species.

We identify a number of potential regulatory relationships between cell-cycle and apoptosis linked genes and these are particularly prominent in the sets we identify for

Table 6.2: Enrichment of GO term overlap and true positive bayesian probability estimates for regulatory interactions with $MI > 0.3$

TO	Proportion ($MI > 0.3$)	Proportion (all)	Fold enrichment	P (enrichment)	P (TP interaction)
1	0.98	0.98	1.02	0.7	0.33
2	0.54	0.47	1.15	8.42×10^{-5}	0.37
3	0.38	0.28	1.35	2.32×10^{-9}	0.40
4	0.27	0.18	1.56	2.9×10^{-3}	0.44
5	0.18	0.11	1.77	1.17×10^{-12}	0.47
6	0.13	0.07	1.99	9.67×10^{-14}	0.50
7	0.11	0.05	2.35	2.28×10^{-7}	0.54
8	0.07	0.03	2.18	2.27×10^{-7}	0.52
9	0.05	0.02	2.57	2.05×10^{-8}	0.56
10	0.04	0.01	3.12	4.64×10^{-7}	0.61
11	0.03	< 0.01	3.45	2.60×10^{-4}	0.64
12	0.02	< 0.01	2.98	2.15×10^{-5}	0.60
13	0.02	< 0.01	3.90	4.05×10^{-5}	0.67
14	0.02	< 0.01	4.99	7.37×10^{-6}	0.72
15	0.01	< 0.01	5.72	1.32×10^{-5}	0.75
16	0.01	< 0.01	7.01	7.09×10^{-6}	0.78
17	0.01	< 0.01	9.22	7.48×10^{-7}	0.83
18	0.01	< 0.01	10.12	1.45×10^{-6}	0.84

RMs where more than half of the human genes involved come from cell cycle enriched expression profiles (1, 6 and 14, figure 6.5).

From the small number of inferred high confidence regulatory interactions, it is not possible to report on a genome-wide scale upon the differences among specific gene regulatory events between primate species. However, by pooling potential gene-gene interactions according to their function we can observe more general patterns. By comparing MI values from gene pairs from GO categories that are child terms of the biological process GO terms “cell cycle” and “immune response” (corresponding to the two major functional themes of differentially expressed genes), we identify fifteen GO term pairs for which the MI values are statistically significantly different between RMs and AGMs ($P < 0.05$). All GO terms among these fifteen pairs pertain to immune response. Of the fifteen significant term pairs, one or both of the terms in the pair refer specifically to type I interferon activity (supplement S6.4). This result indicates that the regulatory relationships between genes involved in responding to type I interferon are significantly different between AGMs and RMs during SIV infection, confirming previous observations [189, 48, 49].

6.4.4 Virus protein interactions among differentially expressed genes

To investigate specific effects of individual viral proteins on host gene expression regulation, we included viral protein regulatory interactions described as “upregulates", “downregulates" or “regulates", as obtained from a HIV-1 interaction dataset [27, 26], in the gene regulatory networks (Figure 6.5). Notably, the majority of regulatory interactions with these host proteins involve the Tat protein. Indeed, we find that seven expression profiles are significantly enriched for genes that take part in specific viral protein-protein

Table 6.3: Enrichment for virus protein interactions among expression profiles

Profile	Virus interaction	No. of genes	Fold enrichment	Corrected P
13	Tat upregulates	10	7.0	2.2×10^{-6}
11	Tat upregulates	7	7.3	9.9×10^{-5}
6	Tat upregulates	12	3.2	2.9×10^{-3}
9	Tat upregulates	9	4.0	2.9×10^{-3}
12	Tat upregulates	8	3.4	0.022
1	Gag binds	3	10.0	0.040
8	Nef upregulates	2	23.5	0.044
6	Vpr inactivates	2	18.5	0.044
12	Matrix is stimulated by	2	17.9	0.044
12	Vpr competes with	2	17.9	0.044

interactions (PPIs) ($P < 0.05$) and the most significantly enriched are upregulation by Tat, in profiles 6, 9, 11 and 13 (Table 6.3), four clustered expression profiles that are represented in the regulatory interactions in Figure 6.5, two of which are significantly enriched for genes involved in response to viral infection (table 6.1). This result indicates that regulation of immune activation, relevant to SIV infection in RM and AGMs, might probably be mediated by mechanisms responsive to the viral Tat protein.

6.4.5 Inferring cellular effectors of SIV-induced changes to host gene expression

It appears that some cellular genes which undergo a regulatory interaction with a virus protein may be responsible for effecting infection-induced change in gene expression over a wider subnetwork of cellular factors. For example, PPP2CB, the catalytic subunit of protein phosphatase 2A, is upregulated by HIV-1 Vpr in an interaction that is believed to be part of the process of Vpr-induced G2 arrest [373]. PPP2CB is the only host gene known to be dysregulated by HIV-1 within a subnetwork of seventeen host factors from the RM-PB regulatory network (see Figure 6.4 and supplementary Figure S6.3 for further examples). In order to identify those host factors and virus-host interactions that might be playing an important role in dysregulation of cellular gene expression, we calculated the statistical significance of the HIV-interacting cellular gene subnetwork sizes, for both RMs and AGMs over both LN and PB CD4+ cell types. Our method takes the estimated likelihood of cellular regulatory interactions (Table 6.2) into account. We find significant cellular genes ($P < 0.05$, by permutation test, including correction for multiple tests) and corresponding subnetworks for each cell source. Nine of these genes are statistically significant in more than one source, five which are statistically significant across both primate species and eight across both types of cell types (Table 6.4 and supplement S6.5 for greater detail).

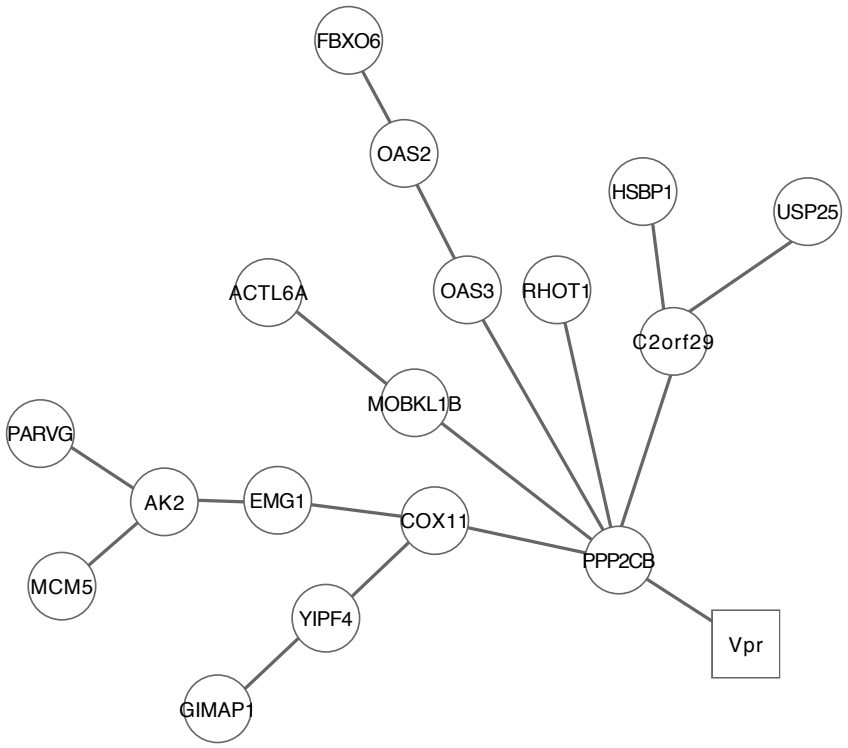

Figure 6.4: Upregulation of PPP2CB by viral protein Vpr potentially effects a changes to host cell gene expression. Genes are represented by nodes. Round nodes represent cellular genes and the square node represents the viral Vpr gene. Potential gene-regulatory interactions are represented by edges.

IRF7 is a virally induced transcription factor that acts as a "master regulator" of type-I interferon-induced responses, as it is essential for ISG induction and is essential for both virus-activated (MyD88 independent) and Toll-like receptor (TLR) activated (MyD88 dependent) pathways [374]. In addition, we infer that IRF7 is a significant effector of virally induced change in gene regulation in AGM-LNs and also RMs during SIV infection. In addition, IRF7 from both RM cell types forms is part of expression profile 6, hence, IRF7 is likely to have a role in driving the progressive continual upregulation of virus response genes in RMs in acute and chronic infection. We find that IRF7 expression at day one post infection is significantly greater in AGMs than RMs ($P = 0.00014$ by T test, Figure 6.6). Therefore we suggest that IRF7 could be part of a positive feedback loop that enhances the initial interferon response in AGMs. However, as IRF7 was only differentially expressed in four from six AGMs at day one, the relevance of this finding, in contributing to the lack of AIDS onset is unclear. It is possible that IRF7 upregulation in the remaining AGMs did occur between samplings but was, thus, not observed .

6.5 Discussion

In this work we have defined fifteen expression profiles for genes that are differentially expressed during SIV infection of RMs (pathogenic) and AGMs (non-pathogenic).

Table 6.4: Virally regulated cellular genes that are part of a significantly sized and robust subnetwork of regulatory interactions.

Cellular gene	Virus protein	ML	MP	AL	AP
ISG15	Tat	•	•	•	•
ISG20	ISG20			•	
IRF7	Tat	•	•	•	
IFI27	Tat	•	•		
IFIT3	Tat	•	•		
IL1B	Tat, Nef, gp160, gp120, gp41		•		
IL8	Tat, Nef, gp120, Vpr		•		
CXCR4	Tat, Nef, gp120		•		•
CXCL10	Tat, Nef, gp120	•			
CCL3	Tat, Nef, gp120, gp41, Vpr, Matrix		•		
MX1	Tat	•	•		•
BIRC5	Vpr	•	•		
CDC2	gp120	•			
CDC20	Tat	•			
HSPA1A	gp120			•	•
HSP90AA1	Tat				•
STAT1	Tat	•	•		•
PSMB10	Tat		•		
PPP2CB	Vpr		•		

Several of these profiles capture the expression of specific biological sub-functions, notably genes of the immune response to viral infection, such as type I ISGs, and cell cycle control genes.

We observe patterns of gene expression pertaining to virus response genes in AGMs and RMs that were previously identified for both this data set [189] and also in an independent study of pathogenic versus non pathogenic SIV infection in SMs [48]: Expression profiles 11 and 12 appear to capture the innate virus defense response in AGMs. Indeed these two profiles are similar, differing only in their relative fold-changes, with a large initial upregulation at day one post-infection, followed by rapid downmodulation during the remaining acute infection period. From genes that comprise expression profiles 11 and 12, only two and ten genes, respectively, are those differentially expressed in RMs. Furthermore, none of these genes from AGMs are present in expression profiles 11 or 12, rather only two of these genes, KIF2B and TTC39B, are differentially expressed in AGMs, neither of which have any clear immune-response related activity. The strong and rapid antiviral response that can be observed in AGMs, is absent at these early time points in RMs. Conversely, expression profile 13, that also includes a large proportion of virus response genes and shares a statistically significant proportion of the same genes with pro-

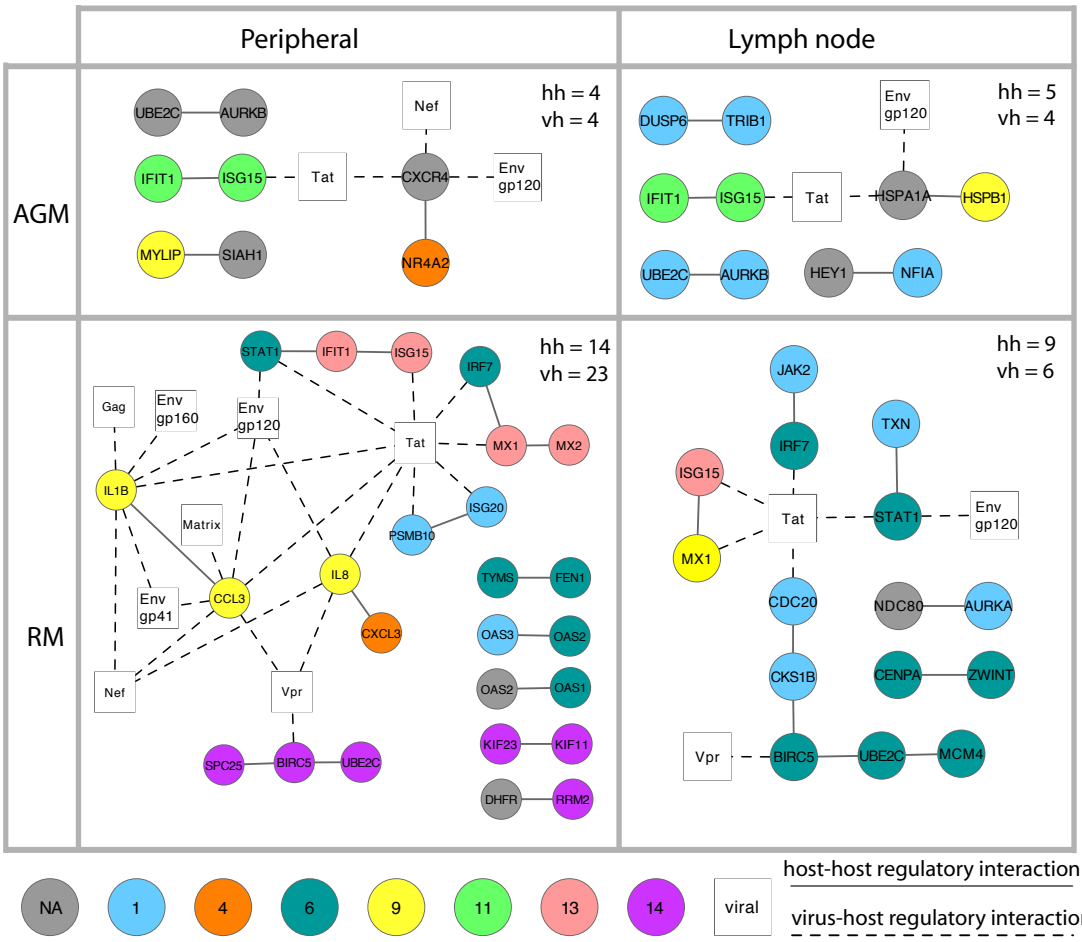

Figure 6.5: Gene regulatory networks, detected for each type of CD4+ cell. Genes are represented by nodes. Round nodes represent cellular genes and square nodes viral genes. Cellular genes are colour coded according to the expression profile from which they are derived, as given in the key at the bottom of figure. Solid edges represent gene regulatory relationships between host cellular genes, obtained using a mutual information method. Broken lines represent established host-virus gene regulatory interactions. Interaction counts for virus-host (vh) and host-host (hh) interactions are displayed for each network.

file 11 for RM infection (Figure 6.3), undergoes upregulation > 1 days post-infection and these genes continue to be upregulated in RMs throughout acute infection and through transition to chronic infection.

We previously found that low levels of interferon-α can induce strong ISG upregulation in AGM cells leading to amplification of interferon-α production, and thus proposed that a positive feedback loop is present [189]. Indeed, IFN control during SIV infection appears to differ between AGMs and RMs, and leads to contrasting expression patterns for initial induction of innate immune response genes [189, 48]. However, it remains unclear whether rapidity of innate immune response is likely to be a determining factor in SIV pathogenesis. Indeed, it also remains to be clarified whether differences between antiviral response kinetics between AGMs and RMs is due to a differences in immune activation independent of the SIV serotypes used in the initial study.

Assuming aspects of initial triggering (including rapidity and strength) of ISG in-

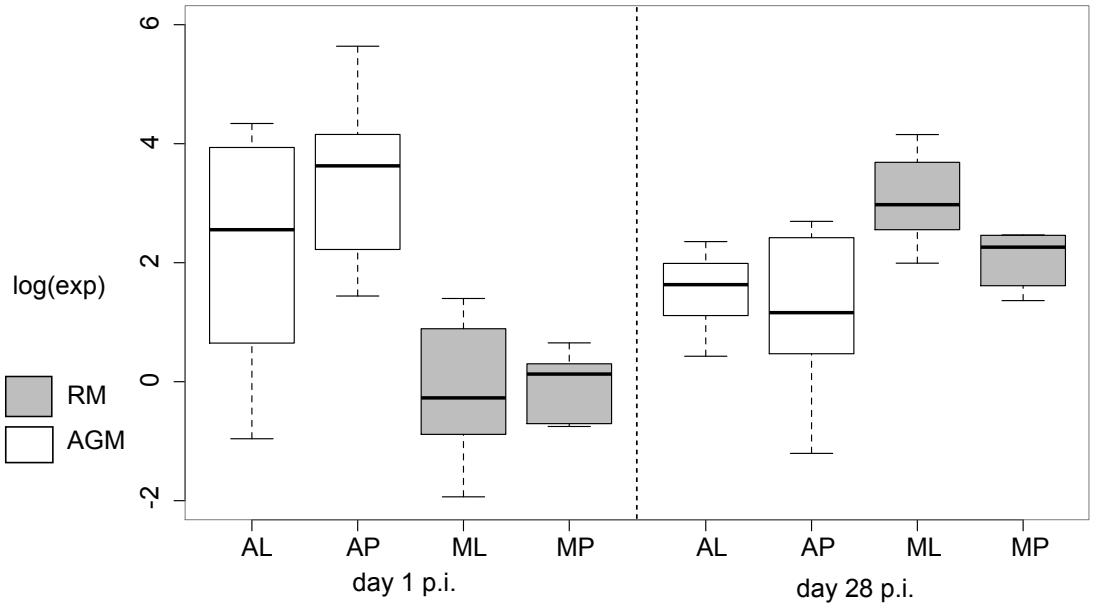

Figure 6.6: IRF7 gene expression in AGMs and RMs at days one and 28 post-infection (p.i.). Log expression values, relative to baseline expression, for all RM and AGM animals in both LN and PB CD4+ cells is shown. AL = AGM-LN, AP = AGM-PB, ML = RM-LN, MP = RM-PB. At day one, IRF7 gene expression is clearly upregulated in AGMs but remains at baseline level in RMs, in both cell types. At day 28, IRF7 expression is greater in RMs, particularly in the LN cell type.

duction is important to pathogenicity of infection, we propose that two major questions relating to type I ISG induction in pathogenic versus non-pathogenic SIV infection remain to be answered. Firstly, what is the mechanism of activation and positive feedback that is linked to ISG expression and why does it differ between the progressor and the non-progressor species? Secondly, what is the mechanism of attenuation of the immune response that is present in non-pathogenic but not pathogenic SIV infection? By combining expression profile information with inferred gene regulatory relationships and virus protein-induced cellular gene regulation data we are able to capture regulatory events that are potentially important for disease progression during acute SIV infection.

The initial activation of the innate immune response by the virus is likely to involve viral proteins and their expression in host cells. The viral Tat gene has been suggested to play a central role in perturbing the host immune response, particularly through dysregulation and activation of cytokines including type I interferons and regulators of inflammation, such as interleukins and TNF-α [31]. Furthermore, expression of HIV-1 Tat in dendritic cells can induce expression of many interferon-inducible genes [115]. However, other viral proteins are also implicated in immune dysregulation including envelope gycoproteins, Nef and Vpr (e.g., [375, 376, 377]). The expression profiles from both RMs and AGMs that comprise immune-related genes are enriched for cellular factors whose expression is altered by HIV-1 Tat. Hence, it appears that immune dysregulation in both pathogenic and non-pathogenic infections involve Tat. Using MI networks, we infer cel-

lular effectors of virally induced changes to cellular gene regulation (Table 6.4). Indeed, fifteen from nineteen of these effectors are regulated by Tat and nine are regulated by at least one other HIV-1 protein. All but six of these fifteen genes (HSP90AA1, CDC20 and CXCR4) are ISGs [114]. Therefore, it is probable that these genes may be part of a greater interferon stimulated network of regulatory interactions.

Interferon-regulatory factor 7 (IRF7) is perhaps the best documented positive regulator of IFN-α induction [378, 374, 379]. Our results highlight that IRF7 potentially has a role in driving the initial interferon response in AGMs and also the progressive continual upregulation of virus response genes in RMs, possibly via a positive feedback loop. However, what remains unclear is why the feedback loop that includes IRF7 and interferon-α, has a different response in RMs compared to AGMs. Multiple IRF7-dependent sensory mechanisms are active during HIV-1 infection including TLR7 activation and cytosolic sensors [374, 380]. Therefore, the role that these different sensory mechanisms have and also the impact that viral proteins (including Tat and also Vpr that is known to stimulate TLR4 [381]) have upon both initial and progressive immune activation during pathogenic and non-pathogenic SIV infection is of great interest and should be subject to further study.

Our results highlight that significant differences exist between gene regulatory relationships of type I interferon related genes between RMs and AGMs, suggesting that type I interferon mediated responses could be important for determining the outcome of SIV infection. We identify a significant relationship between IFIT1 and ISG15 in all but RM LN cells. ISG15 is a type I interferon induced ubiquitin-like protein that is conjugated to other proteins in a process known as ISGylation as a mechanism in cellular antiviral response. ISGylation can lead to gain or loss of function among target substrates that are known to include antiviral ISGs such as RIG-I [305], proteins that have a role in HIV-1 release from the cell via endosomal trafficking [382] and also viral proteins such as influenza NS1 [383, 384]. Indeed, IFIT1 is among the targets of ISG15, although the precise nature of IFIT1 ISGylation remains to be determined [383]. IFIT1 (otherwise known as ISG56), is expressed during virus infection and in response to type I interferon [385]. IFIT1 has been shown to regulate cellular antiviral responses through a negative-feedback mechanism involving inhibition of type I IFN [372] and is therefore potentially important in mediating SIV pathogenesis. Indeed, virally induced expression of ISG15 is enhanced when IFIT1 is knocked down via RNA interference [372]. Although initial expression profiling does suggest that IFIT1 is upregulated in concert with other viral response genes in RMs (being of expression cluster 13), expression of IFIT1 in LN CD4+ cells, unlike the other cell types studied, is not statistically significantly upregulated ($P > 0.1$) during acute SIV infection in RMs, e.g., 0-28 days post-infection, indicating that the levels of IFIT1 are inconsistent between samples. A lack of significant induction of IFIT1 could curtail immune modulation in these cells and contribute to SIV pathogenicity. Therefore, we suggest that the roles of ISG15 and IFIT1 during SIV infection and in particular any

regulatory relationship that exists between these two genes should be further investigated.

Both SIV and HIV exert an intricate system of control over the cell-cycle. The Vpr protein of both HIV-1 and SIV induces cell cycle arrest in the G2 phase of mitosis. It is believed that Vpr-induced cell-cycle arrest results in increased viral replication by upregulating transcription and translation of viral genomic material. Vpr also causes apoptosis through induction of the intrinsic pathway, an effect that is causally linked to cell-G2 cycle arrest and T cell loss [40]. Vpr carries out these activities through a sizable set of direct physical and regulatory interactions with host proteins (179 host protein interaction partners are catalogued in the HIV-1 human protein interaction database [27, 26]). Furthermore, as with immune dysregulation, perturbation of the cell-cycle and host cell apoptosis by HIV-1 is not limited to the activity of a single viral protein as other HIV-1 proteins (additional to Vpr) are also implicated, including Tat [386, 387] and envelope glycoproteins [388]. Therefore, the presence of expression profiles that illustrate changes in regulation to cell-cycle related host genes in our results was not unexpected.

