# DEVELOPMENT OF A HEPATITIS C VIRUS KNOWLEDGEBASE WITH COMPUTATIONAL PREDICTION OF FUNCTIONAL HYPOTHESIS OF THERAPEUTIC RELEVANCE

## KOJO KWOFIE SAMUEL

Thesis presented in fulfillment of the requirements for the Degree of *Doctor Philosophiae* at the South African National Bioinformatics Institute, Faculty of Natural Sciences, University of the Western Cape

May 2011

UNIVERSITY *of the*
WESTERN CAPE

Supervisor: Prof. Vladimir Bajic

Co-supervisor: Prof. Alan Christoffels

# Keywords

Abstract

Association

Biomedical concepts

Database

Dictionaries

Hepatitis C Virus

Hepatocellular carcinoma

Hypothesis generation

Protein-protein interactions

Text mining

# Abstract

To ameliorate Hepatitis C Virus (HCV) therapeutic and diagnostic challenges requires robust intervention strategies, including approaches that leverage the plethora of rich data published in biomedical literature to gain greater understanding of HCV pathobiological mechanisms. The multitudes of metadata originating from HCV clinical trials as well as low and high-throughput experiments embedded in text corpora can be mined as data sources for the implementation of HCV-specific resources. HCV-customized resources may support the generation of worthy and testable hypothesis and reveal potential research clues to augment the pursuit of efficient diagnostic biomarkers and therapeutic targets. This research thesis report the development of two freely available HCV-specific web-based resources: (i) Dragon Exploratory System on Hepatitis C Virus (DESHCV) accessible via http://apps.sanbi.ac.za/DESHCV/ or http://cbrc.kaust.edu.sa/deshcv/ and (ii) Hepatitis C Virus Protein Interaction Database (HCVpro) accessible via http://apps.sanbi.ac.za/hcvpro/ or http://cbrc.kaust.edu.sa/hcvpro/.

DESHCV is a text mining system implemented using named concept recognition and co-occurrence based approaches to computationally analyze about 32, 000 HCV related abstracts obtained from PubMed. As part of DESHCV development, the pre-constructed dictionaries of the Dragon Exploratory System (DES) were enriched with HCV biomedical concepts, including HCV proteins, name variants and symbols to enable HCV knowledge specific exploration. The DESHCV query inputs consist of user-defined keywords, phrases and concepts. DESHCV is therefore an information extraction tool that enables users to computationally generate association between concepts and support the prediction of potential hypothesis with diagnostic and therapeutic relevance. Additionally, users can retrieve a list of abstracts containing tagged concepts that can be used to overcome the herculean task of manual biocuration. DESHCV has been used to simulate previously reported thalidomide-chronic hepatitis C hypothesis and also to model a potentially novel thalidomide-amantadine hypothesis.

HCVpro is a relational knowledgebase dedicated to housing experimentally detected

HCV-HCV and HCV-human protein interaction information obtained from other databases and curated from biomedical journal articles. Additionally, the database contains consolidated biological information consisting of hepatocellular carcinoma (HCC) related genes, comprehensive reviews on HCV biology and drug development, functional genomics and molecular biology data, and cross-referenced links to canonical pathways and other essential biomedical databases. Users can retrieve enriched information including interaction metadata from HCVpro by using protein identifiers, gene chromosomal locations, experiment types used in detecting the interactions, PubMed IDs of journal articles reporting the interactions, annotated protein interaction IDs from external databases, and via "string searches". The utility of HCVpro has been demonstrated by harnessing integrated data to suggest putative baseline clues that seem to support current diagnostic exploratory efforts directed towards vimentin. Furthermore, eight genes comprising of *ACLY*, *AZGP1*, *DDX3X*, *FGG*, *H19*, *SIAH1*, *SERPING1* and *THBS1* have been recommended for possible investigation to evaluate their diagnostic potential. The data archived in HCVpro can be utilized to support protein-protein interaction network-based candidate HCC gene prioritization for possible validation by experimental biologists.

# Declaration

I declare that "Development of a Hepatitis C Virus knowledgebase with computational prediction of functional hypothesis of therapeutic relevance" is my own work, that it has not been submitted for degree or examination at any other university, and that all the resources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

Samuel Kojo Kwofie          May 2011

# Acknowledgement

thank you for the administrative support. To my fellow students and other Staff not mentioned here, I greatly appreciate your role as part of my family in South Africa.

This work would not have been possible without the support of my family and friends and will like to give special thanks to all of them. To the following crew I love you all: Samuelle Marie Ama Kwofie (daughter), Agnes Aba Quarshie (mum), Mary Amoah (grandma), Therasa Nana Quartey (Aunt), Mabel Asibon (Aunt), Faustina Amoakwa (Aunt), Samuel Kwofie (Dad), Mr. Mensah (Dad), Margin-Enimil (Mentor Dad), Jeffery Mensah (bro), and Samuel Quartey (bro). To my uncles (Emmanuel and Anthony) and sisters (Mabel and Lorna), you are all remembered. And I also thank my dearest friends Felix Adusei-Danso (USA Navy and UCT/UWC Alumnus) and Chantyclaire Tiba (CPUT).

Finally, I thank the Almighty God for the gift of wisdom, life and perseverance. May the Good Lord bless my supervisors (Vlad and Alan), Rev. Fr. Mike Hagan (UWC Chaplain), family and friends in all their future endeavors.
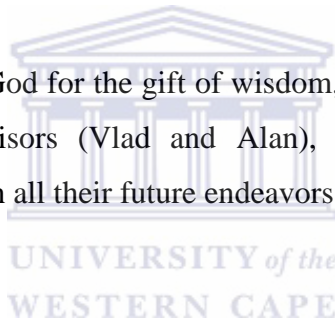
# Journal articles arising out of this thesis

**Kwofie, S. K.,** Radovanovic, A., Sundararajan, V. S., Maqungo, M., Christoffels, A., & Bajic, V. B. (2011). Dragon exploratory system on hepatitis C virus (DESHCV). *Infect Genet Evol, 11*(4), 734-739.

Maqungo, M., Kaur, M., **Kwofie, S. K**., Radovanovic, A., Schaefer, U., Schmeier, S., et al. (2011). DDPC: Dragon Database of Genes associated with Prostate Cancer. *Nucleic Acids Res, 39*(Database issue), D980-985.

**Kwofie, S. K.,** Schaefer, U., Sundararajan, V. S., Bajic, V. B., & Christoffels, A., (2011). Hepatitis C Virus Protein Interaction Database (HCVpro). *Infect Genet Evol*. **Submitted.**

# Conference proceedings arising out of this thesis

**Kwofie, S. K.,** Radovanovic, A., Maqungo, M., Bajic, V.B., Christoffels, A. (2009, December 01–03, 2009). *Hepatitis C Virus discovery database (HCVdd): a biomedical text mining and relationship exploring knowledgebase.* Abstract presented at the ISCB Africa ASBCB Joint Conference on Bioinformatics of Infectious Diseases. Bamako, Mali.

**Kwofie, S. K.,** Radovanovic, A., Maqungo, M., Bajic, V.B., Christoffels, A. (2010, December 01–03, 2010). *DESHCV: Hepatitis C Virus web-based software for biomedical text mining.* Abstract presented at the 22nd International CODATA Conference on CODATA on Scientific Data and Sustainable Development. Cape Town, South Africa.

Maqungo, M., Kaur, M., **Kwofie, S. K**., Radovanovic, A., Schaefer, U., Schmeier, S., et al. (2010, December 01–13, 2010). *Dragon database of genes associated with prostate cancer (ddpc).* Abstract presented at the 22nd International CODATA Conference on CODATA on Scientific Data and Sustainable Development. Cape Town, South Africa.

**Kwofie, S. K.,** Radovanovic, A., Bajic, V.B., Christoffels, A. (2011, March 09–11, 2011). *Inferring enriched biological information from graphs composed of text-derived biomedical concepts of ontologies related to Hepatitis C Virus.* Abstract presented at the ISCB Africa ASBCB Joint Conference on Bioinformatics of Infectious Diseases. Cape Town, South Africa.

# Table of Contents

UNIVERSITY *of the*

WESTERN CAPE

# List of Tables

# List of Figures

# List of Appendix

# Abbreviations

DES: Dragon Exploratory System

DESHCV: Dragon Exploratory System on Hepatitis C Virus

GO: Gene Ontology

HCV: Hepatitis C Virus

HCC: Hepatocellular carcinoma

HCVpro: Hepatitis C Virus Protein Interaction Database

ID: Identification

IE: Information extraction

IR: Information retrieval

LBD: Literature-Based Discovery

NER: Named Entity Recognition

NLP: Natural language processing

PPI: Protein-Protein Interaction

PPIN: Protein-Protein Interaction Network

**Chapter 1: General Introduction**
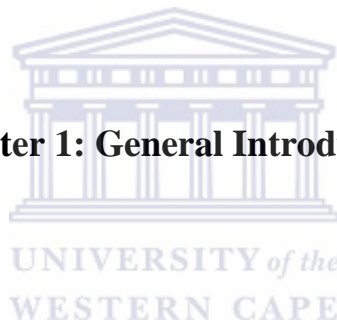
## 1.1 Overview

This chapter provides a brief overview on Hepatitis C Virus biology as well as therapeutic challenges. It also elaborates on the basic principles underlying text mining and provides a brief discussion on concerted efforts directed toward harvesting and harnessing protein-protein interactions. Furthermore, it provides the rationale underpinning the research thesis reported here.