Our results indicate that dysregulation of transition to M phase of the cell-cycle is disrupted to a far greater extent in RMs than AGMs during SIV infection. In addition, we find that cell-cycle and apoptosis linked genes are particularly prominent among the gene regulatory relationships that we detect for RMs (Figure 6.5). BIRC5 encodes Survivin, a negative regulator of apoptosis [389]. BIRC5 expression is upregulated by HIV-1 Vpr, possibly in order to increase the viability of HIV-1 infected cells [390]. UBE2C encodes a ubiquitin-conjugating enzyme that is also known to have anti-apoptotic characteristics [391]. AURKB encodes one of three known members of the aurora kinase family, that play a role in mitosis, the over-expression of which can induce defects in spindle formation and lead to apoptosis [392]. UBE2C, BIRC5 and AURKB are all regulated through the cell cycle and the existence of regulatory relationships between these genes during SIV infection is highly plausible and could be linked to the Vpr activity.

Interestingly, despite our inference of a greater number of intra cell-cycle regulatory relationships in RMs than AGMs, we do not detect significant differences between their gene regulatory relationships. This result suggests that in general, gene regulatory relationships between cell-cycle linked genes act in the same manner in both AGMs and RMs during SIV infection, though in the latter, cell-cycle regulation is more active. Therefore, we propose that while cell-cycle components are indicative and probably also causative of pathogenicity through promotion of increased virus replication and apoptosis, these genes do not explain the mechanism through which AGMs and RMs differ in their response to SIV infection.

6.6 Conclusion

Our study confirms that viral response genes undergo strong upregulation in AGMs SIV infection and a more progressive continuous upregulation in RM infection. Our re-

sults indicate that significant differences in cellular gene regulatory relationships exists between AGM and RMs and we surmise that these differences cause the contrasting responses between the AGM natural SIV host and RMs. In addition, cell-cycle components are progressively upregulated to a far greater extent in RMs than AGMs following infection and that this could be synonymous with a pathogenic phenotype. Importantly, we identify several gene-regulatory relationships that could potentially play an important role during SIV infection, including a potential negative feedback mechanism involving ISG15 and IFIT1 and relationships between cell-cycle linked genes BIRC5, UBE2C and AURKB, whose expression could be altered via the viral Vpr protein. Our work adds to a growing base of knowledge on describe retrovirus-host interaction that be applied to HIV-1-human system in order to better characterise and treat this pathogen.

6.7 Supporting material

Supporting file S6.1. A folder containing the results from DAVID functional enrichment charts, for all clusters.

Supporting file S6.2. A folder containing plots of true-prositive:false-positive ratios at different MI thresholds, estimated using a yeast model, given data sets equal in size to the differentially expressed gene sets for each primate species and cell type.

Supporting file S6.3. A folder containing visualisations of regulatory networks for each primate species and cell type, where all edges with an MI value > 0.3 are included. Probes are represented by nodes and gene-regulatory interactions are represented by edges. Node colour represents the membership of the gene to a particular expression profile cluster, as designated in the key file within the folder.

Supporting file S6.4. A tab-delimited text file with results used to identify regulatory relationships that differ between primate species. Columns show: (i) a unique id for the test; (ii) the name of the first GO term; (iii) the name of the second GO term; (iv) the number of gene pairs that match the two terms that appear in the data for both AGMs and RMs; (v) mean MI values for AGM gene pairs; (vi) mean MI values for RM gene pairs; the permuted P value; the corrected P value.

Supporting file S6.5. Tab-delimited text results for the permutation test to assess what virus-interacting genes might be controlling virally activated cellular regulatory subnetworks, for each primate species and cell type. Columns show: (i) the probe ID for the gene; (ii) the Entrez gene ID and gene symbol; (iii) the average subnetwork size for the original network following edge removal; (iv) the average subnetwork size for the random network following edge removal; (v) the P value; (vi) the corrected P value.

Supporting file S6.6. Tab delimited text file showing the assignment of probes to clustered expression profiles. A cluster ID of zero denotes that the probe did not sufficiently match any one of the 15 clusters.

CHAPTER

SEVEN

NETWORK-DRIVEN PERSPECTIVES OF BIOLOGICAL FUNCTION

7.1 Abstract

Systems biology research concerns the dynamic network of interactions that take place between functional subunits such as genes and proteins. Genome-scale interaction data can effectively show the "wiring" of cellular systems, such as protein-protein interactions (PPIs) networks, or epistatic relationships between genes (genetic interactions). However, these perspectives are undoubtedly limited by the scope of the given data type. As a result, contribution of any one unit to a given functional process remains difficult to ascertain. Thus, studies that integrate multiple data types may provide greater insight into biological function. Currently, it is unclear exactly what facets of biological function are captured by single-perspective interaction networks, let alone what added insight can be gained by integrating multiple networks. In this work, we identify functional subnetworks that can be captured from the interaction networks of yeast, Saccharomyces cerevisiae derived from (i) PPI, (ii) genetic and (iii) gene co-expression data. We explore the intersections between subnetworks and investigate an integrated network composed from all three types of interaction. We find striking differences in both the ability of different networks to capture certain areas of function and also the total functional space that is covered by each network. In paricular, we identify transcendent functions that require integrated information in order to be accurately represented. Thus, capture of biological function by network clustering is heavily influenced by choice of data set. Crucially, integration of interactions from multiple sources is essential to attain a comprehensive map of the complex and modular nature of function at the molecular level.

7.2 Background

In their seminal article, Hartwell and colleagues set out the case for the modular and interconnected nature of molecular function [393]. This systems-level understanding of

function has been confirmed by a multitude of biological studies in which networks have become the primary paradigm of representation, reviewed in [160]. Indeed computational analysis of large-scale data sets is undoubtedly revealing an increasingly complete functional map of the cell [394]. Usually functional modules and sub-networks are assumed to be one and the same. For example, a range of graph-properties based approaches have been developed that identify clusters in protein-protein interaction [395], metabolomic [396], gene expression [397] and genetic interaction data sets [398]. Unfortunately direct physical interactions will not necessarily occur between all molecules in a functional module and different data types will capture different types of biological event which can have varying degrees of contribution to a specific function. For example, genetic interaction data sets are informative in terms of understanding epistatic interactions, gene expression data for studying cells under different condition and protein interactions for determining physical binding and protein complex membership.

The available data sets have, however, for the most part been studied independently and, despite some attempts to integrate distinct data types, e.g., [399], an integrated understanding of molecular and cellular function remains elusive. For example, there is currently reported to be very low overlap between genetic interaction and protein interaction data [398], despite both being clearly linked to molecular phenotype. While it is clear that genetic interactions, are best explained by considering epistasis within and between modules rather than individual genes or their individual gene-products [396, 400], too much emphasis has been placed on reconciling data types as opposed to delimiting the molecules (interacting or not) that comprise a specific function. For this reason a greater effort must be applied to integrating different data types.

Biological annotation, such as Gene Ontology (GO) terms [84] are widely used to analyse functional characteristics of data sets. For example, annotation enrichment methods for characterising protein or gene sets are widespread [401] and have also been used for determining the functions of network clusters [402]. However, biological annotations are, necessarily, a proxy for true function, derived from observable traits that have been deemed important by the research community. Also, the resolution and depth of annotation schemes rely mainly on the number and structure of annotating terms. It is, therefore, import to ascertain whether network topologies can reliably "capture" the model functions defined by annotation and if so, to what extent. Furthermore, as a range of large-scale network-based data sources are available, the functional capture of different sources, including composite, multi-source data can be assessed.

Here we use a clustering approach combined with annotation enrichment to identify how different interaction data types capture function at the molecular level, using *S.cerevisiae* as a model. Three interaction networks were constructed using: (i) protein, (ii) genetic and (iii) gene coregulation interaction data. In addition, a combined network was created by integrating interactions from these networks. Each network was exhaustively clustered using graph partitioning. Biological functions represented by each

cluster were identified using GO annotation. A Voronoi tree mapping method was used to visualise annotation coverage by cluster sets. Our results show striking differences in both the ability of different networks to capture certain areas of biological function and also total functional space that is covered by each network. We demonstrate that PPI networks currently present the most functionally information-rich networks, though our results also demonstrate that genetic interactions also provide reasonable means for capturing process- and component-related functions. Furthermore, we show that integration of network data is essential for gaining the greatest coverage over the modular and often transcendent nature of biological function.

7.3 Methods

7.3.1 Network Generation

Four interaction networks were assembled where nodes represent genes and edges represent interactions between genes:

(1) A PPI network was assembled with physical interaction data from the BioGRID database [83]. Interactions were only included in the network if there was evidence for that interaction from multiple sources. The PPI network therefore represents a high confidence set of physical interactions.

(2) A genetic network was built using data from [398], which was downloaded from SGD. The network was built using a stringent P value cutoff for a genetic interaction of P < 0.001.

(3) A coregulation network was built using expression profiling data from [403] where 300 separate treatments were performed and gene expression was recorded. Two genes were defined to be coregulated if they achieved a P value of < 0.01 for expression using gene-specific error model from a single treatment. Gene nodes were connected by an edge if they were coregulated. Edges were weighted according to the frequency with which they are coregulated across all treatments, where a greater weight denotes a greater frequency, defined by:

$$\text{weight} = 1000 \times \frac{2 \times |c(a,b)|}{|c(a,!b)| + |c(b,!a)|}$$

Where a and b are coregulated genes and $|c(a,b)|$ is the number of times a and b are coregulated over all treatments and $|c(a,!b)|$ is the number of times a is coregulated but not with gene b over all treatments. For the purposes of clustering, weight values were rounded to the nearest whole number.

(4) A combined network was created by pooling all data from the PPI, genetic and coregulation networks. Edges from the different networks were weighted and weights were normalised so that the sum of edge weights contributed by each network was equal. Edges in the combined network were assigned a weight equal to the sum of weights for that edge in all contributing networks.

Note, genes were only included in the networks if they corresponded to an open reading frame in *S.cerevisiae* Genome Database (SGD, www.yeastgenome.org).

7.3.2 Cluster generation

Clusters were generated from networks using a k-way graph partitioning algorithm, kmetis [404]. For a network $G(V, E)$, with V nodes and E edges, kmetis aims to partition nodes into sets of roughly equal size and minimise the number of edges that connect node sets. Resulting node-sets and the edges that link those nodes comprise a subnetwork. For a given network we identified the set of average node-set sizes S for every given partition that could be obtained using k-way partitioning where $k \in \mathbb{N}$ and $0 < k < |V|$. For all $i \in \mathbb{N}$ where $2 < i < \frac{|V|}{2}$ we selected $s \in S$ nearest in value to i and recorded the value for k corresponding to s. We performed k-way partitioning on the network using all distinct recorded values of k. Clusters were obtained by selecting non-redundant largest-connected-components that comprise more than two nodes from the subnetworks produced by partitioning.

Kmetis will always partition the whole graph into k parts, therefore it is likely that some of the clusters we produce do not represent *bona fide* localities within the network. Therefore, clusters were scored based on comparison between mean internal path length and path lengths to other clusters from the same partition. Specifically, the path length between all nodes was calculated using the Dijkstra method [170]. The mean intra-cluster path length for all nodes was computed for all clusters and mean inter-cluster path lengths were computed between every pair of clusters from the same partition. Following this, a one-tailed/one-sample t-test was used to ascertain whether the mean intra-cluster path length is significantly smaller than the mean inter-cluster path lengths, for a given cluster. Any cluster that did not achieve a P value of < 0.05 was discarded.

The edge density, d, for a cluster, with e edges and n nodes, from an unweighted network was defined as the proportion of all possible gene-gene interactions that are present, calculated by:

$$d = \frac{2e}{n(n-1)}$$

Similarly, weighted density, h, for a cluster, with sum edge weight w edges and n nodes, from a weighted network that has mean edge weight $\bar{w}$, was defined as:

$$h = \frac{2w}{\bar{w}n(n-1)}$$

7.3.3 Functional enrichment in network clusters

We assigned function to identified clusters using the Gene Ontology (GO) [84]. GO annotation was retrieved from the GO download site. We used Fisher's exact test to identify overrepresented GO terms for each cluster. All *P* values were false discovery

rate corrected using the method described in [370] with a significance cutoff of $P < 0.05$. Additionally, we used the Matthews correlation coefficient (MCC) [405] as a measure of accuracy of our clusters for each overrepresented term.

Relative enrichment of GO terms with respect to the number of genes represented by a term was calculated for clusters from each network. First, terms were binned according to the number of genes they represent in the network data set and the proportion represented by each bin was calculated. Next, the same process was carried out for enriched terms represented by clusters with MCC> 0.2. Enrichment was defined as the proportion for enriched terms minus the proportion for all terms, for each bin. Hence, enrichment values across all bins sum to exactly zero.

In order to visually compare network coverage, semantic similarity (Lord et al., 2003) was used to determine the functional distance between genes and a tree-structure generated using neighbor-joining and represented in two dimensions using Voronoi Treemaps (Balzer and Deussen 2005; Balzer et al. 2005), implemented with GLASS (available at http://www.bioinformatics.ic.ac.uk/glass/). In this visualisation each cell represents a GO term, whose location within the panel is determined by the semantic distance to all other terms. A cell is coloured if one or more clusters from a particular network display enrichment for that term. The intensity of the colour is determined by the MCC of that cluster for the enriched term.

7.3.4 Identification of congruent network clusters

Clusters from different networks (excluding the combined network) were cross-referenced against one another and the statistical significance of the intersection in genes between two clusters was calculated by Fisher's exact test. To limit the number of comparisons to those of reasonable validity, two clusters were only compared when the size of the two gene sets was not greater than ten-fold different. MCC values were calculated to quantify the precision and accuracy of the cluster intersection. To reduce the number of statistical tests performed, P values were only calculated if the intersect between clusters (true positives) was > 2 genes and the MCC was > 0.2. Resulting P values were corrected for having performed multiple tests [370]. For each cluster from a network that had at least one statistically significant intersection (corrected P< 0.05) with clusters from another network a "best hit" was assigned to the cluster intersection with greatest MCC score. Reciprocal best hits were defined as two best hits between clusters from different networks.

7.3.5 Network visualisation and analysis

All network visualisations were produced using Cytoscape [185]. Edge and node betweenness coefficients were calculated using the NetworkAnalyzer Cytoscape plugin [165].

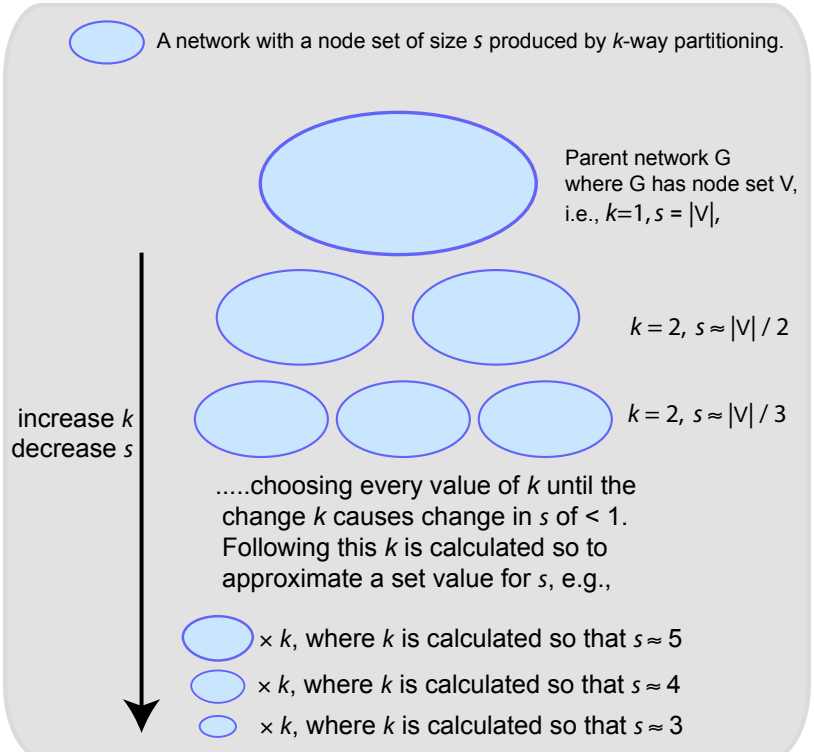

Figure 7.1: Network partitioning methodology. Interaction networks were partitioned by using k-way partitioning. Many different values for k were used in order to produce an extensive set of partitions with a wide range of sizes. Clusters were selected from network partitions (see Methods for more detail).

7.4 Results

7.4.1 Interaction networks and clustering

Four interaction networks were assembled from large-scale *S.cerevisiae* data: a protein-protein interaction (PPI) network consisting of 12182 interactions between 3339 genes, a genetic interaction network consisting of 42546 interactions between 3529 genes, a coregulation network consisting of 3006725 weighted interactions between 4358 genes, and a combined network consisting of 3052053 unique weighted edges between 5489 genes. Network clusters, corresponding to connected subnetworks, were produced for all these interaction networks (Figure 7.1). A total of 9590, 9227, 12889 and 11383 unique clusters were obtained from the PPI, genetic, coregulation and combined interaction networks, respectively. These clusters include from 3 to ~ 3000 genes and represent an extensive and thorough breakdown of each network into connected subnetworks. Furthermore, these subnetworks were filtered to ensure that each subnetwork represents a genuine locality within the network.

In order to validate the clusters we investigated their edge density. Cluster density plots are shown in Figure 7.2. Figures 7.2B and 7.2D highlight that more than 5% of clusters of up to ∼25 genes from both the genetic and PPI networks consist of densely connected cliques where around half of all possible gene-gene interactions are present. In the case of the coregulation and combined (weighted) network clusters, we define a

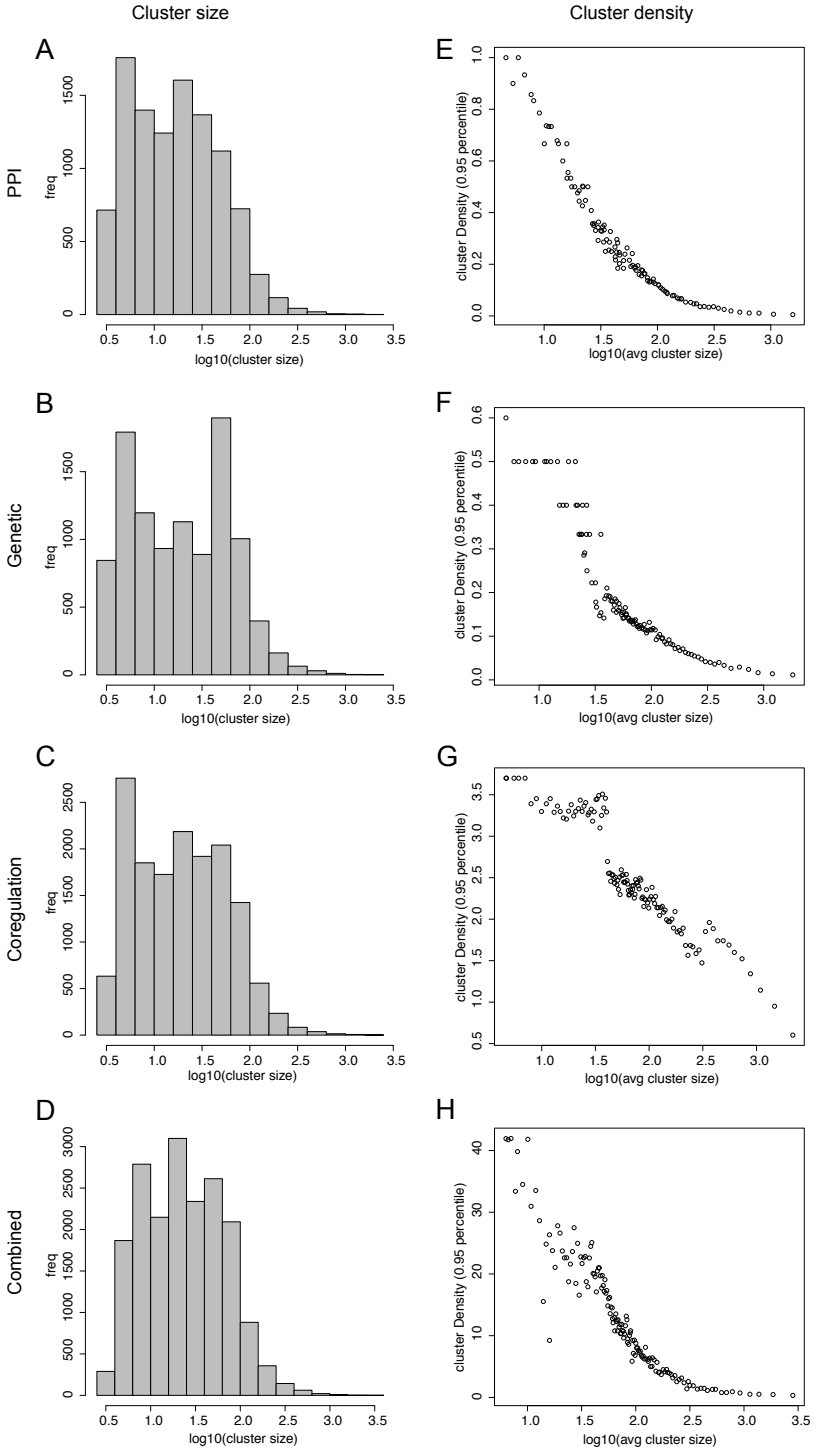

Figure 7.2: Summary of network cluster characteristics. Two types of plot are shown: (i) cluster size against cluster frequency (A, B, C and D), and (ii) cluster size against cluster density (top 95th percentile, charts E, F, G and H).

weighted density measure that takes intra-cluster edge weights and overall mean edge weight into account (Figures 7.2F and 7.2H, respectively). An accellerated decrease in weighted density, as cluster size increases, is clearly apparent for clusters with $>\sim$40 genes. Therefore, clusters of a range of sizes successfully capture cohesive subgroups of interacting genes.

Table 7.1: Average MCC scores of clusters with overrepresented GO terms

		Network			
		PPI	Genetic	Coregulation	Combined
BP	Enriched terms	2271	2265	2169	2639
	Total terms	2710	2736	2655	2694
	Coverage (%)	84	83	82	98
	Average MCC	0.43	0.25	0.19	0.31
MF	Enriched terms	893	1006	1012	1237
	Total terms	1541	1457	1346	1333
	Coverage (%)	58	69	75	93
	Average MCC	0.42	0.25	0.22	0.29
CC	Enriched terms	682	558	556	651
	Total terms	660	632	629	713
	Coverage (%)	93	80	80	97
	Average MCC	0.60	0.26	0.19	0.45

7.4.2 Functional enrichment in network clusters

To investigate what biological functions are captured by network clusters we performed GO enrichment analysis on each cluster. Many network clusters consist of gene sets that are enriched for specific biological functions. Table 7.1 summarises functional enrichment for clusters from each network, for each of the three ontologies: biological process, molecular function and cellular component. Matthews correlation coefficient (MCC) was used to measure the accuracy with which clusters capture specific GO term annotations. MCCs were significantly different for enriched terms for different networks across all three ontologies (Kruskal-Wallis rank sum, PPI, Genetic and Coregulation all P $<$ 2.2e-16). This result indicates that the network data sources have significantly different abilities to capture the biological functions represented by GO. Clearly, the PPI network captures functional annotations with the greatest accuracy, however, the greatest coverage of GO is captured using a combination of data sources.

GO terms can refer to very common functions, i.e, be assigned to a considerable fraction of all genes; refer to specialist functions, i.e., be assigned to very few genes, or lie somewhere between these two extremes. In order to ascertain whether the functional categories enriched in clusters from different networks were biased towards capturing general (or more specialist) functions, using the number of genes represented by a terms as a measure. Two types of plot were produced: first, bar plots A-D in Figure 7.3 show the relative abundance of enriched GO terms against the number of genes represented by the term, for clusters of each network. Here, a positive value represents relative over-representation while a negative value represents relative under-representation. Second, density plots E-H

in Figure 7.3 show how accurately captured enriched GO terms are by the clusters of each network, using the maximum MCC score achieved by each enriched term as a measure of accuracy. Dense shading represents a greater frequency of GO terms. The bar plots show that PPI clusters and clusters from the combined network capture functions with relatively less bias than either the genetic or coregulation clusters; the genetic network clusters displaying a bias for capturing medium specificity functions at the expense of highly specialist functions, whereas a roughly opposite trait can be observed for coregulation network clusters. Density plots indicate that the PPI network and to a lesser extent the combined network clusters, outperform the genetic and especially the coregulation network clusters at accurately capturing more general GO terms, i.e., assigned to > 10 genes. However, from the bar plots it is clear that very general functions with a membership of over $\sim$100 genes are difficult to capture from any of the networks, relative to more specialist functions with fewer members. Furthermore, the density plots indicate the accuracy with which the function is captured diminishes where the function is more general.