## 1.2 Hepatitis C Virus Pathobiology

The identification and characterization of Hepatitis C Virus (HCV) had proved elusive in the 1970s and early 1980s, and HCV was inadvertently referred to as non-A, non-B associated hepatitis. However, with the advent of sophisticated molecular cloning techniques, HCV cDNA was finally isolated and characterized in 1988 (Choo et al., 1998). Although this significant discovery has proved useful, thorough understanding of HCV infection, replication and propagation in cellular hosts remains elusive. Furthermore, elucidation of the molecular pathogenesis and design of appropriate therapeutic interventions are hindered by lack of suitable animal models for HCV infection and appropriate medium for *in vitro* cultivation or propagation of HCV. Even though certain small animal models are permissive to HCV infection or allow protein expression, there is no single suitable small animal model that supports appropriately the full replication cycle of HCV. Nevertheless, rodent models obtained from rat and mice appear to have enhanced our current understanding of some aspects of the molecular etiology underlying HCV infection. Indeed significant progress could be made by developing rodent models which are capable of harboring both human immune system and human liver cells that are susceptible to HCV infection (Kremsdorf and Brezillon, 2007). The chimpanzee offers a suitable animal model susceptible to HCV infections (Lanford et al., 2001) but its usage is expensive, laborious and besieged with ethical furore. Recent development of cell cultures that support HCV infection (Wakita et al., 2005) and replication of subgenomic HCV RNAs in hepatoma cell lines (Lohmann et al.,

1999; Long et al., 2011) may contribute to understanding of HCV pathophysiological mechanisms.

## 1.2.1 Molecular Biology of HCV

Hepatitis C Virus (HCV) is a small, spherical, enveloped, hepatotropic RNA virus, belonging to the family *Flaviviridae*, in the genus of hepacivirus (Figure 1.1). The HCV genome is a single-stranded positive sense RNA of about 9600 bases, flanked by 5' and 3' non-coding regions (Figure 1.2). The genome is made up of a single, large open reading frame encoding a single polyprotein consisting of about 3000 amino acid residues. The polyprotein is cleaved by both viral proteases and host cellular signalases into 10 distinct proteins in the order as follows: NH2-Core-E1-E2-p7-NS2-NS3-NS4A-NS4B-NS5A-NS5B-COOH (Suzuki et al., 1999; Figure 1.3). The 5' non-coding region contains the internal ribosomal entry site (IRES), which is implicated in translation initiation. An additional HCV protein, alternative reading frame protein (ARFP) or otherwise known as F protein produced by ribosomal frameshift within the capsid-encoding region has been described (Xu et al., 2001). The HCV proteins Core, E1 and E2 constitute the structural proteins whilst NS2, NS3, NS4A, NS4B, NS5A and NS5B are nonstructural proteins. The structural and nonstructural proteins are cleaved by host and viral proteinases respectively. The p7 protein, a member of the viroporin family forms a bridge connecting the structural and nonstructural proteins and is cleaved by host proteinases. As pestivirus p7 was suggested not to be a major structural constituent, it can be presumed that HCV p7 protein is not a structural protein (Elbers, 1996). The major roles played by the various HCV proteins have been summarized in Table 1.1 and detailed reviews on the biochemical and functional proteins can be found elsewhere (Bartenschlager, 2004; Penin et al., 2004; Chevaliez and Pawlotsky, 2006; Dubuisson, 2007; Vassilaki and Mavromara, 2009).

**TABLE 1.1.** HCV PROTEINS, THEIR MOLECULAR WEIGHTS AND FUNCTIONS.
(Table adapted from Penin et al., 2004; Chaveliez and Pawlotsky, 2006; Dubuisson 2007).

| HCV Protein | Function | Molecular mass by SDS page (kDa) | Apparent molecular weight (kDa) |
|---|---|---|---|
| Core | RNA binding; nucleocapsid | 21 | 23 (precursor); 21 (mature) |
| E1 | Envelope glycoprotein; associate with E2 | 31-35 | 33-35 |
| E2 | Envelope glycoprotein; receptor binding; associate with E1 | 70 | 70-72 |
| P7 | Ion channel | 7 | 7 |
| NS2 | Component of NS2-3 proteinase | 21 | 21-23 |
| NS3 | NS2/3 proteinase; NS3/4 proteinase; NTPase; RNA helicase; RNA binding | 69 | 69 |
| NS4A | NS3-4A proteinase cofactor | 6 | 6 |
| NS4B | NS5A phosphorylation; membranous web induction | 27 | 27 |
| NS5A | Phosphoprotein; RNA replication by formation of replication complexes; possibly involved in inhibition of INF-alpha; inhibition of apoptosis | 56-58 | 56 (basal form); 58 (hyperphosphorylated) |
| NS5B | RNA-dependent RNA polymerase | 68 | 68 |
| F/ARF-protein | Possibly involved in HCV core protein RNA binding activities | 17 | 16-17 |
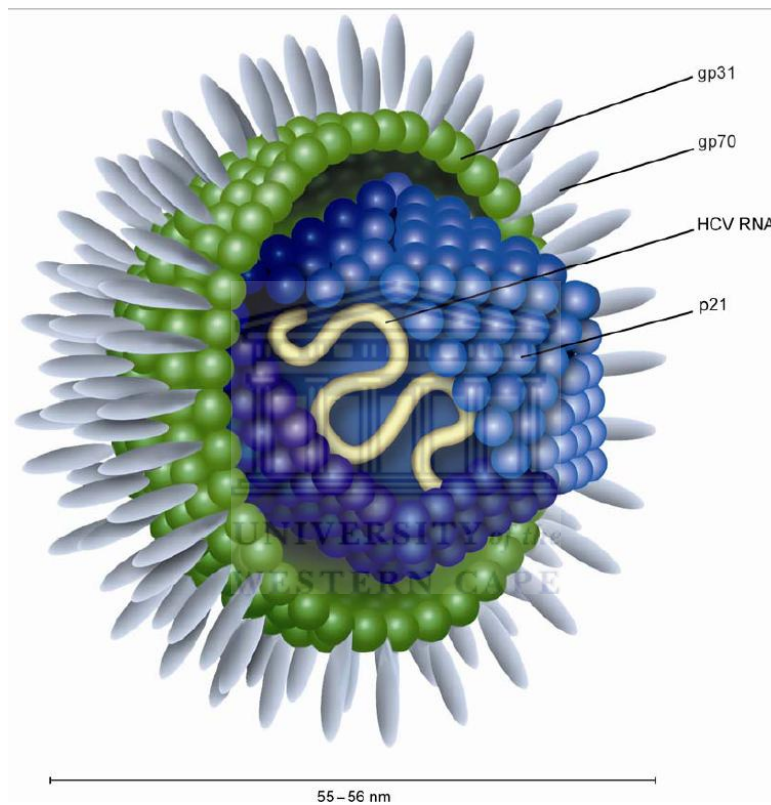
**FIGURE 1.1.** A MODEL OF HCV PARTICLE.
A diagram showing the molecular organization involving viral proteins and RNA (Krekulová et al., 2006).

**FIGURE 1.2.** HCV GENOME ORGANIZATION.
The blue and red coloured segments of the genome encode structural and nonstructural proteins respectively (Dubuisson, 2007).
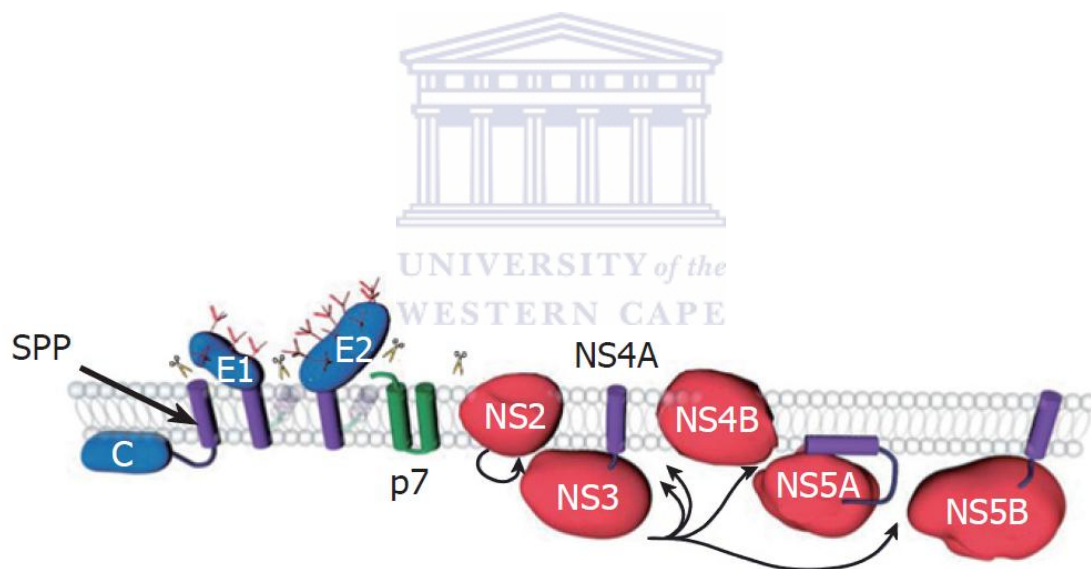


**FIGURE 1.3.** POST TRANSLATIONALLY PROCESSED HCV POLYPROTEIN
Diagram shows the 10 distinct structural and nonstructural proteins as well as junction protein p7. The scissors represent proteolytic cleavages by host peptidases, arrows indicate NS2-3 and NS3-4A cleavages. The transmembrane arrows show cleavage by host signal peptidases (Dubuisson, 2007).