To further investigate the capture of functional categories by network clusters, we visualised gene ontology terms using a Voronoi tree-mapping approach (Figure 7.4, see Methods for details). In the tree maps, each cell represents a GO term, where the layout is determined by the semantic similarity between the terms, i.e., cells that are grouped together represent a similar function. Tree-maps were created for each GO ontology and identical maps were used for clusters of each network. Thus, maps from different networks are directly comprarable, the equivalently positioned cells from each representing the same GO terms. The intensity of cell shading indicates the greatest accuracy with which the GO term is captured by a network, using MCC score as the accuracy measure.

These maps highlight the disparity between the ability of network clusters to capture certain types of functional data: cellular component annotation appears to be the easiest type of biological function to capture, using any any type of network data. Conversely molecular function is more difficult to capture. However, this tree-mapping approach also highlights that certain functional areas within each ontology can be either successfully captured, or are difficult to capture, using the network data. For example, a section of ~ 20 cells near the top left corner of the map for cellular component functions captured by genetic network clusters are not shaded, whereas the equivalent cells are predominantly shaded in both maps for coregulation and PPI networks. In contrast, some areas are clearly shaded in all maps from the same ontology. Therefore, the ability of different networks to capture functional relationships, is not only related to the type of network data but the functional area in question.

Creation of a composite tree-map, where the cell colours represent the network from which the terms are most accurately captured, allows for further comparison (Figure 7.5). From the trees on which the maps are based, we can identify "clades", defined as monophyletic areas within an ontology, that are best characterised by clusters from a single

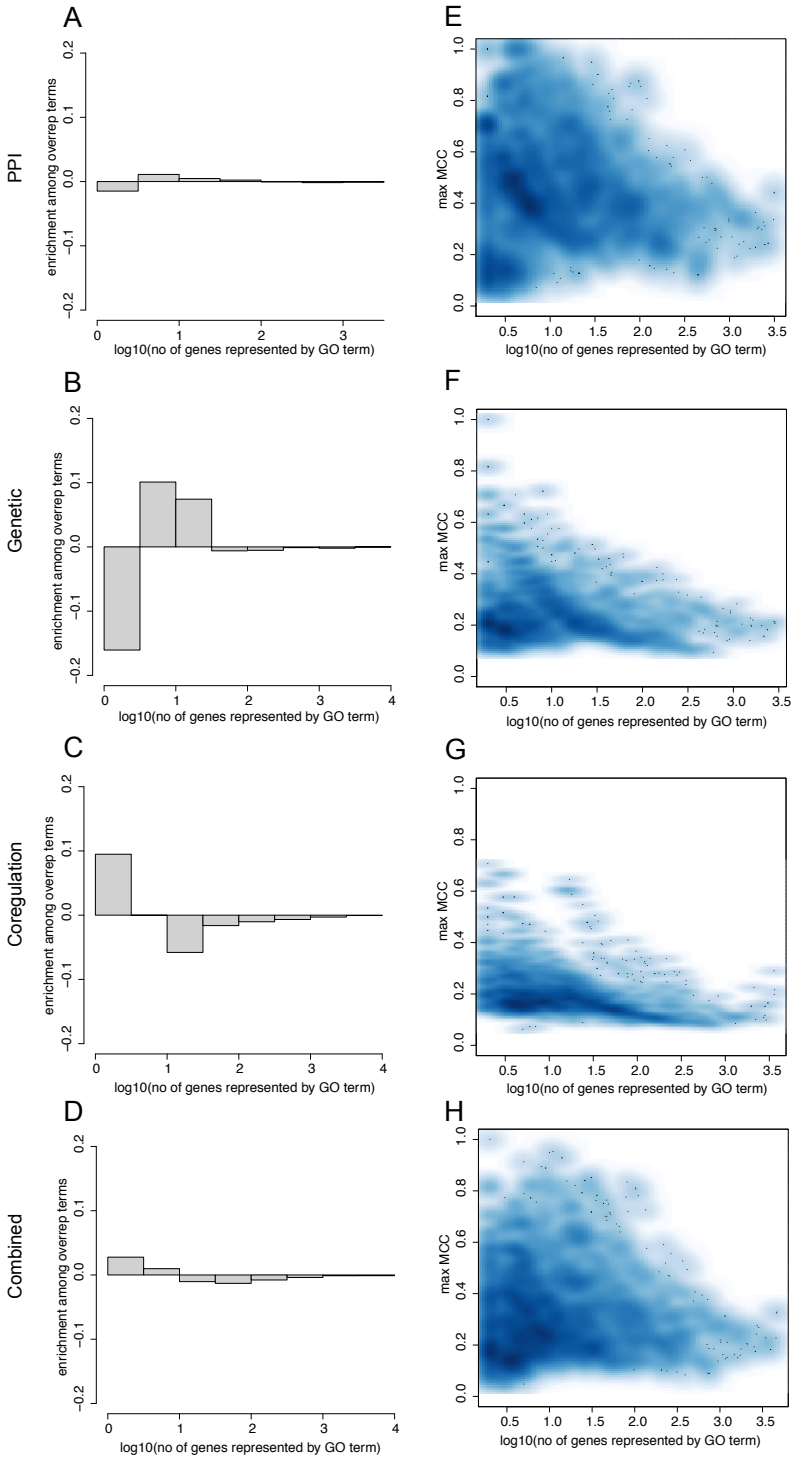

Figure 7.3: GO enrichment among network clusters. Two types of graph are shown: (i) bar plots A, B, C and D show the relative level of enrichment of GO terms pertaining to more specialist, or general functions, measured by the number of genes represented, from each network. Here, a positive value represents relative enrichment of GO terms of the given size, while a negative value represents relative lack of GO terms of the given size. (ii) Density plots E, F, G and H show overall relationship between the number of genes represented by the GO term (x-axis) and the maximum accuracy with which the term is captured by clusters from each network, measured using MCC (y-axis). Denser of shading represents a greater number of GO terms.

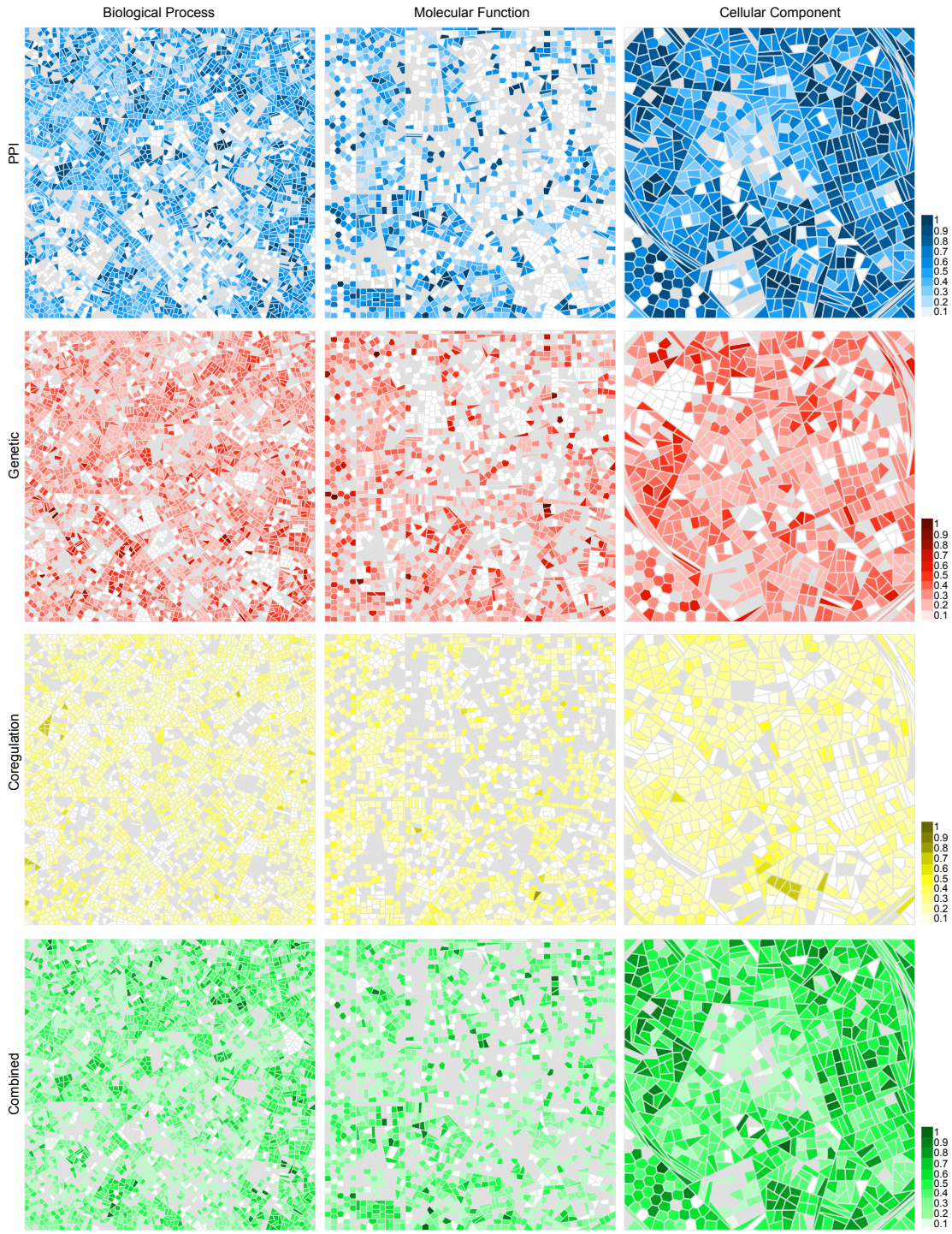

Figure 7.4: GLASS visualisation of enriched GO terms. Each cell represents a GO term and is coloured blue, red, yellow or green if one or more clusters are enriched for that GO term in the PPI, genetic, coregulation or combined networks, respectively. The intensity of each coloured cell shows the best MCC of the clusters with enrichment for that term. Grey coloured cells are those GO terms which have only one or no associated genes in that network.

network. Examples of these clades are outlined in Figure 7.5. Both biological process and cellular component ontologies are both predominantly best represented by clusters identified in the PPI network. Indeed, in the biological process ontology the largest clade that is characterised by a single network consists of 14 terms associated with gene silencing all of which are enriched in clusters from the PPI network. Likewise, 20 terms associated with the proton transporting ATPase complex of the cellular component ontology are all enriched in clusters from the PPI network. These findings indicate that annotation of biological function using GO is centered around physical PPIs.

Despite the dominance of PPI network clusters for accurately capturing functional annotation, there are also clades from each ontology that are most accurately captured by clusters from other networks. Clusters from the combined network best represent 12 terms associated with nucleoside and ribonucleoside biosynthesis in the biological process ontology, whereas in the molecular function ontology we can identify a clade of four terms associated with oxidoreductase activity that are best characterised by clusters in the coregulation network. These findings indicate different areas of biological function my be best represented by different types of biological data, including protein interactions, genetic interactions and coregulatory relationships. Interestingly, all three ontologies have areas that are best represented by clusters from the combined network. Therefore, the integration of data from multiple biological networks can improve recapitulation of certain biological functions, when compared to analysis of any network in isolation.

Examples of clusters that best characterise a single GO term can be found in Figure 7.6. The mitochondrial small ribosomal subunit cellular component term is well characterised by a PPI cluster with 28/30 members annotated with the term and a MCC of 0.94 (Figure 7.6A). The Inosine monophosphate (IMP) biosynthetic process term is best characterised by a cluster from the genetic network (Figure 7.6B). Here, four out of six genes in the cluster are annotated with the term and represent enzymes in the purine biosynthesis pathway. Terms from the molecular function ontology are enriched with the greatest coverage by the clusters from the combined network (Table 7.1). The synaptosomal-associated protein (SNAP) receptor activity molecular function term is best represented by a combined network cluster which incorporates edges from all other networks (Figure 7.6C). These examples further demonstrate that successful capture of functional relationships by network data depends on a combination of the specific biological function being sought and the type of network data being interrogated.

7.4.3 Congruent network clusters

As a second line of study, and omitting the combined network, we investigated whether the clustering of different networks had resulted in the production of congruent clusters, i.e., pairs of clusters from different networks that have significantly intersecting gene sets. By comparing clusters from the PPI, genetic and coregulation networks, we identified statistically significant gene intersections and subsequent “best hits” and “best reciprocal

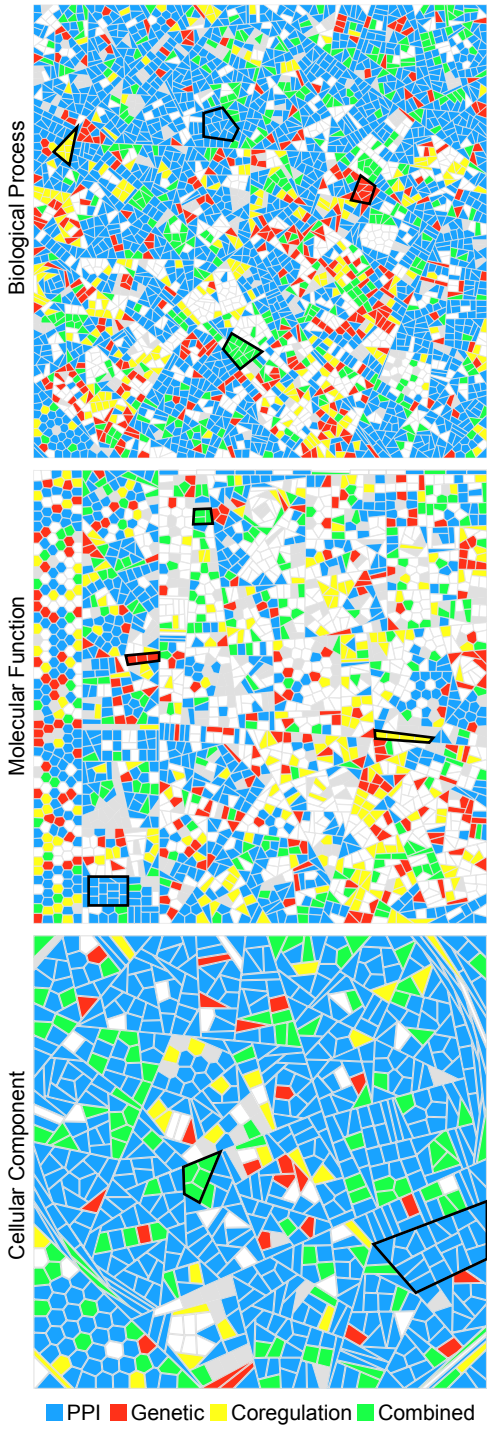

Figure 7.5: GLASS visualisation of enriched GO terms. Each cell represents a GO term and is coloured according according to clusters that have the highest MCC for the enriched term. Blue, red, yellow and green colours indicate that the cluster with the highest MCC is from the PPI, genetic, coregulation or combined network, respectively. Grey coloured cells are those GO terms which have only one or no associated genes in any network. Areas ringed in black show complete areas of the ontology which are best characterised by a single network.

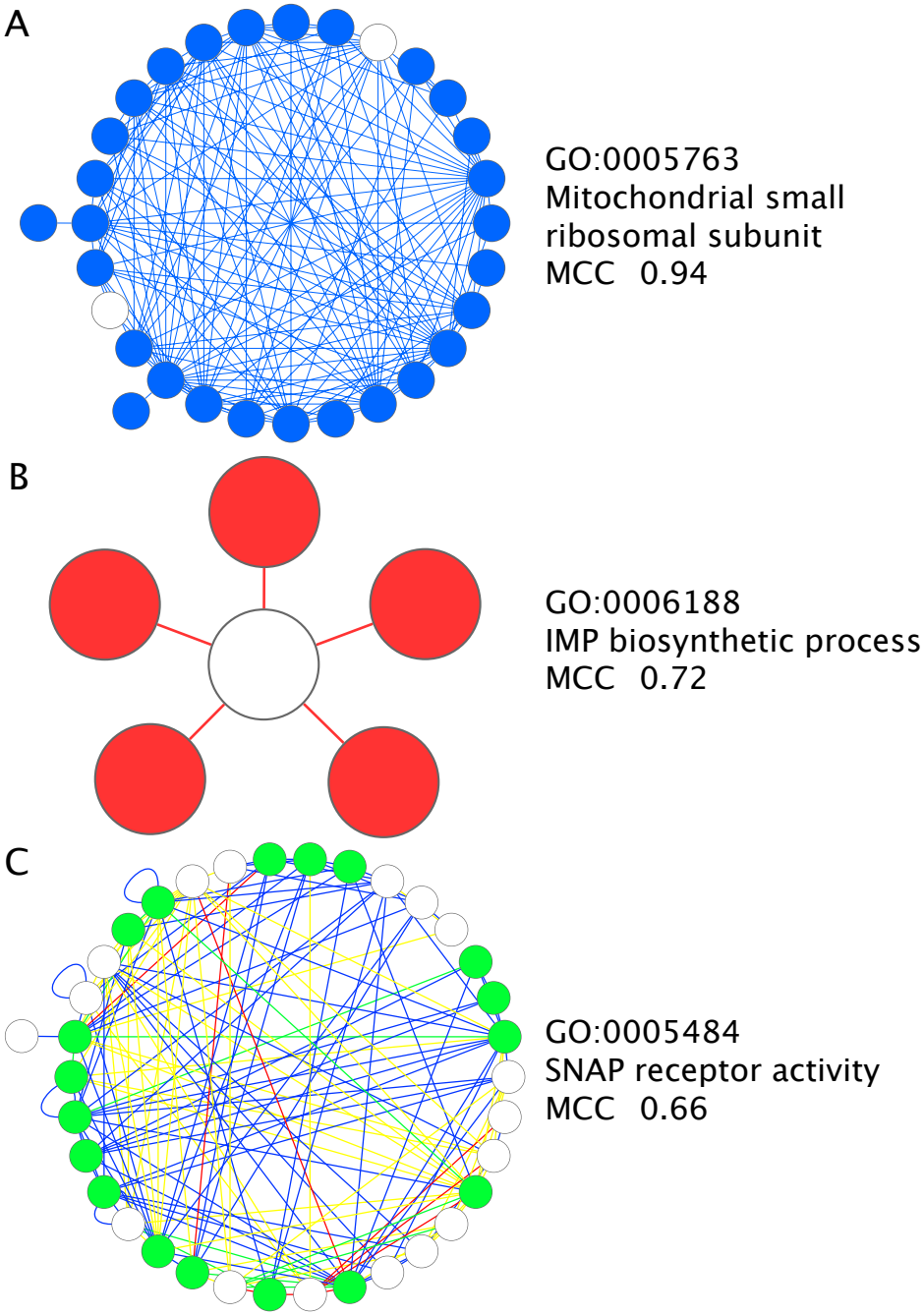

Figure 7.6: An example of functions best characterised by a certain type of network data. The mitochondrial small ribosomal subunit GO term is best represented by a cluster from the PPI network (A). A genetic cluster best represents the IMP biosynthetic process term (B). Finally, the GO term, SNAP receptor activity, is best represented by a cluster in the combined network, created from all nodes and edges in the PPI, genetic and coregulation networks (C). Nodes are coloured blue, red or green if they are present in the PPI, genetic or combined network respectively and are associated with the enriched GO term. White nodes represent nodes in a cluster that are not associated with the enriched GO term. Edges are coloured blue, red, yellow or green if they are present in the PPI, genetic, coregulation or combined network.

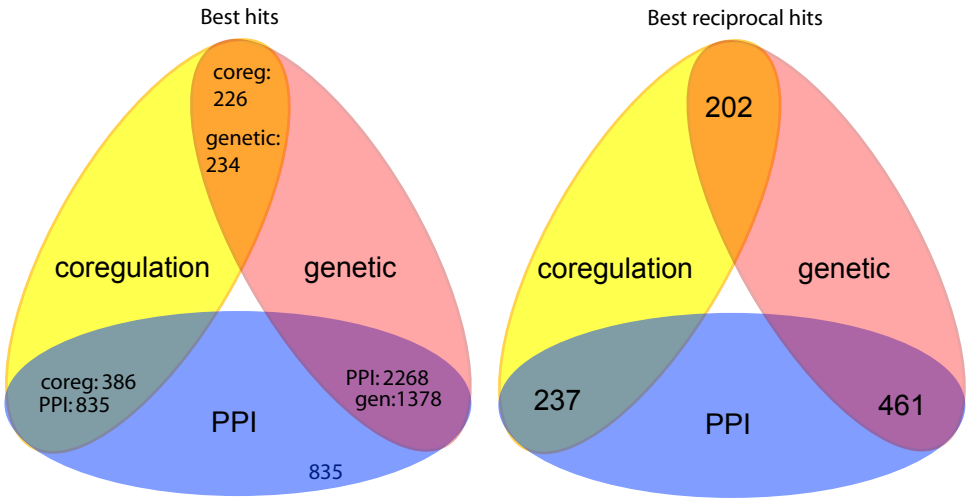

Figure 7.7: Results from best hits analysis. (A) Number of clusters from one network (named outside of the intersection) that are a best hit to a cluster from another network (named within the intersection). (B) Number of best reciprocal hits between clusters from two networks.

hits" (see methods for details) between the clusters of two networks, using MCC score as a measure for accuracy. A best hit represents a significant gene intersection between two clusters where one cluster best matches the other (determined using the maximal MCC). A best reciprocal hit, again, represents a significant gene intersection, though in this case both clusters are the best match to one another. Thus, best reciprocal hits indicate the strongest congruence between clusters from different networks. A summary of best hits and best reciprocal hits are given in Figure 7.7.

To obtain a high-level insight into the congruence relationships between clusters from different networks, we visualised best hits (and best reciprocal hits) using a network, where nodes represent clusters and edges represent the hits (Figure 7.8). From a total of 4669 clusters that are involved in a best hit with one or more clusters, 3689 clusters are involved in a best hit with just one other cluster, however, some clusters have many more best hits. Indeed, the node degree fits a power-law distribution (figure 7.9).