## 1.2.2 HCV LIFE CYCLE

HCV host cell entry and tissue tropism seems poorly understood even though significant advances have been made by elucidating viral-cellular protein interactions as well as associated essential pathways. HCV glycoproteins E1 and E2 binds host cellular membranous constituents enabling viral interaction with a complex set of cell-surface molecules or cellular co-receptors, including tetraspanin (CD81) (Pileri et al., 1998; Flint et al., 1999; Petracca et al., 2000), human scavenger receptor class B type I (SR-BI) (Scarselli et al., 2002; Bartosch et al., 2003), dendritic cell-specific intercellular adhesion molecule-3-grabbing nonintegrin (DC-SIGN) (Gardner et al., 2003; Lozach et al., 2004), liver/lymph node-specific intracellular adhesion molecule-3 (ICAM-3)-grabbing integrin (L-SIGN) (Gardner et al., 2003; Lozach et al., 2004), low-density lipoprotein receptor (LDL-R) (Agnello et al., 1999; Monazahian et al., 1999), asialoglycoprotein receptor (ASGP-R) (Saunier et al., 2003), glycosaminoglycans (HSPG) (Barth et al., 2003), claudin-1 and occludin (Figure 1.4). Tight junction proteins claudin-1 and occludin are thought to control hepatitis C virus entry by mediating virion internalization processes in host cells (Yang et al., 2008; Liu et al., 2009; Ploss et al., 2009). Viral-cell entry also depends on clathrin-mediated endocytic pathway mechanisms, accompanied by a fusion process that appears to require an acidic endosomal environment (Codran et al., 2006). The fusion arises from conformational changes in the glycoproteins triggering decapsidation of viral nulceocapsids to release free positive-strand RNA into the cytoplasm. The liberated RNAs undergoes translation to produce a large HCV polyprotein precursor via the internal ribosome entry site located in the 5' untranslated region of the HCV genome, after which viral proteases and host signalases process the obtained polyprotein into distinct structural and nonstructural proteins through co-translation and post-translation mechanisms. Thereafter, a NS4B-induced membranous web, consisting of a membrane-associated multiprotein complex constitutes the viral replication complex (Egger et al., 2002; Elazar et al., 2004). HCV replication then proceeds via two stages catalyzed by NS5B (Chevaliez and Pawlotsky, 2006): (1) synthesis of a negative-strand RNA intermediate from a template composed of the positive-strand genome RNA and (2) negative-strand RNA intermediate then serves as template for synthesis of several progeny positive-strand RNA, which are then packaged

into new virions (Bartenschlager et al, 2004) assembled on lipid droplets (Gouttenoire et al., 2010). The matured viral particles are then released. However, the mechanisms underlying the maturation and release are still unclear. Comprehensive reviews on the HCV life and intricacies have been discussed elsewhere (Rychłowska and Bieńkowska-Szewczyk, 2007; Dubuisson et al., 2008).
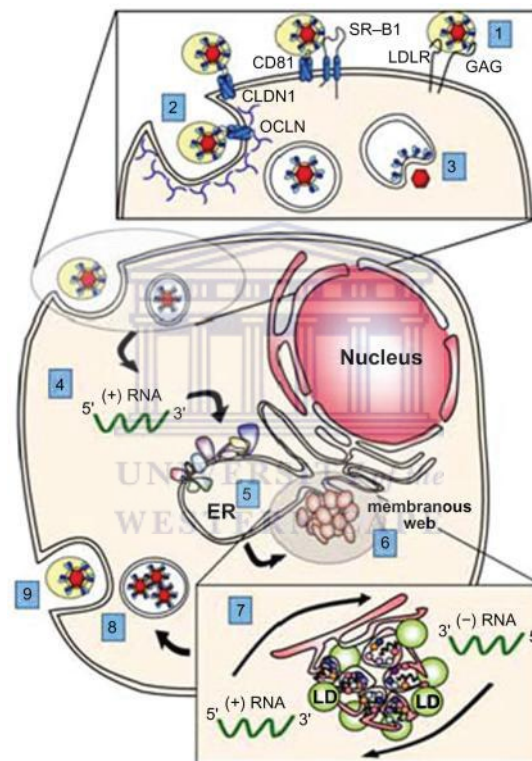
**FIGURE 1.4.** DETAILS OF HCV LIFECYCLE.
(1) Binding of HCV to cell surface receptors; (2) entry of HCV into the cell via endocytosis; (3) viral genome is released into cytoplasm; (4) internal ribosomal entry site-mediated polyprotein translation; (5) polyprotein processing; (6) formation of membranous web as part of replication complexes); (7) viral RNA replication; (8) packaging and assembly of progeny virions; and (9) matured virions are released via the host cellular secretory system (Tencate et al., 2010).

## 1.2.3 HCV nomenclature and classification

Ever since the HCV genome sequence was elucidated and vast amount of sequenced nucleotide, proteins and epitope data became available, several recommendations have emanated from various fora with the sole aim of unifying the nomenclature and classification of HCV isolates to conform to the guidelines stipulated by the International Committee on Virus Taxonomy (ICVT) (Simmonds et al., 1993; Simmonds et al., 1994; Robertson et al., 1998; Simmonds et al., 2005; Kuiken and Simmonds, 2009). This taxonomic framework guides the deposition of HCV sequences in online databases for further exploration by researchers. Furthermore, it allows the resolution of conflict pertaining to assignment of genotypes and any novel isolate that may be discovered in the future. As a result of HCV sequence variability and genetic heterogeneity, the standardization is relevant to research relating to the epidemiology, evolution and pathobiology of HCV associated infections (Kuiken and Simmonds, 2009).

By using comparative techniques including phylogenetic analysis, numerous distinct relationships have been established between HCV isolates thereby leading to the identification of six major genotypes (Simmonds et al., 1993; Simmonds et al., 1994). These divisions are further elaborated using phylogenetic trees (Figures 1.5). Additionally, the system of nomenclature has adopted the full-length genome sequence of isolate H77 (accession number AF009606) as a reference strain for annotation of other variants. The quasispecies nature of HCV allows the coexistence of multiple genetically different but closely related variants in infected individuals (Martell et al., 1992). The genotypes differ in their nucleotide sequence by 31-33% whilst the subtypes differ by 20-25%. The proteins are also known to differ by about 30% in amino acid sequences (Suzuki et al., 2007). The high degree of genetic variability exhibited by HCV could be attributed to the inability of the error-prone RNA-dependent RNA polymerase to proofread, high rate of viral replication and typical larger population size.
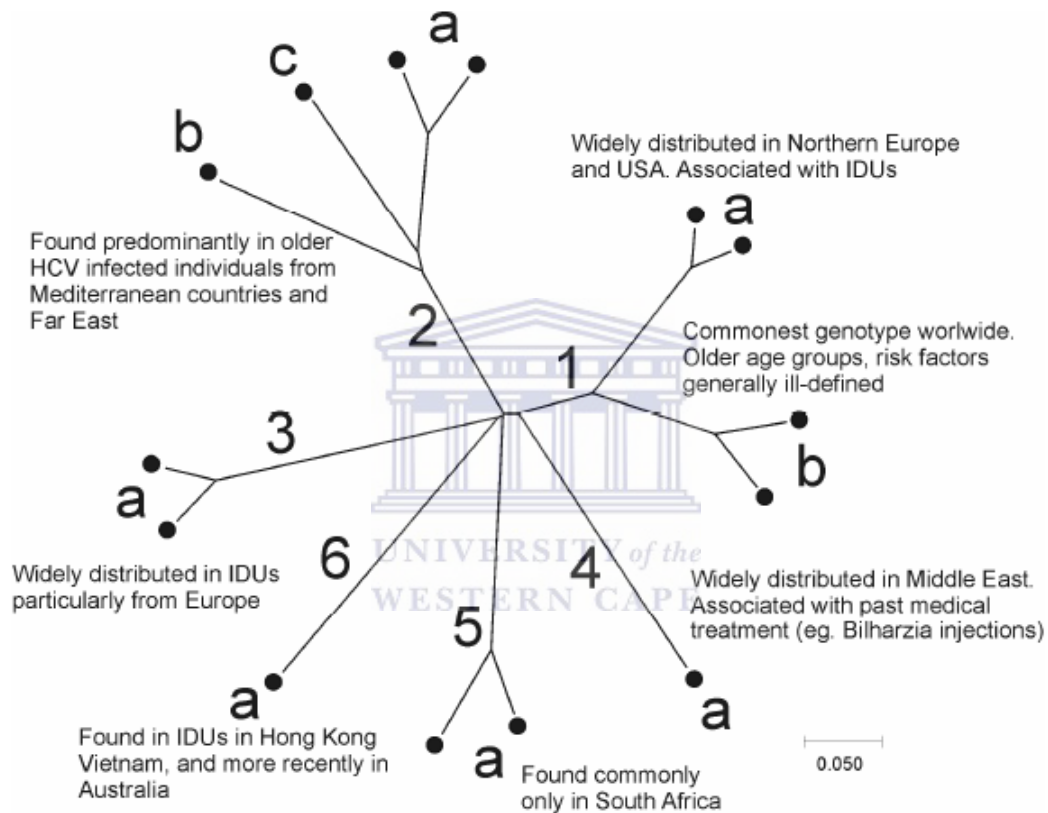
**FIGURE 1.5.** AN EVOLUTIONARY TREE DEPICTING THE MAJOR HCV GENOTYPES.
It shows prevalence according to different geographic regions or countries. (Kuiken and Simmonds, 2009).

## 1.2.4 HCV epidemiology and therapeutic challenges

Currently, about 3% of the global population is infected with HCV, a foremost risk factor and major causative agent for development of acute and chronic hepatic infections. The chronification rate is about 70-80% (Krekulová et al., 2006). HCV infected individuals can develop chronic hepatitis, which may progress from fibrosis to HCV-induced cirrhosis with some subsequently developing hepatocellular carcinoma (HCC) (Figure 1.6). HCC is the 3rd most deadly and 5th most frequent cancer globally (Zender and Kubicka, 2008) and the epidemiology is further exacerbated by the asymptomatic as well as occasionally undetectable nature of HCV infections. HCV infected individuals in the long term may ultimately require liver transplant to abort death from liver impairment. As reported by the Eurasian Harm Reduction Network (EHRN), the World Health Organization (WHO) in 2002 estimated 53,700 annual global deaths could be directly associated with HCV infection. The scale of the infection is further exacerbated by the revelation that liver cancer caused by HCV may be likely linked to 308,000 deaths annually. In total, HCV could be implicated in about 500,000 global deaths annually (EHRN, 2007) and the potential exists for these mortality and morbidity rates to increase in the near future.

The HCV prevalence have been shown to differ amongst various geographic regions in the world (Sy and Jamal, 2006; Baldo et al., 2008). Infection figures from the World Health Organization (2009) are reported here. HCV prevalence was less than 1% in Australia, Canada and Northern Europe; about 1% in USA and most of Europe; and more than 2% in most countries found in Africa, Latin America and Central and South-Eastern Asia. Furthermore, HCV prevalence figures between 5 and 10% are usually reported in certain countries in the aforementioned regions. Countries with relatively high prevalence rates include Egypt (15%), Pakistan (4.7%) and Taiwan (4.4%) (Sievert et al., 2011). Particularly, the Nile delta region of Egypt has significantly higher seroprevalence of HCV, which was revealed to increase with age from 19% in persons of 10-19 years old to approximately 60% in persons of 30 years of age.   The Egyptian HCV crisis could be attributed to the earlier lapses that existed during the nationwide campaign to help reduce the prevalence of schistosomiasis (Yahia, 2011). Intravenous drug use and blood

transfusion-related infections appear as the emerging risk factors in both developed and developing countries, respectively (Sy and Jamal, 2006).

Diagnostically, histologic liver biopsy, ultrasound, and biomarkers are among the variety of techniques employed to evaluate progression of chronic infections to HCC and hepatic steatosis (Kim et al., 2002; Mas et al. 2009; Yu et al., 2010). Unfortunately, liver biopsy is invasive with possible complications (Caillot et al., 2009) whilst reduced sensitivity and specificity can hamper the effectiveness of diagnostic techniques (Outwater et al., 2010). The need for diagnoses or evaluation of liver disease progression with a combination of biomarkers has been proposed (Gangadharan et al., 2007).

The most effective treatment for HCV infection is a combination therapy of pharmacokinetically enhanced pegylated interferon alpha (peginterferon alpha) and nucleoside analogue ribavirin (Fried et al., 2002; Krekulová et al., 2006), but it is sometimes accompanied by undesirable side effects and reduced quality of life during medication. Under standard care medication, this combination treatment cures about 80% of individuals infected with genotype 2 or 3, and 40% of genotype 1 (Strader et al., 2004; Suzuki et al., 2007; Gambarin-Gelwan and Jacobson, 2008; WHO, 2010). It is perceived that about half of infected individuals do not derive significant long term benefit from the combination therapy. It is proposed that the current treatment options can be augmented with an immunotherapeutic vaccine which may be beneficial in immunocompromised individuals (Sharma, 2010). Recently, the search for HCV therapy was vigorously focused on the development of anti-viral drugs targeting host cellular proteins and also protease/polymerase inhibitors specifically designed to disrupt the role of the HCV proteins during viral propagation and replication (Tencate et al., 2010). Some of the HCV or host protein-specific drugs under considerations include lectin cyanovirin-N (CV-N) (Helle et al., 2006), alpha-glucosidase I inhibitor (Celgosivir) (Durantel, 2009), n-butyl deoxynojirimycin (nB-DNJ) (Ouzounov et al., 2002), class of compounds with a thiazolidinone core structure (BMS-824) (Lemm et al., 2010), zinc mesoporphyrin (ZnMP) (Hou et al., 2010) and valyl ester prodrug (NM283) (Gardelli et al, 2009) whilst non HCV-specific anti-viral drugs comprises of nitazoxanide (Rossignol et al., 2010),

cyclophilin inhibitors (NIM811) (Ma et al., 2006; Lawitz et al, 2011) and silibinin (Payer et al., 2010).
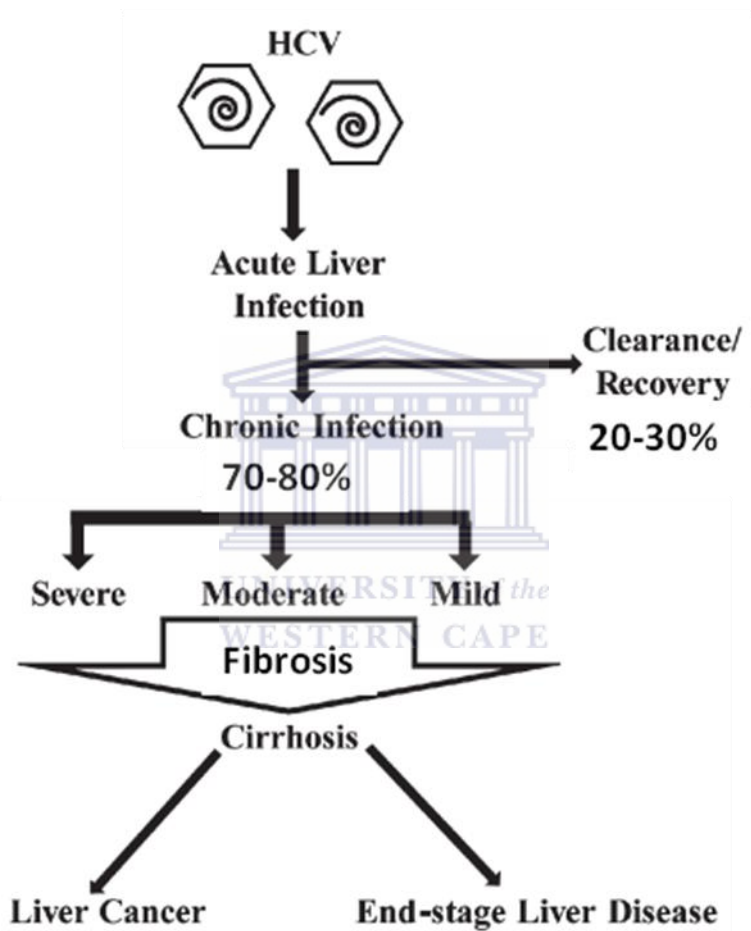
**FIGURE 1.6.** VARIOUS DISEASE STAGES FOLLOWING HCV INFECTIONS
(This diagram was modified from Sharma, 2010)

## 1.3 Text mining

Nowadays, most biomedical research institutions around the world have been expending significant proportions of their budgetary allocations in both material and human resource capacity development, thereby making it easier for them to generate large amount of potentially meaningful results for exploration in a short span of time through the usage of high throughput techniques such as genomics, proteomics, and microarray. The myriad of information derived from both low and high-throughput experiments published in journals can be located in the main body text or as supplementary information for easy electronic accessibility. As a result, biomedical researchers have adopted integrated natural language processing approaches to harness the available information in the published text as part of the strategies used in answering relevant biological questions. This is due to the fact that it is practically impossible to manually sift through the huge volumes of biomedical literature in order to unravel any meaningful latent semantic connections between different parts of the text due to potential human curation errors. The advent of online manuscript submission for journal publications may have contributed in the reported exponential growth in the amount of published biomedical literature recently (Hunter and Cohen, 2006). Currently, PubMed database (Sayers et al., 2011) has grown rapidly to over 20 million indexed citations (Figure 1.7) and houses abstracts from over 5000 life science journals from biomedical articles dating back to 1948 (Lu, 2011). Similarly, the National Library of Medicine expects 1 million journal articles to be indexed annually by 2015, indicating a 50% potential rise in number of processed articles when compared to 2004 (Neveol et al., 2006). By searching for global trends, the number of new journal articles reporting results on certain research topics such as "cell cycle" has also increased appreciably over the years (Jensen et al., 2006). Interestingly, new articles reporting previous popular topics, notable "protein Cdc28" seems to have declined. The rise in scientific review articles and concomitant explosion of scientific publications require the use of automated strategies such as text mining to computationally harvest the vast quantities of knowledge or otherwise may be referred to as knowledge about knowledge (metaknowledge) embedded in the text (Evans and Foster, 2011). Although biomedical text mining techniques and related application tools are increasingly becoming popular with bench biologists, they still face significant