A repeated topological pattern of this network is for the cluster of one network to be connected to a large number of clusters from one other network e.g., Figure 7.8, sections A-D. There are 115 clusters that have a degree > 7 (top $\sim 2\%$). These clusters, that we refer to as *high-degree clusters* are a particularly good hit to a core gene set that is repeatedly identified by k-way clustering, for different values of k, from another network. Therefore, high-degree clusters and their hits appear to be robust subnetworks of genes that transcend multiple networks. Thus, we hypothesised that high-degree clusters might have particular functional significance. Indeed, high-degree clusters and the clusters that are their best hits (together termed *high-degree neighbourhoods*) are: (i) significantly more likely to be enriched for one or more GO term and (ii) capture GO functions with significantly better accuracy than clusters that are not conserved, in all networks ($P < 2.2 \times 10^{-16}$, two-tailed Mann Whitney U test, in all cases). This result indicates that

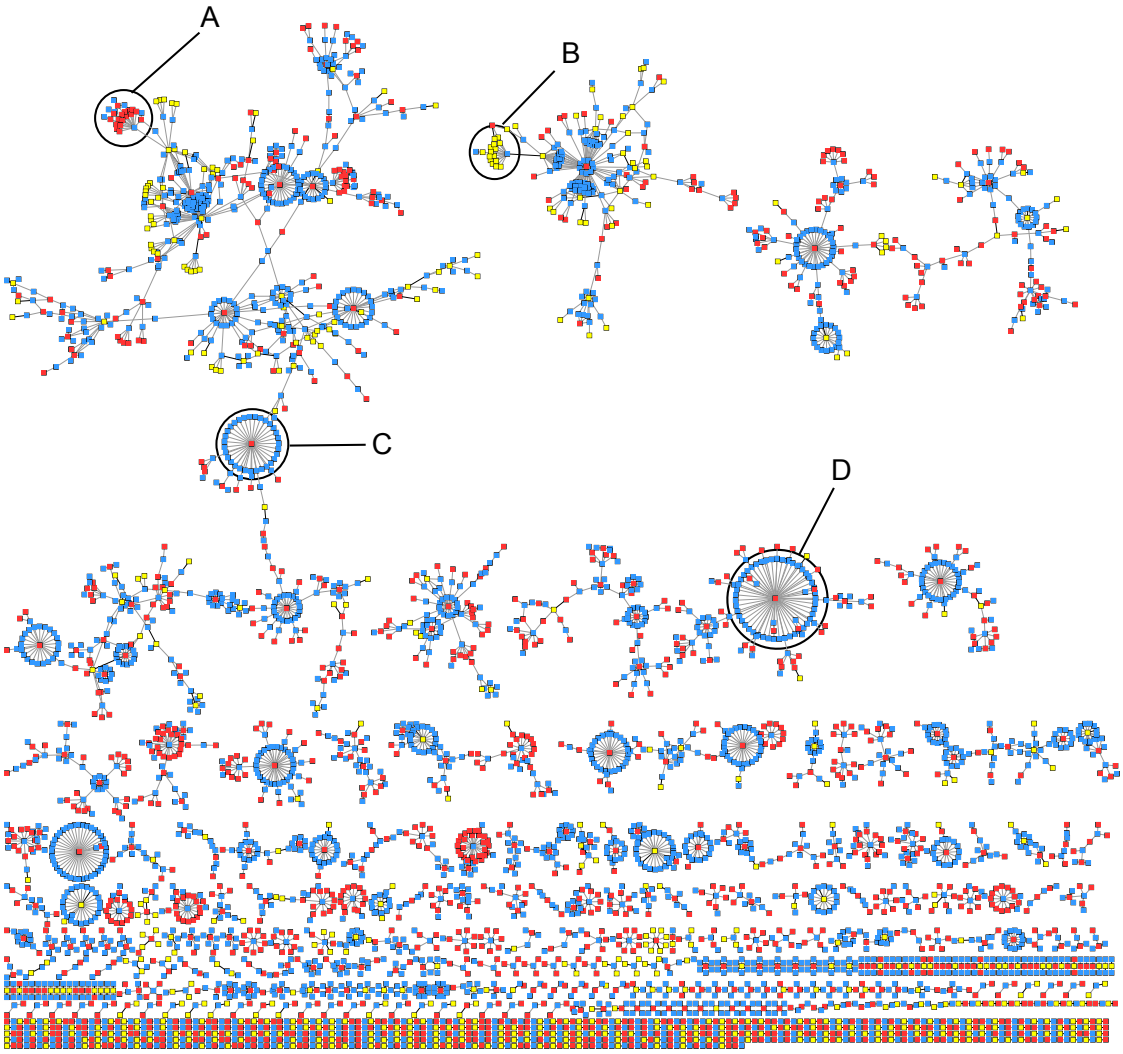

Figure 7.8: Network of "best hits" between clusters of PPI, genetic and coregulation networks. Nodes represent clusters and edges between clusters represent a statistically significant intersection of > 2 genes with an MCC > 0.2. Only the best intersection between each network comparison, defined by MCC score, is shown. A-D highlight high-degree neighbourhoods that consist of a node with degree > 7 and all neighbours of that node.

the conserved clusters are more likely to be *bona fide* functional modules. Furthermore this result highlights the value of integrating information between networks in order to validate network clusters.

To further investigate the merits of integrating network data, we devised a method for testing whether new, biologically relevant functional links can be made by merging strongly congruent clusters. Network clusters from all networks are frequently enriched for multiple biological functions (Table 7.2), we term this co-enrichment. Interestingly, many pairs of GO terms are co-enriched in each network, including pairs from the same and different ontologies and also from both related (descendent or ascendent) and unrelated GO terms from the same ontology (table 7.3). New co-enriched GO term pairs are produced by merging best reciprocal hits from each network combination ((table 7.2).

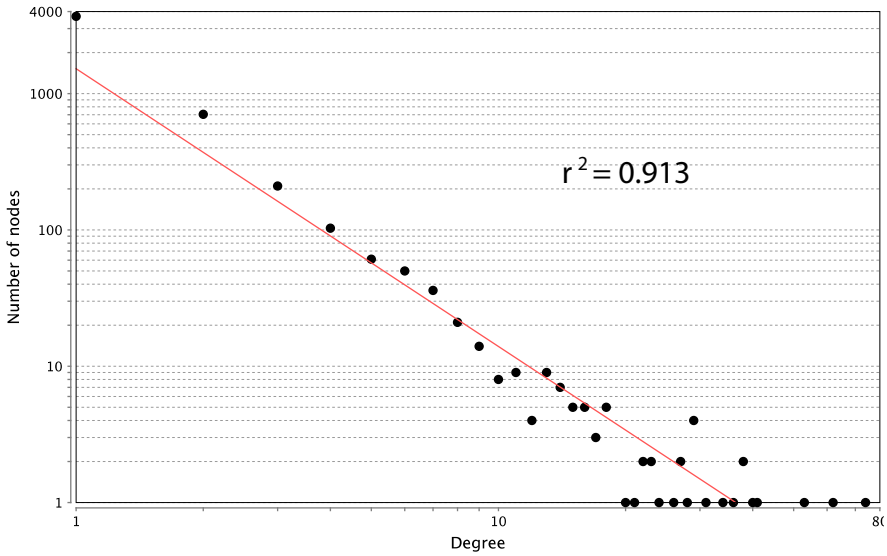

Figure 7.9: Node degrees of the network of best hits fit a power-law distribution. The power law distribution indicates that a large number of nodes have a small degree and only a small proportion of nodes have a high degree in the network.

Table 7.2: GO term enrichment among conserved clusters

Source network	Clusters	Number	Mean no. of enriched GO terms	Mean max MCC per cluster
PPI	high-degree neighbourhoods	2801	135.2	0.509
	all	9591	101.4	0.468
Genetic	high-degree neighbourhoods	1567	81.4	0.256
	all	9228	24.9	0.157
Coregulation	high-degree neighbourhoods	601	15.4	0.138
	all	12889	7.1	0.095

These new pairs represent biological functions that are only co-enriched in subnetworks that comprise interaction data from more than one source.

For example, Figure 7.10 shows a PPI and a genetic cluster that are best reciprocal hits, merged into a subnetwork. Several of the nodes identified by both clusters are clearly highly central to this subnetwork and have high node betweenness coefficients, e.g, YCL061C, YMR048W, YLR288C and YPL194W. Furthermore several genetic interactions between these central genes also have high edge betweenness coefficients. Individually, both clusters are significantly enriched for genes involved in DNA

Table 7.3: GO terms that are co-enriched in network clusters

			Co-enriched GO term pairs		
Source	Network	All	Same ontology, related	Same ontology, unrelated	Different ontology
Whole network	PPI	3910331	1.1%	41.8%	57.0%
	Genetic	1035343	3.0%	41.2%	55.8%
	Coregulation	224188	6.6%	39.2%	54.3%
Best reciprocal hits	PPI vs. Genetic	56154	0.3%	38.5%	61.3%
	PPI vs. Coreg	17201	0.2%	34.4%	65.5%
	Genetic vs. Coreg	3817	0.03%	41.2%	58.8%

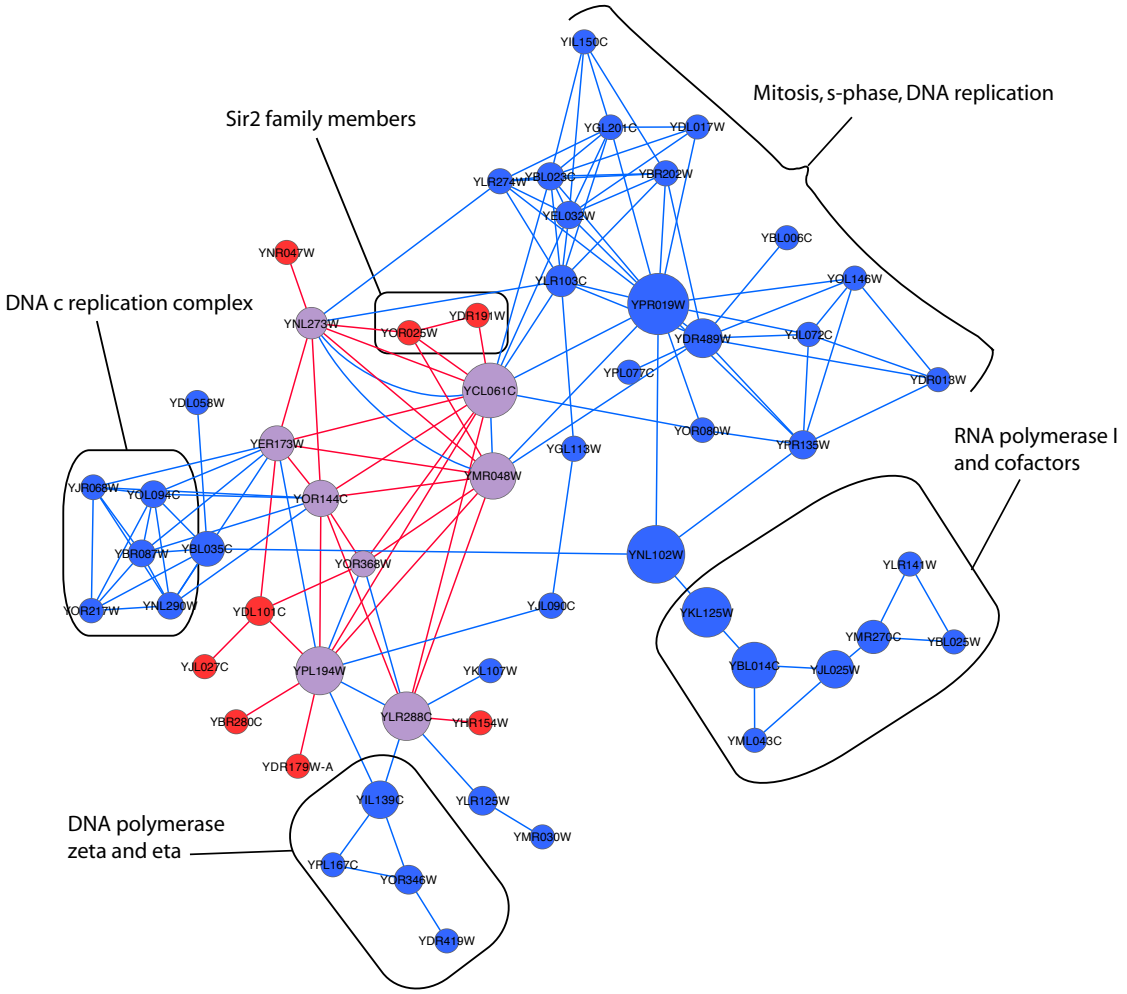

Figure 7.10: A subnetwork that represents merged clusters that are best reciprocal hits. Nodes represent coding genes. Blue nodes genes from a PPI network cluster, red nodes represent nodes from a genetic network cluster. Nodes shown in purple represent genes that are present in both the PPI and genetic clusters. Node diameter is proportional to the node betweenness. Blue edges represent PPIs between encoded proteins and red edges represent genetic interactions between genes. Edge diameter is proportional to the edge betweenness coefficient. This network is a functional module involved in control of DNA replication.

replication and cell cycle control (GO:0006260 and GO:0007049). However, by combining these two clusters 81 new functional links are made between GO terms that are not co-enriched in clusters from any single network. Specifically, presence of Sir2 family genes (YOR025W and YDR191W) that are NAD(+)-dependent histone deacetylases involved in cell cycle progression [406] – cause the new links, such as linking NAD binding (GO:0070403) to S phase of mitotic cell cycle (GO:0000084) and DNA replication factor C complex (GO:0005663). The Sir2 family members genetically interact with several proteins that are central to the subnetwork, including YCL061C (S-phase checkpoint protein) and YMR048W (replication fork associated factor). Hence, by integrating network data novel, biologically meaningful functional links are made.

7.4.4 Discussion

From the three networks containing a uniform type of data, biological functions, whether specialist or general, are most accurately and completely captured by the PPI network (Table 7.1 and Figure 7.5). Unsurprisingly, this includes cellular component GO annotations – an ontology that characterises physical cellular components, such as protein complexes [84]. In contrast, the molecular function ontology, that gives enzymatic and biochemical properties of gene products is only partially captured by PPI interactions, indicating that subnetworks that represent physical components and protein complexes contain both biochemically similar or unrelated subunits. More remarkable however, is the almost complete extent to which PPI clusters captured functional annotations from both the cellular component (93%) and biological process (84%) ontologies (Table 7.1). The biological process ontology embodies quite pragmatic, tangible notions of cellular function, such as apoptosis (GO:0006915) and glucose metabolism (GO:0006006) [84]. Our results show that the combinatorial activity of proteins is prevalent in the majority of established GO biological processes. Moreover, this finding shows that the content of the PPI network for *S.cerevisiae* is sufficient in coverage and functional information to capture subnetworks from the majority of biological processes and cellular components.

All explanations for *bona fide* genetic interactions imply a functional relationship between genes. Undeniably, these relationships have, from a organismal perspective, some functional relevance, as they are expressed phenotypically. For example, Chen and Thorner [407] show that epistatic interactions exist both within and between genes of two given pathways; a mitogen-activated protein kinase pathway and a protein kinase A pathway. The functional processes influenced by these pathways, include cell elongation and cell adhesion.

A systematic study to detect within- and between-pathway genetic interactions in *S.cerevisiae* was performed Kelley and Ideker [399]. In their work, the "pathways" they assess actually refer to a cohesive subnetwork of proteins in a PPI network, and thus, their pathways are theoretically equivalent to the PPI network clusters that we investigated. Kelley and Ideker classify genetic interactions as within-pathway or between-pathway, the former indicating a genetic interaction between elements of the same subnetwork and the latter indicating a genetic interaction between elements from a separate subnetwork. Thus, within-pathway genetic interactions are indicative of a functional PPI subnetwork, such as a protein complex. Kelley and Ideker [399] identify that many between-pathway interactions link interdependent functional relationships. Thus, between- pathway interactions indicate functional PPI submodules that are essential agents for achieving a single, greater functional process. Theoretically, if this notion were entirely upheld by GO annotation, genes involved in *bona fide* genetic interactions should always both be attributed with a given process or component annotation that captures their common cellular activity, be that either a relatively specialist or a very general function. Indeed, we show that 83% of the biological process terms and 80% of cellular component terms are cap-

tured by genetic interactions. Therefore, the interaction network derived from genetic interactions is clearly a reasonable choice of data for capturing physically interacting and process-related functions.

The PPI data is a from a compendium of experimentally validated PPIs, whereas the other data sources are extrapolated from high-throughput experiments. Therefore, the quality of the derived networks, in terms of type I and type II error rates, are unlikely to be equivalent. Hence, direct comparison of the performance of each network at capturing aspects of biological function will undoubtedly not only reflect the information available from the type of interaction but also the error rate.

Clearly, the co-regulation network performed least effectively overall at capturing biological functions. However, this may be an unavoidable feature of gene expression data. As a measure of transcript abundance, gene expression data can only provide an estimate for the relative change at the level of the protein. Any errors are likely to affect the co-expression network. Despite these drawbacks, like the genetic and PPI networks, the co-expression network captures the majority of functions embodied by biological process and cellular component GO ontologies. Interestingly, the co-expression network captures more molecular function GO terms than either the PPI or genetic networks (Table 7.1). Indeed, studies that investigate genes encoding enzymatically similar proteins may benefit from utilising gene co-expression information.

Our combined network is a weighted union of the PPI, genetic and co-expression networks. Although a more refined method for integrating these data could be developed, this network permits direct comparison between the integrated and uniform networks. The most notable aspect of the integrated network is that the coverage of captured annotations is almost complete for each GO ontology. Furthermore, molecular function GO annotations especially, are more successfully depicted by combined data than by any other network we investigated (Table 7.1). However, the accuracy with which these functions are captured is generally not as great as for the PPI network (Table 7.1). This is perhaps due to a greater level of noise in the combined than PPI network, stemming from the co-expression and genetic interaction data. Yet it is conceivable that a more refined data integration method, involving, for example, machine-learning of *bona fide* functional links, could attenuate the error rate. Both the clusters derived from the combined network, and congruent clusters represent functional subsystems that transcend multiple types of interaction data. Investigation of congruent network clusters revealed that biologically relevant functional links can be identified by integrating data from multiple sources. Therefore, complete perspectives on the integrated nature of biological functionality is only accessible using a composite data source.

In conclusion, despite the potential for different error rates in the datasets, our results show that each network is capable of caturing certain areas of biological function with greater accuracy than the other networks that were investigated (Figure 7.5). Thus, the choice of interaction dataset directly influences the ability of networks to depict spe-

cific functional relationships. Importantly, we have shown that combined network data can represent a greater range of biological functions than networks that utilise a single, uniform data source.

CHAPTER

EIGHT

FINAL DISCUSSION

The majority of work in this thesis concerns the interaction between viruses and their host cells. Specifically we researched HIV-1, HCV and SIV, using large-scale data sources and involving computational analysis. We have presented biological insights into virus infection, generated reusable data sets, produced publicly available software, developed novel methodologies for analysis of host-virus interaction data and made findings that could potentially be important for the development of novel or improved antiviral strategies. Specific contributions to these areas are briefly summarised:

Insights into virus infection have been made throughout the research in this thesis. In chapter 3 section 3.4, our use-case study for JNets, we explore the co-targeting of human proteins by FDA-approved drugs and HIV-1 proteins. We find that HIV-1 targets many of the same host factors as antineoplastic agents and immunosupressive drugs, a result that illustrates the pathogenic mechanisms by which HIV-1 disrupts the immune response and possibly also promotes malignancy. In chapter 4, using a clustering approach, we discovered that there are certain patterns of interaction and HIV-1-induced perturbation associated with specific cellular subsystems and provide simplified and coherent networks that describe the interaction of HIV-1 with these subsystems. In chapter 5 we use microarray analysis to identify differences in gene expression between HCV infected and uninfected cell types, from which we make novel findings, both in terms of the genes and the functions that are differentially regulated due to infection, such as an enrichment for genes encoding zinc-finger transcription factors. Furthermore, we identify novel host factors that are potentially anti- and pro-viral. Finally, in chapter 6 we identify gene expression profiles specific to SIV infection in natural and pathogenic SIV infections, to which we link specific functional significance and we also identify active gene-regulatory relationships between differentially expressed genes that play a role in regulating the immune response of infected monkeys. We show that immune gene-regulatory relationships are likely to be important factors for determining the outcome of SIV infection, a feature of infection that is likely to have parallels in HIV-1 infection of human hosts.

Generation of new, publicly available data sets is an important part of scientific research, as it allows subsequent studies to integrate new information to provide additional

perspectives on data elements and improved insight into results. For example, in our research we have used results from siRNA screens in chapters 4 and 5, in order to indicate host factors essential for virus replication and throughout we have used Gene Ontology [84, 140] annotation to infer the functions of gene sets.

We have generated gene expression data from HCV infected and non-infected hepatoma cells has been made publicly available in a MIAME compliant format [408]. This gene expression data could be used for research into HCV infection and also research that considers the differences in gene expression between hepatoma cells used in HCV cell-culture. In addition, we have published a list of host factors that are present in HCV replicase complexes which is the first large-scale data set of this kind and a valuable resource for further investigating HCV replication.

In addition to these primary data sources, secondary data sets have been produced such as HIV-1 interaction patterns and their associated subsystems – this meta-data may be useful in subsequent studies that require classification of HIV-1 interacting host factors, or HIV-1 interactions into functionally related subsets. For example, in the case where an additional set of host genes has been identified as important to HIV-1 infection, cross referencing the genes from this set against the host subsystems that we define would indicate what HIV-linked host-cellular functions are present in the data and also indicate prominent sets of HIV-1 interactions. As a second example, we define lists of potentially anti- and pro-viral host factors, though these are not experimentally confirmed, they could still be useful in integrative analysis, or alternatively, to contribute to the scientific rationale behind more intensive, small-scale experimental research of host factors involved in HCV infection.

Production of publicly available software is also hugely valuable to researchers. Indeed in this research we have employed many publicly available software tools and packages, including Cytoscape [185], R statistical software including several biology-specific Bioconductor packages [312, 313] and tools from the database for annotation, visualization, and integrated discovery (DAVID) [137]. In this thesis we have presented JNets, a software tool for network visualisation with a linked annotation enrichment analysis function. As a stand-alone application JNets was not intended to compete with Cytoscape, that outperforms JNets both in terms of functionality and extensibility. However, until the release of Cytoscape Web [409], JNets was unique in its functionality as a web-deployable network-visualisation applet, with the ability to customise visualisations and inspect annotation. However, JNets does remain the only web-deployable network visualisation software that performs enrichment analysis.

In addition to new software, we have developed novel methodologies for data analysis, that may be copied, adapted or simply inspire a related research method in the future. In chapter 4 we use a novel implementation of biclustering combined with a network-based permutation test to identify statistically significant patterns of HIV-1-host interaction patterns, we then apply a distance-based tree inference algorithm to the significant clusters

in order group them into higher level functional sets (figure 4.1). Though there are established methods within this pipeline, this is a unique scheme for analysis of interactions and clearly demonstrates that HIV-1 interactions are specific to certain host subsystems. Likewise, in chapter 5, we use an established functional enrichment measure [137] combined with a network approach as a simple method to visualise and identify functional overlap between sets of genes (figure 5.5).

Chapters 4, 5 and 6 of this thesis discuss specific host functional subsystems, cellular pathways or individual factors that may play an important role in viral infection, with a focus towards identifying prospective drug targets. Though it is merely speculative to suggest how these chapters may contribute to the development of new antiviral drugs, future work may be influenced by the results presented. Thus, we briefly highlight aspects of these chapters that are most likely to provide relevant insight in the field of target discovery. In chapter 4, we identify host subsystems that engage with HIV-1 through certain patterns of protein interaction that comprise host factors enriched for those essential for HIV-1 replication (e.g., mRNA transport components, see section 4.4.4). Perturbation of these subsystems could disrupt these essential subsystems and prevent a necessary set of HIV-1 interactions from occurring. Proteins that are part of, or have an influence over these essential host subsystems could be targeted by a novel antiviral agent. In chapter 5, we discuss a number of cellular factors that are potentially pro- and antiviral during *in vitro* HCV infection of hepatoma cells. Though the importance of these factors requires further investigation, our work provides a set of putative genes and proteins that might be exploited by an anti-HCV treatment. Such a treatment could operate by targeting and directly antagonising proviral proteins, or by mimicking or even enhancing the functional activity of antiviral proteins. Lastly, in chapter 6, we highlight potential gene-regulatory interactions that could be important determinants for disease outcome in SIV infection. For example, we highlight an interaction between ISG15 and IFIT1 that appears to have a role in negative regulation of the interferon-mediated immune response. Elucidation of specific pathways and interactions that influence disease progression in SIV infection may provide valuable insight for researching and developing drugs that curtail the onset of AIDS in HIV infected patients.

The research in this thesis has involved the integration of a wide variety of data types and measures, including experimental data, functional annotation, distance measures, protein-protein interaction data and gene-regulatory relationships. Indeed, data of multiple types has been used to perform integrative analyses, where network-based analysis approaches are frequently employed. One major drawback of many current network-based studies, including those described presently, is the static nature of the networks. Typically a single network is used to express the possible interactions being made in a given biological system, when in reality all of these interactions may not be possible, dependent on aspects such as cell-types, environmental factors and temporal aspects of biological processes. For example, HHPID interactions may be cell-type specific or specific to a certain

disease stage, though currently only by manual inspection of the source articles can these aspects be identified. Hence, studies such as that in chapter 6 and studies carried out by other research groups (e.g., [410, 411]) have tended to treat interactions equally, filtering only on the available interaction type description provided by the HHPID. Ideally, network analyses, whether for host-virus interplay, or for single organisms should be more dynamic, enabling relevant temporal, spatial and physiological differences between the states of biological systems to be discerned. However, we have far from exhausted our search for human protein-protein interactions in the simplest binary format (i.e., protein A interacts with protein B) [412], let alone annotated the interactions with the necessary meta-information that would be required to produce specific subsets based on their spatial or temporal activity.

Indeed, in order to successfully merge any data obtained from multiple separate sources (such as public databases), certain variables may be omitted, and simplifications and compromises are inevitably made. Therefore the problem of data-merging is not limited to network-based analyses but to large-scale data integration procedures in general. One method for circumventing this problem is to carry out multiple assays simultaneously. For example, to study HCV replication Blackham *et al.* [113] perform coordinated expression profiling and siRNA screens so that their data sets are complementary. However, this mode of study provokes production of redundant results as it requires that experiments are repeated for each new study. An alternative solution is that newly published biological data is consistently rich in meta-data, such that a wide range of specific details may be computationally retrieved and used for annotating and filtering. This is not a new idea, projects such as Minimum Information About a Microarray Experiment (MIAME) publish specifications for microarray data whilst repositories such as Gene Expression Omnibus (GEO) [87] store MIAME compliant microarray data from a wide variety of sources in a systematic format that allows user selection based on a range of biological and technical criteria. In addition biological pathway data may also come in standardised formats, such as BioPax [413] that allow a wide variety of meta-data to be linked to events. Furthermore, projects such as BioMart [414] provide easy access to current mappings between biological entities and meta-data including as protein, transcript, genes, biological annotation, allowing cross-data-type queries to be handled with relative ease and importantly, with accuracy.

Resources that provide subject-specific insight from a range of large-scale data types are also possible. For example, GPS-Prot has effectively combined different large-scale HIV-1 data sources, including siRNA screen data and protein interaction data, into a single integrated research platform, from which additional custom information may be overlaid and results visualised using a network approach [415]. However, it is yet to be proven whether GPS-Prot will provide a continually up-to date, sufficiently flexible and in-depth resource such that detailed research projects make significant use of this software platform. Regardless, GPS-Prot provides an effective overview of HIV-1 host interaction,

insight into the data types that are available in this field of research and also a window into how the various data sets overlap and integrate to form a more detailed view of HIV-1 infection.

There are some data types and related analysis perspectives that are prominent in computational biology research that have not been significantly exploited in this thesis. Although protein sequence data was briefly used to biologically validate results from bi-clustering analysis (chapter 4), we have not made extensive use of this data type. Sequence data is a hugely valuable resource for biological research to study evolution, particularly as, owing to rapid advance in technology, genomic sequencing has become much cheaper and quicker to perform. Sequence information has been used extensively to study virus evolution and the applications of sequence analysis are too numerous to mention but include identifying the origin of virus strains through reconstruction of phylogenies (e.g., [416]), identification of drug resistance mutations using a deep sequencing approach (e.g., [417]) and studying sequence diversity in relation to protein structural constraints (e.g., [418]). Evolutionary information is certainly appropriate for use in host-virus interaction studies. For example, Sauter *et al.* [419] show that pandemic HIV-1 may be reliant on the evolution of new virus-host interaction. The SIV ancestor of HIV-1 antagonises chimpanzee antiviral tetherin through interaction with the viral Nef protein. However, Sauter *et al.* identify that an evolutionary switch took place whereby tetherin antagonism was effected by viral Vpu, a mechanism that could have promoted the spread of pandemic group M HIV-1. Their study demonstrates that evolutionary studies of host-virus interaction have considerable potential particularly for linking these interactions to viral phenotypes. Thus, it would be both interesting and valuable to ascertain how the wider interaction networks of viruses, like HIV-1 and HCV, have evolved.

Because the scientific literature base is continually growing, text mining, though not directly employed in our methods, is also of increasing importance in biological research. For example, the HIV-1, Human Protein Interaction Database (HHPID) [26, 27] was manually curated from literature where over 100 000 journal articles were screened [27]. However, this feat was extremely time-consuming (the HHPID took 7 years to compile) and inevitably expensive, furthermore (and probably relatedly), no significant database updates with new data have been made since the initial publication. If an accurate text mining procedure were developed to capture host-virus interactions from primary literature, the process of compiling these interactions could be automated. Furthermore, the primary articles that are currently linked by the HHPID to specific protein interactions offer a ready-made curated set on which to train such a text mining method. Such a procedure could be tailored to identify interactions from viruses other than HIV-1 in order to developed wholly new virus-host interaction networks. Indeed, such an initiative has begun in our research group, though the project is in its infancy.

In our study of host-virus interaction, we frequently use interaction network models and systematic functional annotation. Indeed, combining these sorts of data is not an un-

common scheme of analysis for understanding host-virus systems, either in the case of HIV-1 (e.g., [298, 145, 156, 410, 415]), or HCV (e.g., [126, 318]). Studies such as these rely not only on the strength of the virus-host interaction data, but on the strength of the network model for the host cell. The study we presented in chapter 7 is the most thorough investigation of which we are aware into the relationship between function and large-scale interaction data. Perhaps the most important finding from this study was that certain aspects of biological function transcend multiple types of network data. Therefore, improving our cellular network models, and especially developing more refined and accessible integrated models of cellular function, are areas of computational biology that warrant further research. This should include the development of more refined, yet accessible integrated network models of the cell and continued development accurate, large-scale assays to explore the specific molecular characteristics of cellular factors. Such research will not only drive our understanding of host-cellular function but a wider understanding human diseases, including pathogenic viral infections.

In summary, the work presented here has utilised computational analyses of large-scale biological data to provide insight into viral infection of host cells by both HCV and HIV-1 viruses. In particular, the presented research has utilised interaction networks as a basis for data integration and bespoke data models to test hypotheses, in doing so we have developed new software and pioneered novel methodologies for data analysis. Ultimately, a major aim of biomedical research into pathogenic viruses is to develop preventative measures and effective treatment methods. In our work we hope to have contributed to this field by both providing broad insight into virus molecular biology and by highlighting biological subsystems, cellular functions and even specific host factors that may be exploited in the search for new antiviral strategies.

BIBLIOGRAPHY

[1] Koshland D: **The seven pillars of life**. *Science* 2002, **295**(5563):2215–2216.

[2] Breitbart M, Rohwer F: **Here a virus, there a virus, everywhere the same virus?** *Trends in Microbiology* 2005, **13**(6):278–284.

[3] Edwards R, Rohwer F: **Viral metagenomics**. *Nature Reviews Microbiology* 2005, **3**(6):504–510.

[4] Koonin E, Senkevich T, Dolja V: **The ancient Virus World and evolution of cells**. *Biology Direct* 2006, **1**:1–29.

[5] Kay A, Zoulim F: **Hepatitis B virus genetic variability and evolution**. *Virus Research* 2007, **127**(2):164–176.

[6] Claverie J, Ogata H, Audic S, Abergel C, Suhre K, Fournier P: **Mimivirus and the emerging concept of "giant" virus**. *Virus Research* 2006, **117**:133–144.

[7] Baltimore D: **Expression of animal virus genomes**. *Microbiology and Molecular Biology Reviews* 1971, **35**(3):235–241.

[8] Campbell N, Reece J: **Biology. 6th** 2005.

[9] Jones R, Nelson M: **The role of receptors in the HIV-1 entry process**. *European Journal of Medical Research* 2007, **12**(9):391–396.

[10] von Hahn T, Rice C: **Hepatitis C Virus Entry**. *Journal of biological chemistry* 2008, **283**(7):3689–3693.

[11] Permanyer M, Ballana E, Esté J: **Endocytosis of HIV: anything goes**. *Trends in Microbiology* 2010, **18**(12):543–551.

[12] Schelhaas M: **Come in and take your coat off into how host cells provide endocytosis for virus entry**. *Cellular Microbiology* 2010, **12**(10):1378–1388.

[13] Arias C, Escalera-Zamudio M, et al.: **Molecular anatomy of 2009 influenza virus A (H1N1)**. *Archives of Medical Research* 2009, **40**(8):643–654.

[14] Boivin S, Cusack S, Ruigrok R, Hart D: **Influenza A virus polymerase: Structural insights into replication and host adaptation mechanisms**. *Journal of Biological Chemistry* 2010, **285**(37):28411–28417.

[15] Sarafianos S, Marchand B, Das K, Himmel D, Parniak M, Hughes S, Arnold E: **Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition**. *Journal of Molecular Biology* 2009, **385**(3):693–713.

[16] Morton V, Stockley P, Stonehouse N, Ashcroft A: **Insights into virus capsid assembly from non-covalent mass spectrometry**. *Mass Spectrometry Reviews* 2008, **27**(6):575–595.

[17] Chazal N, Gerlier D: **Virus entry, assembly, budding, and membrane rafts**. *Microbiology and Molecular Biology Reviews* 2003, **67**(2):226–237.

[18] Takada K, Ono Y: **Synchronous and sequential activation of latently infected Epstein-Barr virus genomes**. *Journal of Virology* 1989, **63**:445–449.

[19] Young L, Rickinson A: **Epstein into Barr virus: 40 years on**. *Nature Reviews Cancer* 2004, **4**(10):757–768.

[20] Campbell-Yesufu O, Gandhi R: **Update on Human Immunodeficiency Virus (HIV)-2 Infection**. *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America* 2011, **52**(6):780–787.

[21] Swanson C, Malim M: **SnapShot: HIV-1 proteins**. *Cell* 2008, **133**(4):742.

[22] Ganser-Pornillos B, Yeager M, Sundquist W: **The structural biology of HIV assembly**. *Current Opinion in Structural Biology* 2008, **18**(2):203–217.

[23] Bukrinskaya A: **HIV-1 matrix protein: a mysterious regulator of the viral life cycle**. *Virus Research* 2007, **124**(1-2):1–11.

[24] Zhang J, Sharma P, Crumpacker C: **Enhancement of the basal-level activity of HIV-1 long terminal repeat by HIV-1 nucleocapsid protein**. *Virology* 2000, **268**(2):251–263.

[25] Brady J, Kashanchi F: **Tat gets the" green" light on transcription initiation**. *Retrovirology* 2005, **2**:69.

[26] Ptak R, Fu W, Sanders-Beer B, Dickerson J, Pinney J, Robertson D, Rozanov M, Katz K, Maglott D, Pruitt K, et al.: **Cataloguing the HIV Type 1 Human Protein Interaction Network**. *AIDS Research and Human Retroviruses* 2008, **24**(12):1497–1502.

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

[28] Rautonen N, Rautonen J, Martin N, Wara D: **HIV-1 Tat induces cytokine synthesis by uninfected mononuclear cells**. *AIDS* 1994, **8**(10):1504–1506.

[29] Westendorp M, Li-Weber M, Frank R, Krammer P: **Human immunodeficiency virus type 1 Tat upregulates interleukin-2 secretion in activated T cells.** *Journal of Virology* 1994, **68**(7):4177–4185.

[30] Hidalgo-Estévez A, González E, Punzón C, Fresno M: **Human immunodeficiency virus type 1 Tat increases cooperation between AP-1 and NFAT transcription factors in T cells**. *Journal of General Virology* 2006, **87**(6):1603–1612.

[31] Li J, Yim H, Lau A: **Role of HIV-1 Tat in AIDS pathogenesis: its effects on cytokine dysregulation and contributions to the pathogenesis of opportunistic infection**. *AIDS* 2010, **24**(11):1609–1623.

[32] Mischiati C, Pironi F, Milani D, Giacca M, Mirandola P, Capitani S, Zauli G: **Extracellular HIV-1 Tat protein differentially activates the JNK and ERK/MAPK pathways in CD4 T cells**. *AIDS* 1999, **13**(13):1637–1645.

[33] Patrizio M, Colucci M, Levi G: **Human immunodeficiency virus type 1 Tat protein decreases cyclic AMP synthesis in rat microglia cultures**. *Journal of Neurochemistry* 2001, **77**(2):399–407.

[34] Ariumi Y, Kaida A, Hatanaka M, Shimotohno K: **Functional cross-talk of HIV-1 Tat with p53 through its C-terminal domain**. *Biochemical and Biophysical Research Communications* 2001, **287**(2):556–561.

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

[36] Pollard V, Malim M: **The HIV-1 rev protein**. *Annual Reviews in Microbiology* 1998, **52**:491–532.

[37] Neville M, Stutz F, Lee L, Davis L, Rosbash M: **The importin-beta family member Crm1p bridges the interaction between Rev and the nuclear pore complex during nuclear export**. *Current Biology* 1997, **7**(10):767–775.

[38] Truant R, Cullen B: **The arginine-rich domains present in human immunodeficiency virus type 1 Tat and Rev function as direct importin β-dependent**

nuclear localization signals. *Molecular and Cellular Biology* 1999, **19**(2):1210–1217.

[39] Malim M, Emerman M: **HIV-1 accessory proteins into ensuring viral survival in a hostile environment**. *Cell Host & Microbe* 2008, **3**(6):388–398.

[40] Andersen J, Le Rouzic E, Planelles V: **HIV-1 Vpr: Mechanisms of G2 arrest and apoptosis**. *Experimental and Molecular Pathology* 2008, **85**:2–10.

[41] Paxton W, Connor R, Landau N: **Incorporation of Vpr into human immunodeficiency virus type 1 virions: requirement for the p6 region of gag and mutational analysis**. *Journal of Virology* 1993, **67**(12):7229–7237.

[42] Picker L: **Immunopathogenesis of acute AIDS virus infection**. *Current Opinion in Immunology* 2006, **18**(4):399–405.

[43] Schacker T, Collier A, Hughes J, Shea T, Corey L: **Clinical and epidemiologic features of primary HIV infection**. *Annals of Internal Medicine* 1996, **125**(4):257–264.

[44] Grossman Z, Meier-Schellersheim M, Paul W, Picker L: **Pathogenesis of HIV infection: what the virus spares is as important as what it destroys**. *Nature Medicine* 2006, **12**(3):289–295.

[45] Pandrea I, Apetrei C: **Where the wild things are: Pathogenesis of SIV infection in African nonhuman primate hosts**. *Current HIV/AIDS Reports* 2010, **7**:28–36.

[46] Silvestri G: **AIDS pathogenesis: a tale of two monkeys**. *Journal of Medical Primatology* 2008, **37**:6–12.

[47] Mickaël P, Jean-François D, Patricia S, Ivona P, Ousmane D, Cécile B, Michaela M: **Distinct expression profiles of TGF-β1 signaling mediators in pathogenic SIVmac and non-pathogenic SIVagm infections**. *Retrovirology* 2006, **3**:37.

[48] Lederer S, Favre D, Walters K, Proll S, Kanwar B, Kasakow Z, Baskin C, Palermo R, McCune J, Katze M: **Transcriptional profiling in pathogenic and non-pathogenic SIV infections reveals significant distinctions in kinetics and tissue compartmentalization**. *PLoS Pathogens* 2009, **5**(2):e1000296.

[49] Bosinger S, Li Q, Gordon S, Klatt N, Duan L, Xu L, Francella N, Sidahmed A, Smith A, Cramer E, et al.: **Global genomic analysis reveals rapid control of a robust innate response in SIV-infected sooty mangabeys**. *The Journal of clinical investigation* 2009, **119**(12):3556–3572.

[50] **UNAIDS Report on the Global AIDS Epidemic (2010), www.unAIDS.org/en/**. *Joint United Nations Program on HIV/AIDS* 2010.

[51] **AVERT, averting HIV and AIDS, www.avert.org/**.

[52] Haynes B, Liao H, Tomaras G: **Is developing an HIV-1 vaccine possible?** *Current Opinion in HIV and AIDS* 2010, **5**(5):362–367.

[53] Ghosh R, Ghosh S, Chawla S: **Recent advances in antiretroviral drugs**. *Expert Opinion on Pharmacotherapy* 2011, (0):1–16.

[54] Yeni P: **Update on HAART in HIV**. *Journal of Hepatology* 2006, **44**:S100–S103.

[55] Noorali S, Pace D, Bagasra O: **Of lives and livers: emerging responses to the hepatitis C virus**. *The Journal of Infection in Developing Countries* 2010, **5**(01):1–17.

[56] Chisari F: **Unscrambling hepatitis C virus into host interactions**. *Nature* 2005, **436**(7053):930–932.

[57] Chevaliez S, Pawlotsky J: **HCV genome and life cycle**. *Hepatitis C Viruses: Genomes and Molecular Biology* 2006, **1**:5–47.

[58] Sharma S: **Hepatitis C virus: molecular biology & current therapeutic options**. *Indian Journal of Medical Research* 2010, **131**:17–34.

[59] Gottwein J, Bukh J: **Cutting the gordian knot-development and biological relevance of hepatitis C virus cell culture systems**. *Advances in Virus Research* 2008, **71**:51–133.

[60] Popescu C, Dubuisson J: **Role of lipid metabolism in hepatitis C virus assembly and entry**. *Biology of the Cell* 2010, **102**:63–74.

[61] Moradpour D, Penin F, Rice C: **Replication of hepatitis C virus**. *Nature Reviews Microbiology* 2007, **5**(6):453–463.

[62] Takeuchi O, Akira S: **MDA5/RIG-I and virus recognition**. *Current Opinion in Immunology* 2008, **20**:17–22.

[63] Vassilaki N, Mavromara P: **The HCV ARFP/F/core+ 1 protein: production and functional analysis of an unconventional viral product**. *IUBMB life* 2009, **61**(7):739–752.

[64] Shao S, Wu W, Bian Z, Yu J, Zhao P, Zhao L, Zhu S, Qi Z: **Hepatitis C virus F protein inhibits cell apoptosis by activation of intracellular NF-κB pathway**. *Hepatology Research* 2009, **39**(3):282–289.

[65] Tang H, Grisé H: **Cellular and molecular biology of HCV infection and hepatitis**. *Clinical Science* 2009, **117**:49–65.

[66] **Hepatitis C (WHO/CDS/CSR/LYO/2003)**. *World Health Organisation, Geneva.* 2002, :1–69.

[67] Negro F, Alberti A: **The global health burden of hepatitis C virus infection**. *Liver International* 2011, **31**:1–3.

[68] Frank C, Mohamed M, Strickland G, Lavanchy D, Arthur R, Magder L, Khoby T, Abdel-Wahab Y, Ohn E, Anwar W, et al.: **The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt**. *The Lancet* 2000, **355**(9207):887–891.

[69] Cornberg M, Razavi H, Alberti A, Bernasconi E, Buti M, Cooper C, Dalgard O, Dillion J, Flisiak R, Forns X, et al.: **A systematic review of hepatitis C virus epidemiology in Europe, Canada and Israel**. *Liver International* 2011, **31**:30–60.

[70] Munir S, Saleem S, Idrees M, Tariq A, Butt S, Rauff B, Hussain A, Badar S, Naudhani M, Fatima Z, et al.: **Hepatitis C Treatment: current and future perspectives**. *Virology Journal* 2010, **7**:296.

[71] Pawlotsky J: **Therapy of hepatitis C: from empiricism to eradication**. *Hepatology* 2006, **43**(S1):S207–S220.

[72] Lemon S, McKeating J, Pietschmann T, Frick D, Glenn J, Tellinghuisen T, Symons J, Furman P: **Development of novel therapies for hepatitis C**. *Antiviral Research* 2010, **86**:79–92.

[73] Tang H, Peng T, Wong-Staal F: **Novel technologies for studying virus-host interaction and discovering new drug targets for HCV and HIV**. *Current Opinion in Pharmacology* 2002, **2**(5):541–547.

[74] Khattab M: **Targeting host factors: A novel rationale for the management of hepatitis C virus**. *World Journal of Gastroenterology* 2009, **15**(28):3472–3479.

[75] Arhel N, Kirchhoff F: **Host proteins involved in HIV infection: New therapeutic targets**. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 2010, **1802**(3):313–321.

[76] Cherry S: **What have RNAi screens taught us about viral-host interactions?** *Current Opinion in Microbiology* 2009, **12**(4):446–452.

[77] Dorr P, Westby M, Dobbs S, Griffin P, Irvine B, Macartney M, Mori J, Rickett G, Smith-Burchnell C, Napier C, et al.: **Maraviroc (UK-427,857), a Potent, Orally Bioavailable, and Selective Small-Molecule Inhibitor of Chemokine Receptor CCR5 with Broad-Spectrum Anti-Human Immunodeficiency Virus Type 1 Activity**. *Antimicrobial Agents and Chemotherapy* 2005, **49**(11):4721–4732.

[78] Archer J, Braverman M, Taillon B, Desany B, James I, Harrigan P, Lewis M, Robertson D: **Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4-using HIV-1 with ultra-deep pyrosequencing**. *AIDS* 2009, **23**(10):1209–1218.

[79] Coelmont L, Kaptein S, Paeshuyse J, Vliegen I, Dumont J, Vuagniaux G, Neyts J: **Debio 025, a cyclophilin binding molecule, is highly efficient in clearing hepatitis C virus (HCV) replicon-containing cells when used alone or in combination with specifically targeted antiviral therapy for HCV (STAT-C) inhibitors**. *Antimicrobial Agents and Chemotherapy* 2009, **53**(3):967–976.

[80] Young D, Connelly C, Grohmann C, Deiters A: **Small Molecule Modifiers of MicroRNA miR-122 Function for the Treatment of Hepatitis C Virus Infection and Hepatocellular Carcinoma**. *Journal of the American Chemical Society* 2010, **132**(23):7976–7981.

[81] Jangra R, Yi M, Lemon S: **Regulation of hepatitis C virus translation and infectious virus production by the microRNA miR-122**. *Journal of Virology* 2010, **84**(13):6615–6625.

[82] Jares P: **DNA microarray applications in functional genomics**. *Ultrastructural Pathology* 2006, **30**(3):209–219.

[83] Stark C, Breitkreutz B, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Research* 2006, **34**(Database Issue):D535–D539.

[84] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, et al.: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000, **25**:25–29.

[85] Cullum R, Alder O, Hoodless P: **The next generation: Using new sequencing technologies to analyse gene regulation**. *Respirology* 2011, **16**(2):210–222.

[86] Overbergh L, Giulietti A, Valckx D, Decallonne B, Bouillon R, Mathieu C: **The use of real-time reverse transcriptase PCR for the quantification of cytokine gene expression**. *Journal of Biomolecular Techniques: JBT* 2003, **14**:33–43.

[87] Edgar R, Domrachev M, Lash A: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002, **30**:207–210.

[88] Thompson M, Furtado L: **High density oligonucleotide and DNA probe arrays for the analysis of target DNA**. *Analyst* 1999, **124**(8):1133–1136.

[89] Heller M: **DNA microarray technology: devices, systems, and applications**. *Annual Review of Biomedical Engineering* 2002, **4**:129–153.

[90] Rosa G, de Leon N, Rosa A: **Review of microarray experimental design strategies for genetical genomics studies**. *Physiological Genomics* 2006, **28**:15–23.

[91] Liu G, Loraine A, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose M: **NetAffx: Affymetrix probesets and annotations**. *Nucleic Acids Research* 2003, **31**:82–86.