challenges in extracting relevant information from texts. Nevertheless, significant improvement has been achieved through the incorporation of computational linguistics techniques. Text mining has been defined in a broader sense as any system capable of retrieving knowledge from text; and in strictest sense as a system that extracts implicit knowledge not obviously stated in the text (Zweigenbaum et al., 2007). The field of text mining can be divided into the following domains: (1) information retrieval, (2) information extraction, (3) entity recognition, and (4) analysis of text-image data.
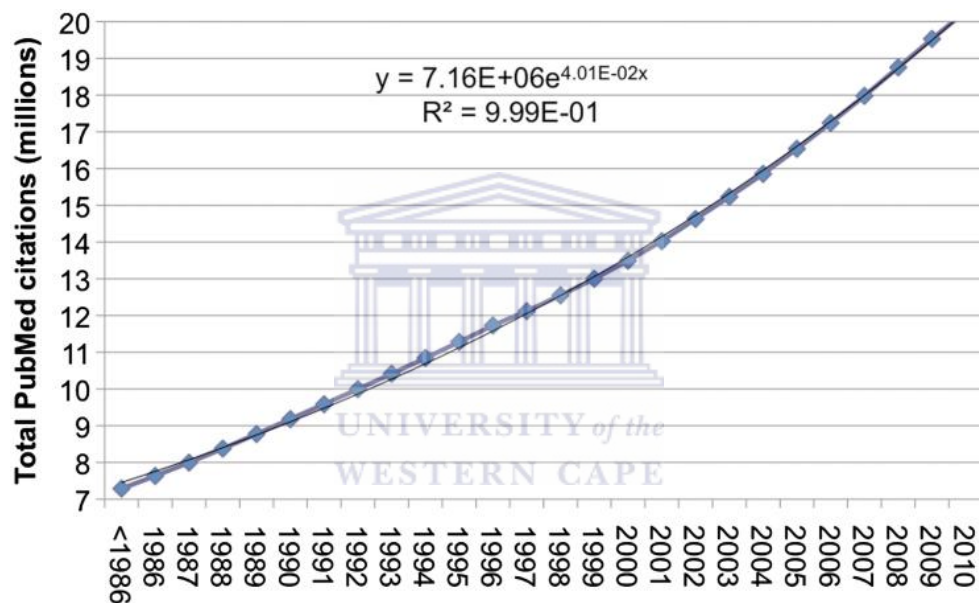
**FIGURE 1.7.** A GRAPH SHOWING THE TOTAL GROWTH OF PUBMED CITATIONS.
The graph was generated with citations from 1986 to 2010. PubMed currently houses over 20-million citations and grows at an annual rate of about 4% (Lu, 2011).

## 1.3.1 Analysis of text-image data

Most non-textual data such as figures and tables in published full-text articles convey varying degrees of experimental outcomes which aid in the interpretation of research findings. Figure legends comprising of text of sentences are used to describe or annotate non-textual data. With the increase in amount of published articles and associated text-image datasets, natural language processing and image processing techniques are employed in concert to automatically extract structured information from text. This approach has been implemented in the development of the Structured Literature Image Finder (SLIF) project, a pipeline that enables the extraction of structured information from full-text articles and also provides a user-friendly web-based tool for accessing information about subcellular location from fluorescence microscopy images (Coelho et al., 2010). The pipeline consists of three main stages, namely caption processing, image processing and latent topic discovery. During the first two stages, figures and their respective captions are extracted from a collection of papers and are used as entries to create a database. The captions are further analysed to obtain biological entities such as proteins, cell types or lines and are then mapped onto external databases. Next, the figures are panel-segmented and fluorescence microscopy images are then classified according to depicted subcellular location. The final stage is to uncover a core set of themes known as latent topics from the processed full-text papers which then form the basis for visualization and semantic interpretation that aid in new knowledge discovery. Another usefulness of a text-image mining system such as SLIF is illustrated by the fact that users can retrieve articles with images for which a topic, for instance "tumorigenesis" is overly represented for future exploration. Other publicly available online resources for retrieving images and associated papers from image caption text are Yale Image Finder (Xu et al., 2008) and GoldMiner (Kahn and Thao, 2007). Comprehensive reviews and research involving text-image dataset processing have been reported elsewhere (Müller et al., 2004a; Shatkay et al., 2006; Uwimana and Ruiz, 2008; Qian and Murphy, 2008).

## 1.3.2 Information retrieval (IR)

IR entails using automated systems, particularly a search engine to identify and rank any list of documents or portion of text relevant to user interest or based on the user's query. Simplistically, IR filters the relevant documents from a large pool of documents. The retrieved text corpora can consist of any documents comprising of full-text articles, abstracts or phrases. IR can be separated into two distinct domains consisting of *ad hoc* and text categorization systems.

## 1.3.2.1 *Ad hoc* information retrieval systems

*Ad hoc* information retrieval involves the use of user-supplied keyword queries to return subset of documents containing the queried keywords (Figure 1.8). A popular *ad hoc* IR system is PubMed, which take as input natural language in the form of free-text keywords to recover a list of citations that match the queried text (Lu, 2011). PubMed search strategy is based on both Boolean model and automatic term mapping (ATM) techniques. For a typical Boolean query, the Boolean operator "AND" is inserted between multi-term user defined query such as 'vulval cell' to retrieve a list of documents containing the search string 'vulva AND cell'. Additionally, stop words such as "the" and "it" are removed while common words ending with "-ing" and "-s" are truncated to enable name variants of the queried word to match documents (Jensen et al., 2006). The ATM system compares and maps the user-constructed keywords to a set of pre-indexed terms, for example the Medical Subject Headings (MeSH). Documents containing the user keywords as well as the mapped MeSH pre-indexed terms are retrieved as the query output (Lu, 2011). Additionally, Current IR systems being developed have been enhanced by their ability to allow flexible ranking of articles (MedlineRanker, Fontaine et al., 2009), prioritize (Pubfocus, Plikus et al., 2006) or generate multi-level relevance feedback upon querying without expert knowledge (RefMed, Yu et al., 2011). Similarly, design of a dynamic user query interface for MEDLINE database is actively being pursued with a notable example being iPubMed (Wang et al., 2010), a system that enables interactive queries with on the spot feedback to users' searches letter by letter as

well as an extra feature for fuzzy search. Also since some users of IR systems possess different language backgrounds, BabelMeSH has been designed as a platform to allow cross-language querying via MEDLINE/PubMed (Liu et al., 2006).
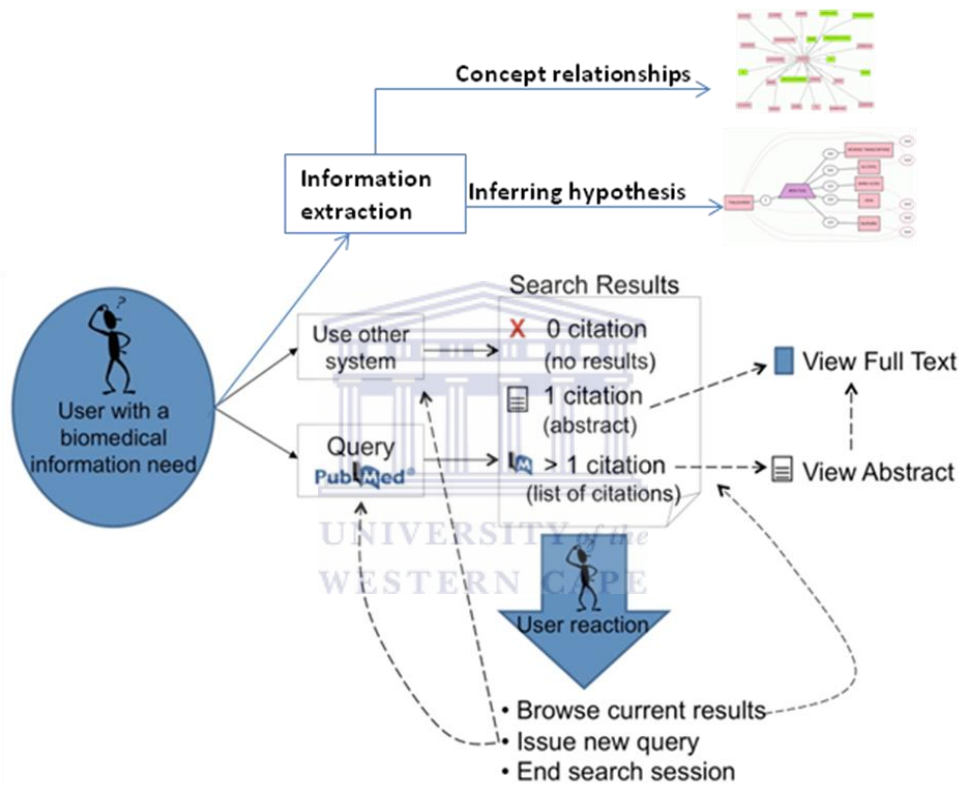
**FIGURE 1.8.** INFORMATION RETRIEVAL AND EXTRACTION SYSTEMS.
Diagram depicting the variety of results obtained by querying both information retrieval and extraction systems (modified from Lu, 2011).

## 1.3.2.2 Text categorization system

It is often not possible to use keyword queries to retrieve a subset of published biomedical documents sharing common topical characteristics among a large set of documents. Biologists confronted with large numbers of routinely published articles may require classifying these papers into smaller topic-specific groups or subgroups to enhance knowledge retrieval (Shatkay et al., 2006). For instance, a researcher may like to retrieve a set of documents pertaining to current or future research interest with ease. Text categorization therefore provides efficient approach to cluster documents into various taxonomies based on pre-defined set of categories. In this way papers reporting research in a similar domain can belong to the same category. The cluster labels assigned to taxonomies aid new researchers to locate new concepts among the pool of papers, whilst allowing experts to explore hierarchically clustered documents to effectively identify papers reporting specific concepts (Chen et al., 2006).

The categories used for clustering could be Gene Ontologies, genomic features or any shared biomedical concepts. Notable projects such as BioCreAtIvE challenge (Critical Assessment of Information Extraction in Biology) and Knowledge Discovery and Data Mining (KDD) Challenge Cup (Yeh et al., 2003) developed systems to locate short text passages that shared similar Gene Ontology annotations for specific proteins found in full-text articles. Text categorization concept has been implemented in the development of the automatic classification software, a document classification engine capable of classifying papers into topic-based hierarchy (Chen et al., 2006). The engine classifies over 7000 papers housed in Textpresso (Müller et al., 2004b) by using two-stage technique composed of a support vector machine-based classifier and a newly developed phrase-based clustering algorithm. Although the text classification has been optimized for *Caenorhabditis elegans* related papers by using human-created rules composed of an inventory of terms associated with each topic, it is amenable to any types of document sources. The usefulness of this system is illustrated as follows: classification of papers based on "Sex Determination category" returned the following top ranked clustered output: "hermaphrodite male", "sex determination", "development gene", "development cells", "cells fate", 'vulval cells", "cell signaling", "fate signaling", "precursor cell" and

"signaling induction". These results allow the user to analyze the papers according to sex determination sub-categories making it easier to avoid the herculean task of manual sorting. Another popular text categorization software platform developed by Vivisimo (Taylor, 2007) uses a different approach by allowing documents to be categorized into descriptive folders without classifying into taxonomies. Elsewhere (Aphinyanaphongs and Aliferis 2003; 2005), text categorization models were built to successfully categorize content-specific and high-quality PubMed articles to aid in Evidence Based Medicine (EBM) and were reported to outperform text mining systems employing preconstructed Boolean-based queries such as PubMed clinical query filters. On the contrary, caution must be exercised when comparing the performances of two different text mining systems since it is very tempting to make conclusive inferences based on recall, precision and F-score values without considering the fact that each system operates under characteristically different environment with unique merit.

## 1.3.4 Named Entity Recognition (NER)

This process identifies and extracts biomedical entity names from published text and assigns them to predefined semantic categories with similar topical characteristics. In certain instances, the terms "entity" and "concept" are interchangeable leading to the analogy "named concept recognition". Therefore concepts or entities can be names of genes, proteins, diseases, pathways, drugs or pharmacological substances and may belong to different categories which can be referred to as dictionaries (Kwofie et al., 2011). There are instances where categories consisting of biomedical entities are organized into ontologies, a catalogue of concepts, objects and their relationship (Müller et al., 2004b). Some of the relations generally present in almost any domains are "-is a" and "part-of", whiles model domain-specific relations are exemplified by "has-location" and "clinically-associated-with" (Spasic et al., 2005). NER is particularly useful since it circumvents the difficult process of biologists having to manually scan through a large amount of published articles to locate biomedical entities. Accurately predicted terms (entities) can be used to augment existing biological database entries or used to infer relationships for

hypothesis generation. Additionally, the terms may provide indices into literature to aid in research result communication (Hirschman et al., 2002). The earlier NER tools were designed based on variety of techniques that sought to tackle some of the major difficulties usually encountered in developing automated text recognition systems, including named entity variant structural characteristics, name ambiguities, the unavailability of common standards and fixed nomenclatures, and different morphological description of entities in the biomedical text (Franzén et al., 2002). Furthermore, systems developed from precompiled dictionaries may not be able to recognize new entities present in recently published text.

Approaches used in developing NER systems have been categorized into 3 broad techniques (Cohen and Hunter, 2008; Torii et al., 2009) consisting of (i) rule/pattern-based systems characterized by hand-written rules, (ii) statistics or machine-learning-based systems and (iii) entity lookup system via pre-compiled dictionary. Some of the systems employing the handcrafted rule-based techniques are PROSPER (Fukuda et al., 1998), Yapex (Franzén et al., 2002), ProMiner (Hanisch et al., 2003; 2005), GAPSCORE (Chang et al., 2004), Text Detective (Tamames, 2005) and another described elsewhere (Narayanaswamy et al., 2003). The dictionary lookup systems are ProtScan (Egorov et al., 2004), HMMs (Kou et al., 2005), NLProt (Mika et al., 2004) and others (Tsuruoka and Tsujii, 2004; Koike et al., 2005; Tsuruoka et al., 2007; Sasak et al., 2008). Examples of machine learning-based systems are BANNER (Leaman and Gonzalez, 2008), PowerBioNE (Zhou et al., 2004), BioCreAtIvE task 1A/II (Yeh et al., 2005; Smith et al., 2008), POSBIOTM-NER (Song et al., 2005) and others (Proux et al., 1998; Lee et al., 2004; Li et al., 2011). Ideally, NER systems are developed by combining all or some of the above-mentioned approaches (BioTagger-GM, Torii et al., 2009).

Additionally, each text mining implementation approach has peculiar advantages or disadvantages over others. For instance, dictionary-based systems intrinsically provide ID information for recognized terms by matching them to identical terms in the pre-constructed dictionary, thereby making this approach the preferable choice within the earlier stages of the named entity extraction pipelines (Tsuruoka and Tsujii, 2004).

Unfortunately, dictionary-based approach is besieged with poor recognition of short names resulting in high rate of false positives with concomitant decrease in overall precision during evaluation, and this hurdle is remedied by using machine learning approaches to heuristically filter out the false positives (Tsuruoka and Tsujii, 2004). Concerning fine-tuning of systems, machine learning techniques are easy to tune to new domains once the tagged training dataset is available but the rule-based system requires laborious human analysis to obtain a transparent system easier to fine-tune, expand and support (Franzén et al., 2002). Interestingly, the supervised machine learning approach has a disadvantage since it sometimes requires a large amount of error-free annotated corpus that needs a lot of time and extensive energy to create (Narayanaswamy et al., 2003). This deficiency has been circumvented by using rule-based approaches which exploits surface clues and simple linguistic as well as domain knowledge to locate relevant named entities in biomedical text (Narayanaswamy et al., 2003).
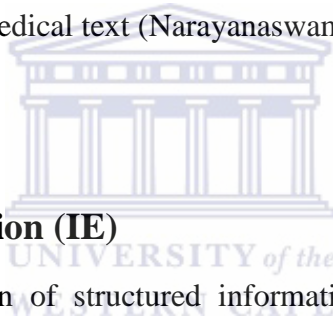
## 1.3.5 Information extraction (IE)

IE is the automated extraction of structured information from bibliome to generate relationships between biomedical concepts and hypothesis (Figure 1.8). The notable approaches for relation extraction are co-occurrence-based, pattern-matching, and natural language processing systems.

## 1.3.5.1 Co-occurrence methods

The simplest IE approach used to infer relationship between entities is the detection of co-occurring entities within an abstract, sentence or a phrase (Ding et al., 2002; Wren and Garner, 2004). If two biomedical concepts such as proteins appear together frequently in separate abstracts, the likelihood that a functionally relevant relationship exist between these proteins is high. Ideally, relationships suggested from co-occurrences are usually vague and it is impossible to infer the type or direction of the relationships (Fundel et al.,

2007). Furthermore, it is not easy to differentiate between either direct or indirect relationships amongst entities (Jensen, 2006).