[92] Dai M, Wang P, Boyd A, Kostov G, Athey B, Jones E, Bunney W, Myers R, Speed T, Akil H, et al.: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Research* 2005, **33**(20):e175.

[93] Lu X, Zhang X: **The effect of GeneChip gene definitions on the microarray study of cancers**. *Bioessays* 2006, **28**(7):739–746.

[94] Sandberg R, Larsson O: **Improved precision and accuracy for microarrays using updated probe set definitions**. *BMC Bioinformatics* 2007, **8**:48.

[95] Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249–264.

[96] Kauffmann A, Huber W: **Microarray data quality control improves the detection of differentially expressed genes**. *Genomics* 2010, **95**(3):138–142.

[97] Kauffmann A, Gentleman R, Huber W: **arrayQualityMetricsÑa bioconductor package for quality assessment of microarray data**. *Bioinformatics* 2009, **25**(3):415–416.

[98] Quackenbush J: **Computational analysis of microarray data**. *Nature Reviews Genetics* 2001, **2**(6):418–427.

[99] Irizarry R, Bolstad B, Collin F, Cope L, Hobbs B, Speed T: **Summaries of Affymetrix GeneChip probe level data**. *Nucleic Acids Research* 2003, **31**(4):e15.

[100] Bolstad B, Irizarry R, Astrand M, Speed T: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185–193.

[101] Rao Y, Lee Y, Jarjoura D, Ruppert A, Liu C, Hsu J, Hagan J: **A comparison of normalization techniques for microRNA microarray data**. *Statistical applications in genetics and molecular biology* 2008, **7**.

[102] Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, Pawitan Y: **Filtering genes to improve sensitivity in oligonucleotide microarray data analysis**. *Nucleic Acids Research* 2007, **35**(16):e102.

[103] Hackstadt A, Hess A: **Filtering for increased power for microarray data analysis**. *BMC Bioinformatics* 2009, **10**.

[104] McClintick J, Edenberg H: **Effects of filtering by Present call on analysis of microarray experiments**. *BMC Bioinformatics* 2006, **7**.

[105] Smyth G: **Linear models and empirical Bayes methods for assessing differential expression in microarray experiments**. *Statistical applications in genetics and molecular biology* 2004, **3**:1027.

[106] Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, Guedj M: **Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies**. *PloS One* 2010, **5**(9):e12336.

[107] Meyer P, Kontos K, Lafitte F, Bontempi G: **Information-theoretic inference of large transcriptional regulatory networks**. *EURASIP Journal on Bioinformatics and Systems Biology* 2007, **2007**:79879.

[108] Da Wei Huang B, Lempicki R: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature Protocols* 2008, **4**:44–57.

[109] Ideker T, Ozier O, Schwikowski B, Siegel A: **Discovering regulatory and signalling circuits in molecular interaction networks**. *Bioinformatics* 2002, **18**(suppl 1):S233.

[110] Giri M, Nebozhyn M, Showe L, Montaner L: **Microarray data on gene modulation by HIV-1 in immune cells: 2000 into 2006**. *Journal of Leukocyte Biology* 2006, **80**(5):1031–1043.

[111] Bigger C, Brasky K, Lanford R: **DNA microarray analysis of chimpanzee liver during acute resolving hepatitis C virus infection**. *Journal of Virology* 2001, **75**(15):7059–7066.

[112] Basu A, Meyer K, Lai K, Saito K, Di Bisceglie A, Grosso L, Ray R, Ray R: **Microarray analyses and molecular profiling of Stat3 signaling pathway induced by hepatitis C virus core protein in human hepatocytes**. *Virology* 2006, **349**(2):347–358.

[113] Blackham S, Baillie A, Al-Hababi F, Remlinger K, You S, Hamatake R, McGarvey M: **Gene Expression Profiling Indicates the Roles of Host Oxidative Stress, Apoptosis, Lipid Metabolism, and Intracellular Transport Genes in the Replication of Hepatitis C Virus**. *Journal of Virology* 2010, **84**(10):5404–5414.

[114] de Veer M, Holko M, Frevel M, Walker E, Der S, Paranjape J, Silverman R, Williams B: **Functional classification of interferon-stimulated genes identified using microarrays**. *Journal of Leukocyte Biology* 2001, **69**(6):912–920.

[115] Izmailova E, Bertley F, Huang Q, Makori N, Miller C, Young R, Aldovini A: **HIV-1 Tat reprograms immature dendritic cells to express chemoattractants for activated T cells and macrophages**. *Nature Medicine* 2003, **9**(2):191–197.

[116] Bansal M, Belcastro V, Ambesi-Impiombato A, Di Bernardo D: **How to infer gene networks from expression profiles**. *Molecular Systems Biology* 2007, **3**:78.

[117] Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data**. *Journal of Computational Biology* 2000, **7**(3-4):601–620.

[118] Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context**. *BMC Bioinformatics* 2006, **7**(Suppl 1):S7.

[119] Meyer P, Lafitte F, Bontempi G: **minet: A r/bioconductor package for inferring large transcriptional networks using mutual information**. *BMC Bioinformatics* 2008, **9**:461.

[120] Faith J, Hayete B, Thaden J, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins J, Gardner T: **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles**. *PLoS Biology* 2007, **5**:e8.

[121] Butte A, Kohane I: **Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements**. In *Pacific Symposium on Biocomputing*, *Volume 5*, Citeseer 2000:418–429.

[122] Cover T, Thomas J: **Elements of Information Theory. Series in Telecommunications** 1991.

[123] Altay G, Emmert-Streib F: **Revealing differences in gene network inference algorithms on the network level by ensemble methods**. *Bioinformatics* 2010, **26**(14):1738–1744.

[124] von Mering C, Krause R, Snel B, Cornell M, Oliver S, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein into protein interactions**. *Nature* 2002, **1**:399–403.

[125] Panchenko A, Przytycka T: **Protein-protein Interaction Networks, Identification, Computer Analysis and Prediction**. *Computational Biology* 2008.

[126] De Chassey B, Navratil V, Tafforeau L, Hiet M, Aublin-Gex A, Agaugue S, Meiffren G, Pradezynski F, Faria B, Chantier T, et al.: **Hepatitis C virus infection protein network**. *Molecular Systems Biology* 2008, **4**:230.

[127] Gong S, Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, Choi H, Cho Y, Han K, Lee S, et al.: **PSIbase: a database of Protein Structural Interactome map (PSIMAP)**. *Bioinformatics* 2005, **21**(10):2541–2543.

[128] Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure1**. *Journal of Molecular Biology* 2001, **313**(4):903–919.

[129] Bhardwaj N, Lu H: **Correlation between gene expression profiles and protein into protein interactions within and across genomes**. *Bioinformatics* 2005, **21**(11):2730–2738.

[130] Mohamed T, Carbonell J, Ganapathiraju M: **Active learning for human protein-protein interaction prediction**. *BMC Bioinformatics* 2010, **11**(Suppl 1):S57.

[131] DYER M, MURALI T, SOBRAL B: **Supervised learning and prediction of physical interactions between human and HIV proteins**. *Infection, genetics and evolution* 2011, **11**(5):917–923.

[132] Bader G, Betel D, Hogue C: **BIND: the biomolecular interaction network database**. *Nucleic Acids Research* 2003, **31**:248–250.

[133] Chatr-Aryamontri A, Ceol A, Palazzi L, Nardelli G, Schneider M, Castagnoli L, Cesareni G: **MINT: the Molecular INTeraction database**. *Nucleic Acids Research* 2006, **35**(suppl 1):D572–D574.

[134] Stark C, Breitkreutz B, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets**. *Nucleic Acids Research* 2006, **34**(Database issue):D535–D539.

[135] Marte B: **Cancer: Super p53**. *Nature* 2002, **420**(6913):279.

[136] Dickerson J, Pinney J, Robertson D: **The biological context of HIV-1 host interactions reveals subtle insights into a system hijack**. *BMC Systems Biology* 2010, **4**:1–13.

[137] Dennis Jr G, Sherman B, Hosack D, Yang J, Gao W, Lane H, Lempicki R: **DAVID: database for annotation, visualization, and integrated discovery**. *Genome Biology* 2003, **4**(5):P3.

[138] Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub T, Lander E, et al.: **Gene set enrichment analysis: a**

knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 2005, **102**(43):15545–15550.

[139] Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady Y, Müller R, Meese E, Lenhof H: **GeneTrail – advanced gene set enrichment analysis**. *Nucleic Acids Research* 2007, **35**(suppl 2):W186–W192.

[140] Barrell D, Dimmer E, Huntley R, Binns D, O'Donovan C, Apweiler R: **The GOA database in 2009 into an integrated Gene Ontology Annotation resource**. *Nucleic Acids Research* 2009, **37**(Database issue):D396–D403.

[141] Fire A, Albertson D, Harrison S, Moerman D: **Production of antisense RNA leads to effective and specific inhibition of gene expression in C. elegans muscle**. *Development* 1991, **113**(2):503–514.

[142] Hannon G, Rossi J: **Unlocking the potential of the human genome with RNA interference**. *Nature* 2004, **431**(7006):371–378.

[143] Prudêncio M, Lehmann M: **Illuminating the host into How RNAi screens shed light on host-pathogen interactions**. *Biotechnology Journal* 2009, **4**(6):826–837.

[144] Brass A, Dykxhoorn D, Benita Y, Yan N, Engelman A, Xavier R, Lieberman J, Elledge S: **Identification of host proteins required for HIV infection through a functional genomic screen**. *Science* 2008, **319**(5865):921–926.

[145] König R, Zhou Y, Elleder D, Diamond T, Bonamy G, Irelan J, Chiang C, Tu B, De Jesus P, Lilley C, et al.: **Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication**. *Cell* 2008, **135**:49–60.

[146] Zhou H, Xu M, Huang Q, Gates A, Zhang X, Castle J, Stec E, Ferrer M, Strulovici B, Hazuda D, et al.: **Genome-scale RNAi screen for host factors required for HIV replication**. *Cell Host & Microbe* 2008, **4**(5):495–504.

[147] Hao L, Sakurai A, Watanabe T, Sorensen E, Nidom C, Newton M, Ahlquist P, Kawaoka Y: **Drosophila RNAi screen identifies host genes important for influenza virus replication**. *Nature* 2008, **454**(7206):890–893.

[148] König R, Stertz S, Zhou Y, Inoue A, Hoffmann H, Bhattacharyya S, Alamares J, Tscherne D, Ortigoza M, Liang Y, et al.: **Human host factors required for influenza virus replication**. *Nature* 2009, **463**(7282):813–817.

[149] Karlas A, Machuy N, Shin Y, Pleissner K, Artarini A, Heuer D, Becker D, Khalil H, Ogilvie L, Hess S, et al.: **Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication**. *Nature* 2010, **463**(7282):818–822.

[150] Ng T, Mo H, Pilot-Matias T, He Y, Koev G, Krishnan P, Mondal R, Pithawalla R, He W, Dekhtyar T, et al.: **Identification of host genes involved in hepatitis C virus replication by small interfering RNA technology**. *Hepatology* 2007, **45**(6):1413–1421.

[151] Randall G, Panis M, Cooper J, Tellinghuisen T, Sukhodolets K, Pfeffer S, Landthaler M, Landgraf P, Kan S, Lindenbach B, et al.: **Cellular cofactors affecting hepatitis C virus infection and replication**. *Proceedings of the National Academy of Sciences* 2007, **104**(31):12884–12889.

[152] Supekova L, Supek F, Lee J, Chen S, Gray N, Pezacki J, Schlapbach A, Schultz P: **Identification of human kinases involved in hepatitis C virus replication by small interference RNA library screening**. *Journal of Biological Chemistry* 2008, **283**:29–36.

[153] Tai A, Benita Y, Peng L, Kim S, Sakamoto N, Xavier R, Chung R: **A functional genomic screen identifies cellular cofactors of hepatitis C virus replication**. *Cell Host & Microbe* 2009, **5**(3):298–307.

[154] Li Q, Brass A, Ng A, Hu Z, Xavier R, Liang T, Elledge S: **A genome-wide genetic screen for host factors required for hepatitis C virus propagation**. *Proceedings of the National Academy of Sciences* 2009, **106**(38):16410–16415.

[155] Krishnan M, Ng A, Sukumaran B, Gilfoy F, Uchil P, Sultana H, Brass A, Adametz R, Tsui M, Qian F, et al.: **RNA interference screen for human genes associated with West Nile virus infection**. *Nature* 2008, **455**(7210):242–245.

[156] Bushman F, Malani N, Fernandes J, D'Orso I, Cagney G, Diamond T, Zhou H, Hazuda D, Espeseth A, König R, et al.: **Host Cell Factors in HIV Replication: Meta-Analysis of Genome-Wide Studies**. *PLoS Pathogens* 2009, **5**(5):e1000437.

[157] Huber W, Carey V, Long L, Falcon S, Gentleman R: **Graphs in molecular biology**. *BMC Bioinformatics* 2007, **8**(Suppl 6):S8.

[158] Barabási A, Oltvai Z: **Network biology: understanding the cell's functional organization**. *Nature Reviews Genetics* 2004, **5**(2):101–113.

[159] Kitano H: **Computational Systems Biology**. *Nature* 2002, **420**(6912):206–210.

[160] Barabási A, Oltvai Z: **Network biology: understanding the cell's functional organization**. *Nature Reviews Genetics* 2004, **5**(2):101–113.

[161] Jeong H, Mason S, Barabasi A, Oltvai Z, et al.: **Lethality and centrality in protein networks**. *Nature* 2001, **411**(6833):41–42.

[162] Spirin V, Mirny L: **Protein complexes and functional modules in molecular networks**. *Proceedings of the National Academy of Sciences* 2003, **100**(21):12123–12128.

[163] Huang C, Cheng C, Sun C: **Bridge and brick network motifs: Identifying significant building blocks from complex biological systems**. *Artificial Intelligence In Medicine* 2007, **41**(2):117–127.

[164] Shannon P, Markiel A, Ozier O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks**. *Genome Research* 2003, **13**:2498–2504.

[165] Assenov Y, Ramírez F, Schelhorn S, Lengauer T, Albrecht M: **Computing topological parameters of biological networks**. *Bioinformatics* 2008, **24**(2):282–284.

[166] Yildirim M, Goh K, Cusick M, Barabasi A, Vidal M: **Drug-target network**. *Nature Biotechnology* 2007, **25**(10):1119–1126.

[167] Yeung K, Medvedovic M, Bumgarner R: **Clustering gene-expression data with repeated measurements**. *Genome Biology* 2003, **4**(5):R34.

[168] Aittokallio T, Schwikowski B: **Graph-based methods for analysing networks in cell biology**. *Briefings in Bioinformatics* 2006, **7**(3):243–255.

[169] Dunn R, Dudbridge F, Sanderson C: **The use of edge-betweenness clustering to investigate biological function in protein interaction networks**. *BMC Bioinformatics* 2005, **6**:39.

[170] Dijkstra E: **A note on two problems in connexion with graphs**. *Numerische Mathematik* 1959, **1**:269–271.

[171] Bader G, Hogue C: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**.

[172] Pesquita C, Faria D, Falcão A, Lord P, Couto F: **Semantic similarity in biomedical ontologies**. *PLoS Computational Biology* 2009, **5**(7):e1000443.

[173] Mistry M, Pavlidis P: **Gene ontology term overlap as a measure of gene functional similarity**. *BMC Bioinformatics* 2008, **9**:327.

[174] Pinney J, Dickerson J, Fu W, Sanders-Beer B, Ptak R, Robertson D: **HIV-host interactions: a map of viral perturbation of the host system**. *AIDS* 2009, **23**(5):549–554.

[175] Holden B, Pinney J, Lovell S, Amoutzias G, Robertson D: **An exploration of alternative visualisations of the basic helix-loop-helix protein interaction network**. *BMC bioinformatics* 2007, **8**:289.

[176] Tao Y, Liu Y, Friedman C, Lussier Y: **Information visualization techniques in bioinformatics during the postgenomic era**. *Drug Discovery Today: Biosilico* 2004, **2**(6):237–245.

[177] Kent W, Sugnet C, Furey T, Roskin K, Pringle T, Zahler A, et al.: **The human genome browser at UCSC**. *Genome Research* 2002, **12**(6):996–1006.

[178] Clamp M, Cuff J, Searle S, Barton G: **The jalview java alignment editor**. *Bioinformatics* 2004, **20**(3):426–427.

[179] Rasmussen M, Karypis G: **gcluto: An interactive clustering, visualization, and analysis system**. *CSE/UMN Technical Report: TR* 2008, **4**.

[180] Barkow S, Bleuler S, Prelić A, Zimmermann P, Zitzler E: **BicAT: a biclustering analysis toolbox**. *Bioinformatics* 2006, **22**(10):1282–1283.

[181] Pavlopoulos G, Moschopoulos C, Hooper S, Schneider R, Kossida S: **jClust: a clustering and visualization toolbox**. *Bioinformatics* 2009, **25**(15):1994–1996.

[182] Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes**. *Nucleic Acids Research* 2000, **28**:27–30.

[183] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath G, Wu G, Matthews L, et al.: **Reactome: a knowledgebase of biological pathways**. *Nucleic Acids Research* 2005, **33**(Database Issue):D428.

[184] Suderman M, Hallett M: **Tools for visually exploring biological networks**. *Bioinformatics* 2007, **23**(20):2651–2659.

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

[186] Breitkreutz B, Stark C, Tyers M: **Osprey: a network visualization system**. *Genome Biology* 2003, **4**(3):R22.

[187] Nikitin A, Egorov S, Daraselia N, Mazo I: **Pathway studio-the analysis and navigation of molecular networks**. *Bioinformatics* 2003, **19**(16):2155–2157.

[188] Bader G, Donaldson I, Wolting C, Ouellette B, Pawson T, Hogue C: **BIND into the biomolecular interaction network database**. *Nucleic Acids Research* 2001, **29**:242–250.

[189] Jacquelin B, Mayau V, Targat B, Liovat A, Kunkel D, Petitjean G, Dillies M, Roques P, Butor C, Silvestri G, et al.: **Nonpathogenic SIV infection of African green monkeys induces a strong but rapidly controlled type I IFN response**. *The Journal of Clinical Investigation* 2009, **119**(12):3544–3555.

[190] Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al.: **A comprehensive analysis of protein into protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**:623–627.

[191] Han K, Ju B, Jung H: **WebInterViewer: visualizing and analyzing molecular interaction networks**. *Nucleic Acids Research* 2004, **32**(Web Server Issue):W89–W95.

[192] Klammer M, Roopra S, Sonnhammer E: **jSquid: a Java applet for graphical on-line network exploration**. *Bioinformatics* 2008, **24**(12):1467–1468.

[193] Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society. Series B. Methodological* 1995, **57**:289.

[194] del Real G, Jiménez-Baranda S, Mira E, Lacalle R, Lucas P, Gómez-Moutón C, Alegret M, Peña J, Rodríguez-Zapata M, Alvarez-Mon M, et al.: **Statins Inhibit HIV-1 Infection by Down-regulating Rho Activity**. *The Journal of Experimental Medicine* 2004, **200**(4):541–547.

[195] Nathans R, Cao H, Sharova N, Ali A, Sharkey M, Stranska R, Stevenson M, Rana T: **Small-molecule inhibition of HIV-1 Vif**. *Nature Biotechnology* 2008, **26**(10):1187–1192.

[196] Harris R, Liddament M: **Retroviral restriction by APOBEC proteins.** *Nature Reviews Immunology* 1911, **2004**:868–877.

[197] Harris R, Liddament M, Bieniasz P, Brown W, Schumacher A, Alce T, Popik W, Gaddis N, Sheehy A, Ahmad K, et al.: **Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein**. *Nature* 2002, **418**(6898):646–650.

[198] Wishart D, Knox C, Guo A, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration**. *Nucleic Acids Research* 2006, **34**(Database Issue):D668–D672.

[199] Wishart D, Knox C, Guo A, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets**. *Nucleic Acids Research* 2007, **36**(Database issue):D901–D906.

[200] Laspia M, Rice A, Mathews M: **HIV-1 Tat protein increases transcriptional initiation and stabilizes elongation.** *Cell* 1989, **59**(2):283–292.

[201] Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, Marzio P, Marmon S, Sutton R, Hill C, et al.: **Identification of a major co-receptor for primary isolates of HIV-1**. *Nature* 1996, **381**(6584):661–666.

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

[203] Schwartz O, Marechal V, Gall S, Lemonnier F, Heard J: **Endocytosis of major histocompatibility complex class I molecules is induced by the HIV into 1 Nef protein**. *Nature Medicine* 1996, **2**(3):338–342.

[204] Hopkins A, Groom C: **Target analysis: a priori assessment of druggability.** *Ernst Schering Research Foundation Workshop* 2003, **42**:11–17.

[205] Wang J, Kiyokawa E, Verdin E, Trono D: **The Nef protein of HIV-1 associates with rafts and primes T cells for activation**. *Proceedings of the National Academy of Sciences* 2000, **97**:394–399.

[206] Ling H, Xiao P, Usami O, Hattori T: **Thrombin activates envelope glycoproteins of HIV type 1 and enhances fusion**. *Microbes and Infection* 2004, **6**(5):414–420.

[207] Albanell J, Codony J, Rovira A, Mellado B, Gascon P: **Mechanism of action of anti-HER2 monoclonal antibodies: scientific update on trastuzumab and 2C4.** *Advances in Experimental Medicine and Biology* 2003, **532**:253–268.

[208] Bonnet F, Lewden C, May T, Heripret L, Jougla E, Bevilacqua S, Costagliola D, Salmon D, Chêne G, Morlat P: **Malignancy-Related Causes of Death in Human Immunodeficiency Virus into Infected Patients in the Era of Highly Active Antiretroviral Therapy**. *Infection* 2004, **101**:317–324.

[209] Louie J, Hsu L, Osmond D, Katz M, Schwarcz S: **Trends in causes of death among persons with acquired immunodeficiency syndrome in the era of highly active antiretroviral therapy, San Francisco, 1994-1998.** *The Journal of Infectious Diseases* 2002, **186**(7):1023–1027.

[210] Dyer M, Murali T, Sobral B: **The landscape of human proteins interacting with viruses and other pathogens**. *PLoS Pathogens* 2008, **4**(2):e32.

[211] Rosario B, Hearst M: **Multi-way relation classification: application to protein-protein interactions**. In *Proceedings of the article on Human Language Technology and Empirical Methods in Natural Language Processing*, Association for Computational Linguistics Morristown, NJ, USA 2005:732–739.

[212] Jabado N, Le Deist F, Fischer A, Hivroz C: **Interaction of HIV gp120 and anti-CD4 antibodies with the CD4 molecule on human CD4+ T cells inhibits the binding activity of NF-AT, NF-χB and AP-1, three nuclear factors regulating interleukin-2 gene enhancer activity**. *European Journal of Immunology* 1994, **24**(11):2646–2652.

[213] MacPherson J, Pinney J, Robertson D: **JNets: Exploring networks by integrating annotation**. *BMC Bioinformatics* 2009, **10**:95.

[214] Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data**. *Bioinformatics* 2006, **22**(9):1122–1129.

[215] Peri S, Navarro J, Kristiansen T, Amanchy R, Surendranath V, Muthusamy B, Gandhi T, Chandrika K, Deshpande N, Suresh S, et al.: **Human protein reference database as a discovery resource for proteomics**. *Nucleic Acids Research* 2004, **32**(Database Issue):D497–D501.

[216] Smith T, Waterman M: **Identification of common molecular subsequences**. *Journal of Molecular Biology* 1981, **147**:195–197.

[217] Jiang J, Conrath D: **Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy** 1997, **1**:9008.

[218] Hakes L, Pinney J, Lovell S, Oliver S, Robertson D: **All duplicates are not equal: the difference between small-scale and genome duplication**. *Genome Biology* 2007, **8**(10):R209.

[219] Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees**. *Molecular Biology and Evolution* 1987, **4**(4):406–425.

[220] Dunn B, Goodenow M, Gustchina A, Wlodawer A: **Retroviral proteases**. *Genome Biology* 2002, **3**(4):3006.

[221] Ventoso I, Blanco R, Perales C, Carrasco L: **HIV-1 protease cleaves eukaryotic initiation factor 4G and inhibits cap-dependent translation**. *Proceedings of the National Academy of Sciences* 2001, **98**(23):12966–12971.

[222] Nie Z, Bren G, Vlahakis S, Schimnich A, Brenchley J, Trushin S, Warren S, Schnepple D, Kovacs C, Loutfy M, et al.: **Human Immunodeficiency Virus Type 1 Protease Cleaves Procaspase 8 In Vivo**. *The Journal of Virology* 2007, **81**(13):6947–6956.

[223] Perales C, Carrasco L, Ventoso I: **Cleavage of eIF4G by HIV-1 protease: effects on translation**. *FEBS letters* 2003, **533**:89–94.

[224] Alvarez E, Castelló A, Menéndez-Arias L, Carrasco L: **HIV protease cleaves poly (A)-binding protein**. *Biochemical Journal* 2006, **396**(Pt 2):219–226.

[225] Shaheduzzaman S, Krishnan V, Petrovic A, Bittner M, Meltzer P, Trent J, Venkatesan S, Zeichner S: **Effects of HIV-1 Nef on cellular gene expression profiles**. *Journal of Biomedical Science* 2002, **9**:82–96.

[226] Benson R, Sanfridson A, Ottinger J, Doyle C, Cullen B: **Downregulation of cell-surface CD4 expression by simian immunodeficiency virus Nef prevents viral super infection**. *Journal of Experimental Medicine* 1993, **177**(6):1561–1566.

[227] Alonso A, Derse D, Peterlin B: **Human chromosome 12 is required for optimal interactions between Tat and TAR of human immunodeficiency virus type 1 in rodent cells.** *Journal of Virology* 1992, **66**(7):4617–4621.

[228] Wei P, Garber M, Fang S, Fischer W, Jones K: **A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA**. *Cell* 1998, **92**(4):451–462.

[229] Zhou Q, Chen D, Pierstorff E, Luo K: **Transcription elongation factor P-TEFb mediates Tat activation of HIV-1 transcription at multiple stages**. *The EMBO Journal* 1998, **17**(13):3681–3691.

[230] Pereira-Leal J, Levy E, Teichmann S: **The origins and evolution of functional modules: lessons from protein complexes**. *Philosophical Transactions B* 2006, **361**(1467):507–517.

[231] Cicala C, Arthos J, Selig S, Dennis G, Hosack D, Van Ryk D, Spangler M, Steenbeke T, Khazanie P, Gupta N, et al.: **HIV envelope induces a cascade of cell signals in non-proliferating target cells that favor virus replication**. *Proceedings of the National Academy of Sciences* 2002, **99**(14):9380–9385.

[232] Mellor H, Parker P: **The extended protein kinase C superfamily**. *Biochemical Journal* 1998, **332**(2):281–292.

[233] Gupta S, Aggarwal S, Kim C, Gollapudi S: **Human immunodeficiency virus-1 recombinant gp 120 induces changes in protein kinase C isozymes. A preliminary report**. *International Journal of Immunopharmacology* 1994, **16**(3):197–204.

[234] Moore J, Kitchen S, Pugach P, Zack J: **The CCR5 and CXCR4 coreceptors-central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection**. *AIDS Research and Human Retroviruses* 2004, **20**:111–126.

[235] Simon M: **Receptor tyrosine kinases specific outcomes from general signals**. *Cell* 2000, **103**:13–15.

[236] Gordus A, Krall J, Beyer E, Kaushansky A, Wolf-Yadlin A, Sevecka M, Chang B, Rush J, MacBeath G: **Linear combinations of docking affinities explain quantitative differences in RTK signaling**. *Molecular Systems Biology* 2009, **5**:235.

[237] Mujtaba S, He Y, Zeng L, Farooq A, Carlson J, Ott M, Verdin E, Zhou M: **Structural basis of lysine-acetylated HIV-1 Tat recognition by PCAF bromodomain**. *Molecular Cell* 2002, **9**(3):575–586.

[238] Jeang K, Xiao H, Rich E: **Multifaceted activities of the HIV-1 transactivator of transcription, Tat**. *Journal of Biological Chemistry* 1999, **274**(41):28837–28840.

[239] Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y: **Predicting function: from genes to genomes and back**. *Journal of Molecular Biology* 1998, **283**(4):707–725.

[240] Linial M: **How incorrect annotations evolve-the case of short ORFs**. *Trends in Biotechnology* 2003, **21**(7):298–300.

[241] Gilks W, Audit B, De Angelis D, Tsoka S, Ouzounis C: **Modeling the percolation of annotation errors in a database of protein sequences**. *Bioinformatics* 2002, **18**(12):1641–1649.

[242] Chirmule N, Wang X, Hu R, Oyaizu N, Roifman C, Pahwa R, Kalyanaraman V, Pahwa S: **Envelope glycoproteins of HIV-1 interfere with T-cell-dependent B cell differentiation: role of CD4-MHC class II interaction in the effector phase of T cell help**. *Cellular Immunology* 1994, **155**:169–182.

[243] Chirmule N, McCloskey T, Hu R, Kalyanaraman V, Pahwa S: **HIV gp120 inhibits T cell activation by interfering with expression of costimulatory molecules CD40 ligand and CD80 (B71)**. *The Journal of Immunology* 1995, **155**(2):917–924.

[244] Cefai D, Debre P, Kaczorek M, Idziorek T, Autran B, Bismuth G: **Human immunodeficiency virus-1 glycoproteins gp120 and gp160 specifically inhibit the CD3/T cell-antigen receptor phosphoinositide transduction pathway**. *Journal of Clinical Investigation* 1990, **86**(6):2117–2124.

[245] Hubert P, Bismuth G, Korner M, Debre P: **HIV-1 glycoprotein gp120 disrupts CD4-p56 lck/CD3-T cell receptor interactions and inhibits CD3 signaling**. *European Journal of Immunology* 1995, **25**(5):1417–1425.

[246] Schwartz O, Maréchal V, Le Gall S, Lemonnier F, Heard J: **Endocytosis of major histocompatibility complex class I molecules is induced by the HIV into 1 Nef protein**. *Nature Medicine* 1996, **2**(3):338–342.

[247] Benichou S, Benmerah A: **The HIV nef and the Kaposi-sarcoma-associated virus K3/K5 proteins: parasites of the endocytosis pathway**. *Médecine Sciences: M/S* 2003, **19**:100–106.

[248] Madrid R, Janvier K, Hitchin D, Day J, Coleman S, Noviello C, Bouchet J, Benmerah A, Guatelli J, Benichou S: **Nef-induced alteration of the early/recycling endosomal compartment correlates with enhancement of HIV-1 infectivity**. *Journal of Biological Chemistry* 2005, **280**(6):5032–5044.

[249] Michel N, Allespach I, Venzke S, Fackler O, Keppler O: **The Nef protein of human immunodeficiency virus establishes superinfection immunity by a dual strategy to downregulate cell-surface CCR5 and CD4**. *Current Biology* 2005, **15**(8):714–723.

[250] Venzke S, Michel N, Allespach I, Fackler O, Keppler O: **Expression of Nef downregulates CXCR4, the major coreceptor of human immunodeficiency virus, from the surfaces of target cells and thereby enhances resistance to superinfection**. *The Journal of Virology* 2006, **80**(22):11141–11152.

[251] Chaudhry A, Das S, Jameel S, George A, Bal V, Mayor S, Rath S: **A two-pronged mechanism for HIV-1 Nef-mediated endocytosis of immune costimulatory molecules CD80 and CD86**. *Cell Host & Microbe* 2007, **1**:37–49.

[252] Nel A, Hanekom C, Rheeder A, Williams K, Pollack S, Katz R, Landreth G: **Stimulation of MAP-2 kinase activity in T lymphocytes by anti-CD3 or anti-Ti monoclonal antibody is partially dependent on protein kinase C**. *The Journal of Immunology* 1990, **144**(7):2683–2689.

[253] Nel A, Pollack S, Landreth G, Ledbetter J, Hultin L, Williams K, Katz R, Akerley B: **CD-3-mediated activation of MAP-2 kinase can be modified by ligation of the CD4 receptor. Evidence for tyrosine phosphorylation during activation of this kinase**. *The Journal of Immunology* 1990, **145**(3):971–979.

[254] Nel A, Hanekom C, Hultin L: **Protein kinase C plays a role in the induction of tyrosine phosphorylation of lymphoid microtubule-associated protein-2 kinase. Evidence for a CD3-associated cascade that includes pp56lck and that is defective in HPB-ALL**. *The Journal of Immunology* 1991, **147**(6):1933–1939.

[255] Gupta S, Weiss A, Kumar G, Wang S, Nel A: **The T-cell antigen receptor utilizes Lck, Raf-1, and MEK-1 for activating mitogen-activated protein kinase. Evidence for the existence of a second protein kinase C-dependent pathway**

in an Lck-negative Jurkat cell mutant. *Journal of Biological Chemistry* 1994, **269**(25):17349–17357.

[256] Tian H, Lempicki R, King L, Donoghue E, Samelson L, Cohen D: **HIV envelope-directed signaling aberrancies and cell death of CD4+ T cells in the absence of TCR co-stimulation**. *International immunology* 1996, **8**:65–74.

[257] Phipps D, Read S, Piovesan J, Mills G, Branch D: **HIV infection in vitro enhances the activity of src-family protein tyrosine kinases**. *AIDS* 1996, **10**(11):1191–1198.

[258] Shatrov V, Ratter F, Gruber A, Droge W, Lehmann V: **HIV type 1 glycoprotein 120 amplifies tumor necrosis factor-induced NF-kappa B activation in Jurkat cells.** *AIDS Research and Human Retroviruses* 1996, **12**(13):1209–1216.

[259] Greenway A, Dutartre H, Allen K, McPhee D, Olive D, Collette Y: **Simian immunodeficiency virus and human immunodeficiency virus type 1 nef proteins show distinct patterns and mechanisms of Src kinase activation**. *Journal of Virology* 1999, **73**(7):6152–6158.

[260] Witte V, Laffert B, Gintschel P, Krautkrämer E, Blume K, Fackler O, Baur A: **Induction of HIV transcription by Nef involves Lck activation and protein kinase C theta raft recruitment leading to activation of ERK1/2 but not NF kappa B**. *Journal of Immunology* 2008, **181**(12):8425–8432.

[261] Fackler O, Luo W, Geyer M, Alberts A, Peterlin B: **Activation of Vav by Nef induces cytoskeletal rearrangements and downstream effector functions**. *Molecular Cell* 1999, **3**(6):729–739.

[262] Marina V, Benedetta M, Francesca S, Elisabetta S, Luciana G, Carlo R, Valter M, Giovanna Q: **HIV-1 Nef triggers Vav-mediated signaling pathway leading to functional and morphological differentiation of dendritic cells**. *Immunology Letters* 2003, **87**(1-3):191 into 192.

[263] Varin A, Decrion A, Sabbah E, Quivy V, Sire J, Van Lint C, Roques B, Aggarwal B, Herbein G: **Synthetic Vpr protein activates activator protein-1, c-Jun N-terminal kinase, and NF-{kappa} B and stimulates HIV-1 transcription in promonocytic cells and primary macrophages**. *Journal of Biological Chemistry* 2005, **280**(52):42557–42560.

[264] Pessler F, Cron R: **Reciprocal regulation of the nuclear factor of activated T cells and HIV-1**. *Genes and Immunity* 2004, **5**(3):158–167.

[265] Malmgaard L: **Induction and regulation of IFNs during viral infections**. *Journal of interferon & cytokine research* 2004, **24**(8):439–454.

[266] di Somma M, Majolini M, Burastero S, Telford J, Baldari C: **Cyclosporin A sensitivity of the HIV-1 long terminal repeat identifies distinct p56lck-dependent pathways activated by CD4 triggering**. *European Journal of Immunology* 1996, **26**(9):2181–2188.

[267] Macian F, Rao A: **Reciprocal modulatory interaction between human immunodeficiency virus type 1 Tat and transcription factor NFAT1**. *Molecular and Cellular Biology* 1999, **19**(5):3645–3653.

[268] Lahti A, Manninen A, Saksela K: **Regulation of T cell activation by HIV-1 accessory proteins: Vpr acts via distinct mechanisms to cooperate with Nef in NFAT-directed gene expression and to promote transactivation by CREB**. *Virology* 2003, **310**:190–196.

[269] Fenard D, Yonemoto W, de Noronha C, Cavrois M, Williams S, Greene W: **Nef Is Physically Recruited into the Immunological Synapse and Potentiates T Cell Activation Early after TCR Engagement 1**. *The Journal of Immunology* 2005, **175**(9):6050–6057.

[270] Cicala C, Arthos J, Censoplano N, Cruz C, Chung E, Martinelli E, Lempicki R, Natarajan V, VanRyk D, Daucher M, et al.: **HIV-1 gp120 induces NFAT nuclear translocation in resting CD4+ T-cells**. *Virology* 2006, **345**:105–114.

[271] Stacey A, Norris P, Qin L, Haygreen E, Taylor E, Heitman J, Lebedeva M, DeCamp A, Li D, Grove D, et al.: **Induction of a striking systemic cytokine cascade prior to peak viremia in acute human immunodeficiency virus type 1 infection, in contrast to more modest and delayed responses in acute hepatitis B and C virus infections**. *The Journal of Virology* 2009, **83**(8):3719–3733.

[272] Gavioli R, Gallerani E, Fortini C, Fabris M, Bottoni A, Canella A, Bonaccorsi A, Marastoni M, Micheletti F, Cafaro A, et al.: **HIV-1 Tat Protein Modulates the Generation of Cytotoxic T Cell Epitopes by Modifying Proteasome Composition and Enzymatic Activity 1**. *The Journal of Immunology* 2004, **173**(6):3838–3843.

[273] Remoli A, Marsili G, Perrotti E, Gallerani E, Ilari R, Nappi F, Cafaro A, Ensoli B, Gavioli R, Battistini A: **Intracellular HIV-1 Tat protein represses constitutive LMP2 transcription increasing proteasome activity by interfering with the binding of IRF-1 to STAT1**. *Biochemistry Journal* 2006, **396**:371–380.

[274] Varbanov M, Espert L, Biard-Piechaczyk M: **Mechanisms of CD4 T-cell depletion triggered by HIV-1 viral proteins**. *AIDS Review* 2006, **8**(4):221–236.

[275] Lum J, Badley A: **Resistance to apoptosis: mechanism for the development of HIV reservoirs**. *Current HIV Research* 2003, **1**(3):261–274.

[276] Capaldi R: **Structure and function of cytochrome c oxidase**. *Annual Review of Biochemistry* 1990, **59**:569–596.

[277] Brazil D, Hemmings B: **Ten years of protein kinase B signalling: a hard Akt to follow**. *Trends in Biochemical Sciences* 2001, **26**(11):657–664.

[278] Yi R, Bogerd H, Cullen B: **Recruitment of the Crm1 nuclear export factor is sufficient to induce cytoplasmic expression of incompletely spliced human immunodeficiency virus mRNAs**. *Journal of Virology* 2002, **76**(5):2036–2042.

[279] Stutz F, Izaurralde E, Mattaj I, Rosbash M: **A role for nucleoporin FG repeat domains in export of human immunodeficiency virus type 1 Rev protein and RNA from the nucleus**. *Molecular and Cellular Biology* 1996, **16**(12):7144–7150.

[280] Zolotukhin A, Felber B: **Nucleoporins nup98 and nup214 participate in nuclear export of human immunodeficiency virus type 1 Rev**. *Journal of Virology* 1999, **73**:120–127.

[281] Chook Y, Blobel G: **Karyopherins and nuclear import**. *Current Opinion in Structural Biology* 2001, **11**(6):703–715.

[282] Sherman M, Greene W: **Slipping through the door: HIV entry into the nucleus**. *Microbes and Infection* 2002, **4**:67–73.

[283] Agostini I, Popov S, Li J, Dubrovsky L, Hao T, Bukrinsky M: **Heat-shock protein 70 can replace viral protein R of HIV-1 during nuclear import of the viral preintegration complex**. *Experimental Cell Research* 2000, **259**(2):398–403.

[284] Christ F, Thys W, De Rijck J, Gijsbers R, Albanese A, Arosio D, Emiliani S, Rain J, Benarous R, Cereseto A, et al.: **Transportin-SR2 imports HIV into the nucleus**. *Current Biology* 2008, **18**(16):1192–1202.

[285] Krishnan L, Matreyek K, Oztop I, Lee K, Tipper C, Li X, Dar M, KewalRamani V, Engelman A: **The Requirement for Cellular Transportin 3 (TNPO3 or TRN-SR2) during Infection Maps to Human Immunodeficiency Virus Type 1 Capsid and Not Integrase**. *The Journal of Virology* 2010, **84**:397–406.

[286] Bieniasz P: **Late budding domains and host proteins in enveloped virus release**. *Virology* 2006, **344**:55–63.

[287] Kino T, Pavlakis G: **Partner molecules of accessory protein Vpr of the human immunodeficiency virus type 1**. *DNA and Cell Biology* 2004, **23**(4):193–205.

[288] Vanegas M, Llano M, Delgado S, Thompson D, Peretz M, Poeschla E: **Identification of the LEDGF/p75 HIV-1 integrase-interaction domain and NLS reveals NLS-independent chromatin tethering**. *Journal of Cell Science* 2005, **118**(8):1733–1743.

[289] Llano M, Saenz D, Meehan A, Wongthida P, Peretz M, Walker W, Teo W, Poeschla E: **An essential role for LEDGF/p75 in HIV integration**. *Science* 2006, **314**(5798):461–464.

[290] Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita K, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG**. *Nucleic Acids Research* 2006, **34**(Database Issue):D354–D537.

[291] Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, et al.: **KEGG for linking genomes to life and the environment**. *Nucleic Acids Research* 2008, **36**(Database issue):D480–D484.

[292] Ford E, Puronen C, Sereti I: **Immunopathogenesis of asymptomatic chronic HIV Infection: the calm before the storm**. *Current Opinion in HIV and AIDS* 2009, **4**(3):206–214.

[293] Goila-Gaur R, Strebel K: **HIV-1 Vif, APOBEC, and intrinsic immunity**. *Retrovirology* 2008, **5**:51.

[294] Lehner T, Wang Y, Pido-Lopez J, Whittall T, Bergmeier L, Babaahmady K: **The emerging role of innate immunity in protection against HIV-1 infection**. *Vaccine* 2008, **26**(24):2997–3001.

[295] Barr S, Smiley J, Bushman F: **The interferon response inhibits HIV particle production by induction of TRIM22**. *PLoS Pathogens* 2008, **4**(2):e1000007.

[296] Tokarev A, Skasko M, Fitzpatrick K, Guatelli J: **Antiviral activity of the interferon-induced cellular protein BST-2/tetherin**. *AIDS Research and Human Retroviruses* 2009, **25**(12):1197–1210.

[297] Dickerson J, Pinney J, Robertson D: **The biological context of HIV-1 host interactions reveals subtle insights into a system hijack**. *BMC Systems Biology* 2010, **4**:80.

[298] Bandyopadhyay S, Kelley R, Ideker T: **Discovering regulated networks during HIV-1 latency and reactivation.** In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 2006:354–366.

[299] Chatterji U, Lim P, Bobardt M, Wieland S, Cordek D, Vuagniaux G, Chisari F, Cameron C, Targett-Adams P, Parkinson T, et al.: **HCV resistance to cyclosporin A does not correlate with a resistance of the NSSA?cyclophilin A interaction to cyclophilin inhibitors**. *Journal of Hepatology* 2010, **53**:50–56.

[300] Nishimura-Sakurai Y, Sakamoto N, Mogushi K, Nagaie S, Nakagawa M, Itsui Y, Tasaka-Fujita M, Onuki-Karakama Y, Suda G, Mishima K, et al.: **Comparison of HCV-associated gene expression and cell signaling pathways in cells with or without HCV replicon and in replicon-cured cells**. *Journal of Gastroenterology* 2010, **45**(5):523–536.

[301] Burlone M, Budkowska A: **Hepatitis C virus cell entry: role of lipoproteins and cellular receptors**. *Journal of General Virology* 2009, **90**(5):1055–1070.

[302] Lohmann V, Körner F, Koch J, Herian U, Theilmann L, Bartenschlager R: **Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line**. *Science* 1999, **285**(5424):110–113.

[303] Blight K, Kolykhalov A, Rice C: **Efficient initiation of HCV RNA replication in cell culture**. *Science* 2000, **290**(5498):1972–1974.

[304] Zhong J, Gastaminza P, Cheng G, Kapadia S, Kato T, Burton D, Wieland S, Uprichard S, Wakita T, Chisari F: **Robust hepatitis C virus infection in vitro**. *Proceedings of the National Academy of Sciences* 2005, **102**(26):9294–9299.

[305] Sumpter Jr R, Loo Y, Foy E, Li K, Yoneyama M, Fujita T, Lemon S, Gale Jr M: **Regulating intracellular antiviral defense and permissiveness to hepatitis C virus RNA replication through a cellular RNA helicase, RIG-I**. *Journal of Virology* 2005, **79**(5):2689–2699.

[306] Pedersen I, Cheng G, Wieland S, Volinia S, Croce C, Chisari F, David M: **Interferon modulation of cellular microRNAs as an antiviral mechanism**. *Nature* 2007, **449**(7164):919–922.

[307] Zhong J, Gastaminza P, Chung J, Stamataki Z, Isogawa M, Cheng G, McKeating J, Chisari F: **Persistent hepatitis C virus infection in vitro: coevolution of virus and host**. *Journal of Virology* 2006, **80**(22):11082–11093.

[308] Pileri P, Uematsu Y, Campagnoli S, Galli G, Falugi F, Petracca R, Weiner A, Houghton M, Rosa D, Grandi G, et al.: **Binding of hepatitis C virus to CD81**. *Science* 1998, **282**(5390):938–941.

[309] Gosert R, Egger D, Lohmann V, Bartenschlager R, Blum H, Bienz K, Moradpour D: **Identification of the hepatitis C virus RNA replication complex in Huh-7 cells harboring subgenomic replicons**. *Journal of Virology* 2003, **77**(9):5487–5492.

[310] Nakabayashi H, Taketa K, Miyano K, Yamane T, Sato J: **Growth of human hepatoma cell lines with differentiated functions in chemically defined medium**. *Cancer Research* 1982, **42**(9):3858–3863.

[311] Kato T, Date T, Miyamoto M, Furusaka A, Tokushige K, Mizokami M, Wakita T: **Efficient replication of the genotype 2a hepatitis C virus subgenomic replicon* 1**. *Gastroenterology* 2003, **125**(6):1808–1817.

[312] **The R Project for Statistical Computing (www.r-project.org/)**.

[313] Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al.: **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biology* 2004, **5**(10):R80.

[314] **QC Report Generation for affyBatch objects (www.bioconductor.org/packages/2.6/bioc/html/affyQCReport.html)**.

[315] **A collection of PCA methods (www.bioconductor.org/packages/2.6/bioc/html/pcaMethods**

[316] **Various R programming tools for plotting data (cran.r-project.org/web/packages/gplots/index.html)**.

[317] Quinkert D, Bartenschlager R, Lohmann V: **Quantitative analysis of the hepatitis C virus replication complex**. *Journal of Virology* 2005, **79**(21):13594–13605.

[318] Tripathi L, Kataoka C, Taguwa S, Moriishi K, Mori Y, Matsuura Y, Mizuguchi K: **Network based analysis of hepatitis C virus Core and NS4B protein interactions**. *Molecular bioSystems* 2010, **6**(12):2539–2553.