Generally, co-occurrence-based methods are combined with other machine learning or natural language processing techniques to detect relationships between entities. For instance, a Hebbbian-type of learning algorithm, known as associative concept space (ACS) has been employed in combination with co-occurrence method to extract biological relationships between genes using meta-analysis of scientific texts (Jelier et al., 2005). Two genes co-occur if they are co-mentioned in the abstract, title or medical subject headings (MeSH) index of the text document. Co-occurrences between genes are weighted by using a co-occurrence matrix containing the number of times genes from a given set co-occur. For integration with ACS, only co-occurrences weighted above a specified minimum threshold are included for further analysis. ACS is a multidimensional Euclidean space composed of thesaurus concepts which are positioned in space and the computed distance existing between concepts is an indication of their supposed relatedness. ACS is ideal for relationship extraction when compared to a simple co-occurrence method because ACS has the potential to identify more detailed functional biological relationships and can also achieve better results with less literature per gene (Jelier et al., 2005). Similarly, co-occurrence method was integrated into the implementation of PubGene, an online database with variety of tools for gene-expression analysis (Jenssen et al., 2001). Pre-constructed gene-article index was used to generate a network composed of nodes denoting genes and edges representing co-occurrence of genes symbols or short gene names in the title or abstract of the same MEDLINE article. The detected associations between genes were mapped onto Gene Ontologies and MeSH index annotations. As usual, co-occurrence support the premise of meaningful biological relatedness and this assertion was validated to a reasonable extent by analyzing data from three large-scale experiments. Other works describing the use of co-occurrence methods for elucidating biological relationships between entities are reported elsewhere (Stapley and Benoit, 2000; Chaussabel and Sher, 2002; Alako et al., 2005; Chun et al., 2006; Gabow et al., 2008). Under certain circumstances, co-occurrence based IE methods can provide high recall but may possess poor precision than natural language processing, and

are employed preferable as baseline method for comparing other techniques (Zweigenbaum et al., 2007). Additionally, co-occurrence method is employed suitably as a component of exploratory tools such as STRING (von Mering et al., 2005) and Prolinks (Bowers et al., 2004) because of their potential to identify almost any kind of relationships (Jensen, 2006).

## 1.3.5.2 Natural language processing

Natural language processing (NLP) employs parser-based techniques that involve automated analysis of sentence syntax and semantic structures to recognize meaningful relationships between concepts or entities found in biomedical text corpus. NLP relationship extraction techniques can be loosely categorized into full parser-based and shallow parser-based systems. Shallow parsers partially reconstruct the structure of an entire sentence and are based on extraction of local dependencies between phrasal components. Shallow parsers are very simple to implement, much more efficient and robust. On the contrary, full parsers consider the structure of an entire sentence and though they appear inherently complex, domain sensitive and time consuming, they can be fine-tuned to provide more accurate results (Daraselia et al., 2004; Huang et al., 2006).

A full parser-based approach was implemented in MedScan, an automated system for extracting interactions between proteins from MEDLINE abstracts (Daraselia et al., 2004). Additionally, the system is capable of extracting different types of protein function information encoded as extendable ontology with high precision. MedScan is implemented in three steps: (i) preprocessor module for tagging protein names, (ii) NLP engine composed of syntactic parser and semantic processors, and (iii) information extraction module which function as a domain-specific filter for sentence structures and present extracted information as concept graph. A total of 2976 human protein interactions have been extracted from MEDLINE abstracts published after 1988 and it achieved precision of 91%. When MedScan interactions were compared with key databases (BIND and DIP), about 96% of extracted information were novel. Interestingly,

the system achieved a recall rate of 21%. Other systems incorporating full parsing techniques are described elsewhere (Park et al., 2001; Leroy and Chen, 2002).

A foremost example of NLP shallow parser-based system was implemented in Genescence, a database housing relations extracted from MEDLINE abstracts pertaining to essential topics such as p53, AP-1 and yeast (Leroy et al., 2003). This parser is composed of semantic module for capturing abstract contents and structure module composed of cascaded finite state automata (FSA) for capturing sentence structure as well as modeling relations. Sentence structure capture involves shallow parsing closed class English words for effective detection of relations between noun phrases in biomedical text. Three cancer researchers from the Arizona Cancer Center evaluated 330 relations extracted from 26 abstracts in their area of expertise. A total of 296 relations were rightly extracted from the abstracts resulting in precision of 90% and overall average of 11 rightly extracted relations per abstract. The system is able to extract variety of relations using small number of rules and also mines more relations per abstract since relations are not restricted to specific verbs or unique entities (Leroy et al., 2003). Relationship extraction has been employed as part of the techniques to successful mine gene/protein biological functions from biomedical literature (Koike et al., 2005). In this system, a text sentence is shallowly parsed after which an adopted rule-based sentence structure analysis approach is used to extract the ACTOR (doer of action) and OBJECT (receiver of action) relationships. This system extracted from PubMed more than 190 000 gene-gene ontology (GO) relationships and 150 000 family-GO relationships for major eukaryotes. Additionally, it achieved recall and precision ranges of 54-64% and 91-94% respectively for actual functions described in abstracts. An earlier system developed based on a shallow parser (EngCG) was able to retrieve interaction between genes and gene products based on frequently seen verbs in MEDLINE abstracts (Sekimizu et al., 1998). It achieved precision rates of 90% and about 73% in identifying noun phrases and the right subject and object for each verb in text corpus respectively. Pustejovsky et al. (2002) developed a robust system for extracting inhibition relations from biomedical text. The system involves the construction of semantic automata from UMLS database to aid extraction of target relations and application of corpus analytics on MEDLINE sentences

pertaining to the target relations. Broadly, it consists of shallow parsing, relation identification and anaphora resolution modules. System evaluation achieved precision of 90%, recall 57% and partial recall 22%, consolidating its potential as an efficient corpus-based linguistic system. A pragmatic approach for the automatic extraction of information concerning gene interactions was reported (Proux et al., 2000). The approach entails the use of linguistics systems consisting of part-of-speech tagger and shallow parser, and knowledge processing module for generating semantic representation from pre-extracted syntactic dependencies. The text corpus consists of a set of 1200 sentences enriched with gene interaction information obtained from FlyBase, a database on *Drosophila melanogaster* (FlyBase Consortium, 2002).

A NLP-based tool using semantic dependency parse trees for extracting physical, genetic and regulatory relations between gene and proteins from MEDLINE abstract is RelEx (Fundel et al., 2007). This system employs dependency, part-of-speech taggers, simple rules and noun phrase-chunker. The dependency parse trees are built using Stanford Lexicalized Parser and they enhance the harnessing of complex phrasal relationships using few tree rules. RelEx identifies potential candidate relations by harvesting paths connecting pairs of nodes consisting of genes in the dependency parse trees. The system was used to mine about 150 000 relations from one million MEDLINE abstracts pertaining to gene and protein relations, and achieved 80% precision and 80% recall. RelEx is simple to implement and reproducible, can deal with complex sentences and also identity various types of interactions.

## 1.3.5.3 Pattern-matching

This approach is based on the premise that word patterns surrounding protein names embedded in text sentences can be harnessed to discover clues pertaining to the relationships amongst them. It matches part-of-speech tags obtained from segmented text sentences against pre-constructed patterns and simple matching rules. The pattern-matching system avoids the complex sentence analysis of NLP by focusing on a

restricted area of interest such as protein-protein interactions, which seem to appear as sources of flaws and decreased performance in NLP. A popular pattern-based system for automated extraction of protein-protein interaction from biological text has been described by (Ono et al., 2001). This system combines protein name lookup via named protein dictionary, surface clues on word pattern and rule-based part-of-speech tagging. The performance of this system was evaluated on text sentences using yeast (recall = 86.8% and precision = 94.3%) and *E.coli* (recall= 82.5% and precision = 93.5%) protein named dictionaries. The high rates of precision and recall achieved pinpoint to the fact that this approach can be adopted for any species-specific name protein dictionary.

SUISEKI is another system for automated detection of protein names as well as biological interactions (Blaschke and Valencia, 2001). The steps here describe the implementation of SUISEKI: (1) text corpus obtained from user query via MEDLINE is segmented and parsed by part-of-speech tagger to serve as data source. Rule-based algorithm and dictionaries are combined with data source to detect protein names, (2) sentences containing a minimum of two detected proteins are matched against predefined protein interaction description frames and (3) the resultant data is housed in a protein-protein interaction database equipped with an integrated user query interface which provide modules for analysis and protein information extraction on synonyms and functional descriptions. The efficiency of SUISEKI is demonstrated by the fact that majority of identified interactions have higher chances of being true interactions. A total of 4657 interactions were extracted from 5283 abstracts pertaining to cell cycle corpus. The system achieved between 80% to 50% precision range for high scoring interactions to low scoring ones, and relatively highly recall of more than 70%.

The HypertenGene system automatically extracts etiological relations among genes and diseases in text sentences by combining the advantages of machine learning models and pattern matching including part-of-speech (POS) patterns and extracted phrase chunk features (Tsai et al., 2009). Both NLP and pattern match-based systems display deficiencies in their output results. Even though NLP defy the simplicity and robust character of pattern matching systems, pattern-matching systems overlooks coordinative,

conjunctive and appositive grammatical structures of text (Huang et al., 2006). A hybrid system has been developed as response to the above deficiencies, by integrating a shallow parser to improve remarkably the performance of pattern matching to extract relations between proteins from biomedical texts (Huang et al., 2006). This system achieved an average F-score of 80% and 66% on individual verbs and all verbs respectively. The shallow parser enabled the system to achieve 7% increase in both F-score and precision when compared to the traditional pattern matching-based systems. As part of the implementation, the shallow parser analyzes the coordinative and appositive structures, followed by the application of pattern matching on segmented long sentences to enhance matching precision. Prior to matching, the coordinative structures are encapsulated and then unfolded after extracting. A dynamic programming algorithm computes unique patterns by aligning relevant sentences and key verbs that describe protein interactions and thereafter, a matching algorithm extract the interaction relations (Huang et al., 2004a).
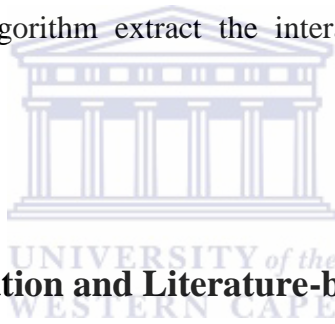
## 1.3.5.4 Hypothesis generation and Literature-based discovery (LBD)