[319] Bigger C, Guerra B, Brasky K, Hubbard G, Beard M, Luxon B, Lemon S, Lanford R: **Intrahepatic gene expression during chronic hepatitis C virus infection in chimpanzees**. *Journal of Virology* 2004, **78**(24):13779–13792.

[320] Woodhouse S, Narayan R, Latham S, Lee S, Antrobus R, Gangadharan B, Luo S, Schroth G, Klenerman P, Zitzmann N: **Transcriptome sequencing, microarray, and proteomic analyses reveal cellular and metabolic impact of hepatitis C virus infection in vitro**. *Hepatology* 2010, **52**(2):443–453.

[321] Bairoch A, Apweiler R, Wu C, Barker W, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al.: **The universal protein resource (UniProt)**. *Nucleic Acids Research* 2005, **33**(Database Issue):D154–D159.

[322] Roohvand F, Maillard P, Lavergne J, Boulant S, Walic M, Andréo U, Goueslain L, Helle F, Mallet A, McLauchlan J, et al.: **Initiation of Hepatitis C Virus Infection Requires the Dynamic Microtubule Network**. *Journal of Biological Chemistry* 2009, **284**(20):13778–13794.

[323] Ariumi Y, Kuroki M, Abe K, Dansako H, Ikeda M, Wakita T, Kato N: **DDX3 DEAD-box RNA helicase is required for hepatitis C virus RNA replication**. *Journal of Virology* 2007, **81**(24):13922–13926.

[324] Schröder M: **Human DEAD-box protein 3 has multiple functions in gene regulation and cell cycle control and is a prime target for viral manipulation**. *Biochemical Pharmacology* 2010, **79**(3):297–306.

[325] Gale M, Foy E: **Evasion of intracellular host defence by hepatitis C virus**. *Nature* 2005, **436**(7053):939–945.

[326] Edelmann M, Kessler B: **Ubiquitin and ubiquitin-like specific proteases targeted by infectious pathogens: emerging patterns and molecular principles**. *Biochimica et Biophysica Acta Molecular Basis of Disease* 2008, **1782**(12):809–816.

[327] Ritchie K, Hahn C, Kim K, Yan M, Rosario D, Li L, de la Torre J, Zhang D: **Role of ISG15 protease UBP43 (USP18) in innate immunity to viral infection**. *Nature Medicine* 2004, **10**(12):1374–1378.

[328] Holowaty M, Zeghouf M, Wu H, Tellam J, Athanasopoulos V, Greenblatt J, Frappier L: **Protein profiling with Epstein-Barr nuclear antigen-1 reveals an interaction with the herpesvirus-associated ubiquitin-specific protease HAUSP/USP7**. *Journal of Biological Chemistry* 2003, **278**(32):29987–29994.

[329] Boutell C, Canning M, Orr A, Everett R: **Reciprocal activities between herpes simplex virus type 1 regulatory protein ICP0, a ubiquitin E3 ligase, and ubiquitin-specific protease USP7**. *Journal of Virology* 2005, **79**(19):12342–12354.

[330] Lu Y, Adegoke O, Nepveu A, Nakayama K, Bedard N, Cheng D, Peng J, Wing S: **USP19 deubiquitinating enzyme supports cell proliferation by stabilizing KPC1, a ubiquitin ligase for p27Kip1**. *Molecular and Cellular Biology* 2009, **29**(2):547–558.

[331] Taniguchi H, Kato N, Otsuka M, Goto T, Yoshida H, Shiratori Y, Omata M: **Hepatitis C virus core protein upregulates transforming growth factor-β1 transcription**. *Journal of medical virology* 2004, **72**:52–59.

[332] Everett H, McFadden G: **Apoptosis: an innate immune response to virus infection**. *Trends in Microbiology* 1999, **7**(4):160–165.

[333] Inoue Y, Murakami K, Hmwe S, Aizaki H, Suzuki T: **Transcriptomic comparison of human hepatoma huh-7 cell clones with different hepatitis C virus replication efficiencies**. *Japanese Journal of Infectious Disease* 2007, **60**(4):173–178.

[334] Evans M, von Hahn T, Tscherne D, Syder A, Panis M, Wölk B, Hatziioannou T, McKeating J, Bieniasz P, Rice C: **Claudin-1 is a hepatitis C virus co-receptor required for a late step in entry**. *Nature* 2007, **446**(7137):801–805.

[335] Chang K, Jiang J, Cai Z, Luo G: **Human apolipoprotein e is required for infectivity and production of hepatitis C virus in cell culture**. *Journal of Virology* 2007, **81**(24):13783–13793.

[336] Merola M, Brazzoli M, Cocchiarella F, Heile J, Helenius A, Weiner A, Houghton M, Abrignani S: **Folding of hepatitis C virus E1 glycoprotein in a cell-free system**. *Journal of Virology* 2001, **75**(22):11205–11217.

[337] Koyama S, Ishii K, Coban C, Akira S: **Innate immune response to viral infection**. *Cytokine* 2008, **43**(3):336–341.

[338] Schuppan D, Krebs A, Bauer M, Hahn E: **Hepatitis C and liver fibrosis**. *Cell Death & Differentiation* 2003, **10**:S59–S67.

[339] Drouet C, Bouillet L, Csopaki F, Colomb M: **Hepatitis C virus NS3 serine protease interacts with the serpin C1 inhibitor**. *FEBS letters* 1999, **458**(3):415–418.

[340] Ghosh S, Zhao B, Bie J, Song J: **Macrophage cholesteryl ester mobilization and atherosclerosis**. *Vascular Pharmacology* 2010, **52**(1-2):1–10.

[341] Braunewell K, Szanto A: **Visinin-like proteins (VSNLs): Interaction partners and emerging functions in signal transduction of a subfamily of neuronal Ca 2+-sensor proteins**. *Cell and Tissue Research* 2009, **335**(2):301–316.

[342] Levy D, Lee C: **What does Stat3 do?** *Journal of Clinical Investigation* 2002, **109**(9):1143–1148.

[343] Yoshida T, Hanada T, Tokuhisa T, Kosai K, Sata M, Kohara M, Yoshimura A: **Activation of STAT3 by the hepatitis C virus core protein leads to cellular transformation**. *The Journal of Experimental Medicine* 2002, **196**(5):641–653.

[344] Zhu H, Shang X, Terada N, Liu C: **STAT3 induces anti-hepatitis C viral activity in liver cells**. *Biochemical and Biophysical Research Communications* 2004, **324**(2):518–528.

[345] Georgel P, Schuster C, Zeisel M, Stoll-Keller F, Berg T, Bahram S, Baumert T: **Virus?host interactions in hepatitis C virus infection: implications for molecular pathogenesis and antiviral strategies**. *Trends in Molecular Medecine* 2010, **16**(6):277–286.

[346] Park K, Choi S, Lee S, Hwang S, Lai M: **Nonstructural 5A protein of hepatitis C virus modulates tumor necrosis factor α-stimulated nuclear factor κB activation**. *Journal of Biological Chemistry* 2002, **277**(15):13122–13128.

[347] Nguyen H, Sankaran S, Dandekar S: **Hepatitis C virus core protein induces expression of genes regulating immune evasion and anti-apoptosis in hepatocytes**. *Virology* 2006, **354**:58–68.

[348] Choi S, Park K, Ahn B, Jung G, Lai M, Hwang S: **Hepatitis C virus nonstructural 5B protein regulates tumor necrosis factor alpha signaling through effects on cellular I {kappa} B kinase**. *Molecular and Cellular Biology* 2006, **26**(8):3048–3059.

[349] Hayden M, Ghosh S: **Shared principles in NF-[kappa] B signaling**. *Cell* 2008, **132**(3):344–362.

[350] Alvarez S, Harikumar K, Hait N, Allegood J, Strub G, Kim E, Maceyka M, Jiang H, Luo C, Kordula T, et al.: **Sphingosine-1-phosphate is a missing cofactor for the E3 ubiquitin ligase TRAF2**. *Nature* 2010, **465**(7301):1084–1088.

[351] Dong B, Zhou Q, Zhao J, Zhou A, Harty R, Bose S, Banerjee A, Slee R, Guenther J, Williams B, et al.: **Phospholipid scramblase 1 potentiates the antiviral activity of interferon**. *Journal of Virology* 2004, **78**(17):8983–8993.

[352] Saddar S, Mineo C, Shaul P: **Signaling by the High-Affinity HDL Receptor Scavenger Receptor B Type I**. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2010, **30**(2):144–150.

[353] Aizaki H, Morikawa K, Fukasawa M, Hara H, Inoue Y, Tani H, Saito K, Nishijima M, Hanada K, Matsuura Y, et al.: **Critical role of virion-associated cholesterol and sphingolipid in hepatitis C virus infection**. *Journal of Virology* 2008, **82**(12):5715–5724.

[354] Yvan-Charvet L, Wang N, Tall A: **Role of HDL, ABCA1, and ABCG1 transporters in cholesterol efflux and immune responses**. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2010, **30**(2):139–143.

[355] Shi S, Polyak S, Tu H, Taylor D, Gretch D, Lai M: **Hepatitis C virus NS5A colocalizes with the core protein on lipid droplets and interacts with apolipoproteins**. *Virology* 2002, **292**(2):198–210.

[356] Raney K, Sharma S, Moustafa I, Cameron C: **Hepatitis C virus non-structural protein 3 (HCV NS3): a multifunctional antiviral target**. *Journal of Biological Chemistry* 2010, **285**(30):22725–22731.

[357] Richer M, Juliano L, Hashimoto C, Jean F: **Serpin Mechanism of Hepatitis C Virus Nonstructural 3 (NS3) Protease Inhibition**. *Journal of Biological Chemistry* 2004, **279**(11):10222–10227.

[358] Huang Y, Cheng J, Zhang S, Wang L, Guo J, Liu Y, Yang Y, Zhang L, Bai G, Gao X, et al.: **Screening of hepatocyte proteins binding to F protein of hepatitis C virus by yeast two-hybrid system**. *World Journal of Gastroenterology* 2005, **11**(36):5659–5665.

[359] Pipe S, Morris J, Shah J, Kaufman R: **Differential interaction of coagulation factor VIII and factor V with protein chaperones calnexin and calreticulin**. *Journal of Biological Chemistry* 1998, **273**(14):8537–8544.

[360] Molinari M, Helenius A: **Chaperone selection during glycoprotein translocation into the endoplasmic reticulum**. *Science* 2000, **288**(5464):331–333.

[361] Dudek J, Greiner M, Müller A, Hendershot L, Kopsch K, Nastainczyk W, Zimmermann R: **ERj1p has a basic role in protein biogenesis at the endoplasmic reticulum**. *Nature Structural & Molecular Biology* 2005, **12**(11):1008–1014.

[362] Chen B, Piel W, Gui L, Bruford E, Monteiro A: **The HSP90 family of genes in the human genome: insights into their divergence and evolution**. *Genomics* 2005, **86**(6):627–637.

[363] Okamoto T, Nishimura Y, Ichimura T, Suzuki K, Miyamura T, Suzuki T, Moriishi K, Matsuura Y: **Hepatitis C virus RNA replication is regulated by FKBP8 and Hsp90**. *The EMBO journal* 2006, **25**(20):5015–5025.

[364] Appay V, Sauce D: **Immune activation and inflammation in HIV-1 infection: causes and consequences**. *The Journal of Pathology* 2008, **214**(2):231–241.

[365] Morgan D, Mahe C, Mayanja B, Whitworth J, Kilmarx P: **Progression to symptomatic disease in people infected with HIV-1 in rural Uganda: prospective cohort studyCommentary: Virus, host, or environment?** *Bmj* 2002, **324**(7331):193–197.

[366] Paiardini M, Pandrea I, Apetrei C, Silvestri G: **Lessons learned from the natural hosts of HIV-related viruses**. *Annual Review of Medicine* 2009, **60**:485–495.

[367] Gardner M: **The history of simian AIDS**. *Journal of Medical Primatology* 1996, **25**(3):148–157.

[368] Milush J, Stefano-Cole K, Schmidt K, Durudas A, Pandrea I, Sodora D: **Mucosal innate immune response associated with a timely humoral immune response and slower disease progression after oral transmission of simian immunodeficiency virus to rhesus macaques**. *Journal of Virology* 2007, **81**(12):6175–6186.

[369] Kumar L, Futschik M: **Mfuzz: a software package for soft clustering of microarray data**. *Bioinformation* 2007, **2**:5.

[370] Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995, **57**:289–300.

[371] Van den Bulcke T, Van Leemput K, Naudts B, Van Remortel P, Ma H, Verschoren A, De Moor B, Marchal K: **SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms**. *BMC Bioinformatics* 2006, **7**:43.

[372] Li Y, Li C, Xue P, Zhong B, Mao A, Ran Y, Chen H, Wang Y, Yang F, Shu H: **ISG56 is a negative-feedback regulator of virus-triggered signaling and cellular antiviral response**. *Proceedings of the National Academy of Sciences* 2009, **106**(19):7945–7950.

[373] Elder R, Yu M, Chen M, Zhu X, Yanagida M, Zhao Y: **HIV-1 Vpr induces cell cycle G2 arrest in fission yeast (Schizosaccharomyces pombe) through a pathway involving regulatory and catalytic subunits of PP2A and acting on both Wee1 and Cdc25**. *Virology* 2001, **287**(2):359–370.

[374] Honda K, Yanai H, Negishi H, Asagiri M, Sato M, Mizutani T, Shimada N, Ohba Y, Takaoka A, Yoshida N, et al.: **IRF-7 is the master regulator of type-I interferon-dependent immune responses**. *Nature* 2005, **434**(7034):772–777.

[375] Ayyavoo V, Mahboubi A, Mahalingam S, Ramalingam R, Kudchodkar S, Williams W, Green D, Weiner D: **HIV-1 Vpr suppresses immune activation and apoptosis through regulation of nuclear factor xB**. *Nature Medicine* 1997, **3**(10):1117–1123.

[376] Conti L, Fantuzzi L, Del Cornŋ M, Belardelli F, Gessani S: **Immunomodulatory effects of the HIV-1 gp120 protein on antigen presenting cells: implications for AIDS pathogenesis**. *Immunobiology* 2004, **209**(1-2):99–115.

[377] Qiao X, He B, Chiu A, Knowles D, Chadburn A, Cerutti A: **Human immunodeficiency virus 1 Nef suppresses CD40-dependent immunoglobulin class switching in bystander B cells**. *Nature Immunology* 2006, **7**(3):302–310.

[378] Sato M, Hata N, Asagiri M, Nakaya T, Taniguchi T, Tanaka N: **Positive feedback regulation of type I IFN genes by the IFN-inducible transcription factor IRF-7**. *FEBS Letters* 1998, **441**:106–110.

[379] OÕBrien M, Manches O, Sabado R, Baranda S, Wang Y, Marie I, Rolnitzky L, Markowitz M, Margolis D, Levy D, et al.: **Spatiotemporal trafficking of HIV in human plasmacytoid dendritic cells defines a persistently IFN-α into producing and partially matured phenotype**. *The Journal of Clinical Investigation* 2011, **121**(3):1088–1101.

[380] Lepelley A, Louis S, Sourisseau M, Law H, Pothlichet J, Schilte C, Chaperot L, Plumas J, Randall R, Si-Tahar M, et al.: **Innate Sensing of HIV-Infected Cells**. *PLoS Pathogens* 2011, **7**(2):e1001284.

[381] Hoshino S, Konishi M, Mori M, Shimura M, Nishitani C, Kuroki Y, Koyanagi Y, Kano S, Itabe H, Ishizaka Y: **HIV-1 Vpr induces TLR4/MyD88-mediated IL-6 production and reactivates viral production from latency**. *Journal of Leukocyte Biology* 2010, **87**(6):1133–1143.

[382] Okumura A, Lu G, Pitha-Rowe I, Pitha P: **Innate antiviral response targets HIV-1 release by the induction of ubiquitin-like protein ISG15**. *Proceedings of the National Academy of Sciences* 2006, **103**(5):1440–1445.

[383] Zhao C, Denison C, Huibregtse J, Gygi S, Krug R: **Human ISG15 conjugation targets both IFN-induced and constitutively expressed proteins functioning in diverse cellular pathways**. *Proceedings of the National Academy of Sciences* 2005, **102**(29):10200–10205.

[384] Skaug B, Chen Z: **Emerging role of ISG15 in antiviral immunity**. *Cell* 2010, **143**(2):187–190.

[385] Guo J, Peters K, Sen G: **Induction of the human protein P56 by interferon, double-stranded RNA, or virus infection**. *Virology* 2000, **267**(2):209–219.

[386] Zhou B, He J: **Proliferation inhibition of astrocytes, neurons, and non-glial cells by intracellularly expressed human immunodeficiency virus type 1 (HIV-1) Tat protein**. *Neuroscience Letters* 2004, **359**(3):155–158.

[387] Bergonzini V, Delbue S, Wang J, Reiss K, Prisco M, Amini S, Khalili K, Peruzzi F: **HIV-Tat promotes cellular proliferation and inhibits NGF-induced differentiation through mechanisms involving Id1 regulation**. *Oncogene* 2004, **23**(46):7701–7711.

[388] Perfettini J, Castedo M, Roumier T, Andreau K, Nardacci R, Piacentini M, Kroemer G: **Mechanisms of apoptosis induction by the HIV-1 envelope**. *Cell Death & Differentiation* 2005, **12**:916–923.

[389] Altieri D: **Survivin and IAP proteins in cell-death mechanisms**. *Biochem. J* 2010, **430**:199–205.

[390] Zhu Y, Roshal M, Li F, Blackett J, Planelles V: **Upregulation of survivin by HIV-1 Vpr**. *Apoptosis* 2003, **8**:71–79.

[391] Jiang L, Bao Y, Luo C, Hu G, Huang C, Ding X, Sun K, Lu Y: **Knockdown of ubiquitin-conjugating enzyme E2C/UbcH10 expression by RNA interference inhibits glioma cell proliferation and enhances cell apoptosis in vitro**. *Journal of Cancer Research and Clinical Oncology* 2010, **136**(2):211–217.

[392] Macarulla T, Ramos F, Tabernero J: **Aurora kinase family: a new target for anticancer drug**. *Recent Patents on Anti-Cancer Drug Discovery* 2008, **3**(2):114–122.

[393] Hartwell L, Hopfield J, Leibler S, Murray A, et al.: **From molecular to modular cell biology**. *Nature* 1999, **402**(6761):C47–C52.

[394] Joyce A, Palsson B: **The model organism as a system: integrating'omics' data sets**. *Nature Reviews Molecular Cell Biology* 2006, **7**(3):198–210.

[395] Wang J, Li M, Deng Y, Pan Y: **Recent advances in clustering methods for protein interaction networks**. *BMC Genomics* 2010, **11**(Suppl 3):S10.

[396] Segre D, DeLuna A, Church G, Kishony R: **Modular epistasis in yeast metabolism**. *Nature Genetics* 2004, **37**:77–83.

[397] Zainudin S, Deris S: **Combining Clustering and Bayesian Network for Gene Network Inference**. *Eighth International Conference on Intelligent Systems Design and Applications* 2008, :557–563.

[398] Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear E, Sevier C, Ding H, Koh J, Toufighi K, Mostafavi S, et al.: **The genetic landscape of a cell**. *Science's STKE* 2010, **327**(5964):425–431.

[399] Kelley R, Ideker T: **Systematic interpretation of genetic interactions using protein networks**. *Nature Biotechnology* 2005, **23**(5):561–566.

[400] Michaut M, Baryshnikova A, Costanzo M, Myers C, Andrews B, Boone C, Bader G: **Protein Complexes are Central in the Yeast Genetic Landscape**. *PLoS Computational Biology* 2011, **7**(2):637–643.

[401] Huang D, Sherman B, Lempicki R: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Research* 2009, **37**:1–13.

[402] Tong A, Lesage G, Bader G, Ding H, Xu H, Xin X, Young J, Berriz G, Brost R, Chang M, et al.: **Global mapping of the yeast genetic interaction network**. *Science* 2004, **303**(5659):808–813.

[403] Hughes T, Marton M, Jones A, Roberts C, Stoughton R, Armour C, Bennett H, Coffey E, Dai H, He Y, et al.: **Functional discovery via a compendium of expression profiles**. *Cell* 2000, **102**:109–126.

[404] Karypis G, Kumar V: **A fast and high quality multilevel scheme for partitioning irregular graphs**. *SIAM Journal on Scientific Computing* 1999, **20**:359–392.

[405] Matthews B: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme**. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 1975, **405**(2):442–451.

[406] Brachmann C, Sherman J, Devine S, Cameron E, Pillus L, Boeke J: **The SIR2 gene family, conserved from bacteria to humans, functions in silencing, cell cycle progression, and chromosome stability.** *Genes & Development* 1995, **9**(23):2888–2902.

[407] Chen R, Thorner J: **Systematic Epistasis Analysis of the Contributions of Protein Kinase A-and Mitogen-Activated Protein Kinase-Dependent Signaling to Nutrient Limitation-Evoked Responses in the Yeast Saccharomyces cerevisiae**. *Genetics* 2010, **185**(3):855–870.

[408] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball C, Causton H, et al.: **Minimum information about a microarray experiment (MIAME)Ñtoward standards for microarray data**. *Nature Genetics* 2001, **29**(4):365–371.

[409] Lopes C, Franz M, Kazi F, Donaldson S, Morris Q, Bader G: **Cytoscape Web: an interactive web-based network browser**. *Bioinformatics* 2010, **26**(18):2347–2348.

[410] van Dijk D, Ertaylan G, Boucher C, Sloot P: **Identifying potential survival strategies of HIV-1 through virus-host protein interaction networks**. *BMC Systems Biology* 2010, **4**:96.

[411] Qian X, Yoon B: **Comparative analysis of protein interaction networks reveals that conserved pathways are susceptible to HIV-1 interception**. *BMC Bioinformatics* 2011, **12**(Suppl 1):S19.

[412] Stumpf M, Thorne T, De Silva E, Stewart R, An H, Lappe M, Wiuf C: **Estimating the size of the human interactome**. *Proceedings of the National Academy of Sciences* 2008, **105**(19):6959–6964.

[413] Demir E, Cary M, Paley S, Fukuda K, Lemer C, Vastrik I, Wu G, D'Eustachio P, Schaefer C, Luciano J, et al.: **The BioPAX community standard for pathway data sharing**. *Nature Biotechnology* 2010, **28**(9):935–942.

[414] Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart into biological queries made easy**. *BMC Genomics* 2009, **10**:1–12.

[415] Fahey M, Bennett M, Mahon C, Jaeger S, Pache L, Kumar D, Shapiro A, Rao K, Chanda S, Craik C, et al.: **GPS-Prot: a web-based visualization platform for integrating host-pathogen interaction data**. *BMC Bioinformatics* 2011, **12**:298.

[416] Plantier J, Leoz M, Dickerson J, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson D, Simon F: **A new human immunodeficiency virus derived from gorillas**. *Nature Medicine* 2009, **15**(8):871–872.

[417] Tilton J, Wilen C, Didigu C, Sinha R, Harrison J, Agrawal-Gamse C, Henning E, Bushman F, Martin J, Deeks S, et al.: **A maraviroc-resistant HIV-1 with narrow cross-resistance to other CCR5 antagonists depends on both N-terminal and extracellular loop domains of drug-bound CCR5**. *Journal of Virology* 2010, **84**(20):10863–10876.

[418] Woo J, Robertson D, Lovell S: **Constraints on HIV-1 diversity from protein structure**. *Journal of Virology* 2010, **84**(24):12995–13003.

[419] Sauter D, Schindler M, Specht A, Landford W, Münch J, Kim K, Votteler J, Schubert U, Bibollet-Ruche F, Keele B, et al.: **Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains**. *Cell Host & Microbe* 2009, **6**(5):409–421.