Relationship extraction deals with harvesting relationships between biomedical entities explicitly stated in a text corpus whiles hypothesis generation aims to unravel implicit relationships existing between disjoint sets of literature pertaining to biomedical concepts. The implicit relations could lead to unraveling of plausible novel knowledge/discovery, which is potentially testable and worthy of further exploration. Swanson originally popularized the concept of text-mined hypothesis in his famous derivation of fish oil/Raynaud's diseases hypothesis. He used disjointed sets of literature pertaining to both fish oil and Raynaud's diseases to propose a therapeutic relationship between them. The fish oil/Raynaud's disease hypothesis relied on the premises that: (a) fish oil decreases platelet activities, vascular reactivity, and blood viscosity, and (b) these same physiological changes ameliorate Raynaud's symptoms (Swanson, 1986). This "premier" hypothesis was confirmed later in a study involving the clinical effects of fish-oil fatty acid ingestion in patients with Raynaud's symptoms (DiGiacomo et al., 1989),

where fish oil was demonstrated to have beneficial effects by delaying the onset of Raynaud's disease complications. Similar hypothesis were proposed for Migraine and Magnesium, as well as Somatomedin C and Arginine (Swanson et al., 1990). The above hypothesized discourses influenced the proposition of the concept known as the complementary structures in disjoint science literatures (CSD) (Swanson, 1991). Complementary means the relationship between two separate scientific arguments, which when consolidated together could yield inferences or insights that are not obvious in the separate arguments. Disjoint literature refers to the scenario when literatures are not co-cited or co-mentioned and have no articles in common. The CSD has been simplified in a concept widely referred to as Swanson's ABC discovery model (Figure 1.9), which stipulates that if a concept A associate with B, and B associate with C, then there could be implicit relationship between concepts A and C (Weeber et al., 2003). A current approach assigns statistical weights to the strength of the hidden relationship (Frijters et al., 2010). The ABC model is further reconstructed separately into hypothesis generating approach referred to as the open discovery system (Figure 1.10) and a testing approach known as the closed discovery system (Weeber et al., 2001; Figure 1.11). The open discovery system is analogous to Swanson's Fish/Raynaud's disease hypothesis, where a starting concept that connects to a linking concept, also connect to the target concept. The open discovery approach leads to hypothesis generation that originate from the quest to answer a specific scientific problem. The closed discovery system involves the use of a bridging concept to establish a link between two other concepts that do not directly co-occur. This also involves the testing and explanation of already generated hypothesis. For example, starting from a disease concept C and substance A, the researcher can intuitively find common intermediate B-terms. If reasonable number of "pathways" are found between A and C, then the propensity for it to be a potentially insightful hypothesis is high (Figure 1.11). The differences between the two discovery approaches are that, literatures pertaining to C and B are interrogated in the open discovery whiles literatures on C and A are interrogated in the closed discovery system. The open and closed discovery processes have been simplified in Figure 1.12 (Cohen et al., 2010a).
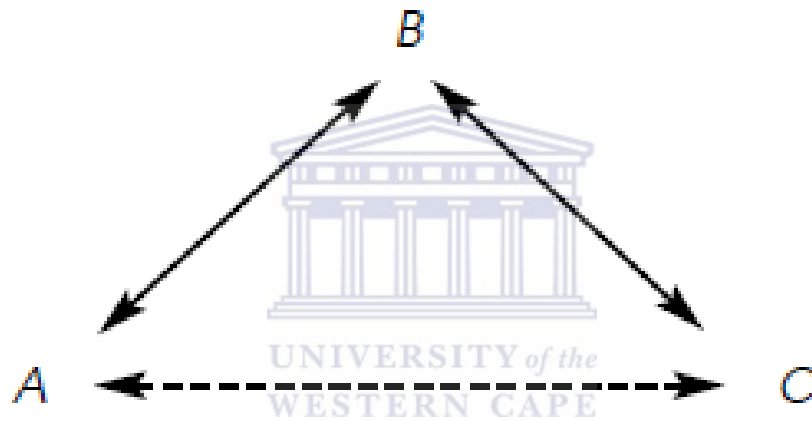
**FIGURE 1.9.** THE SWANSON'S ABC DISCOVERY MODEL.
The solid arrows denoted by AB and BC represent explicit relationships. The dashed arrows denoted by AC represent an implicit relationship and may serve as a potential source of novel hypothesis (Weeber et al., 2003).
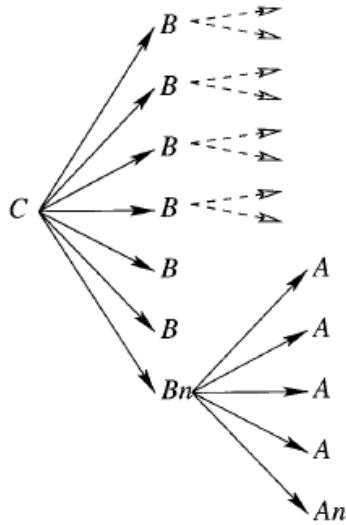
**FIGURE 1.10.** OPEN DISCOVERY SYSTEM.
The starting, linking and target concepts are CBA respectively. Solid arrows indicate "pathways" worthy of further exploration for discovery, and the dashed ones are "pathways" that do not lead to discoveries (Weeber et al., 2001).
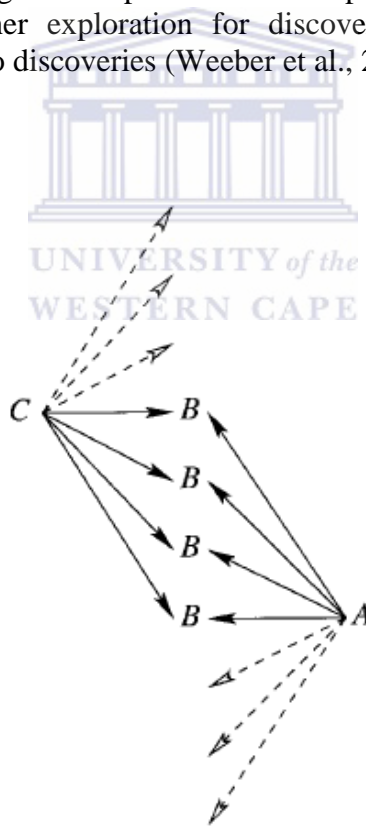


**FIGURE 1.11.** CLOSED DISCOVERY SYSTEM.
The starting concepts are C and A. Concept Bs result from C and A overlapping. Solid arrows indicate "pathways" worthy of further exploration for discovery, and the dashed ones are "pathways" that do not lead to discoveries (Weeber et al., 2001).
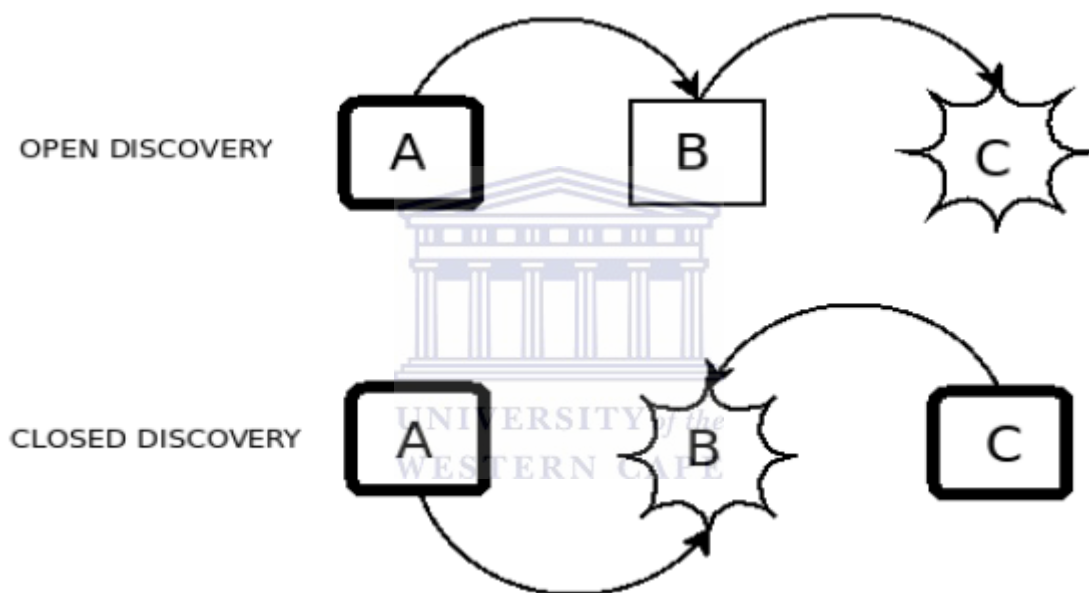
**FIGURE 1.12**. A SIMPLIFIED VERSION OF THE OPEN AND CLOSED DISCOVERY SYSTEMS.
In both instances, the starting concept A for generating discoveries has thick borders, and
the target or endpoint concepts are surrounded by starbursts (Cohen et al., 2010a).

A typical biological example of closed discovery system is illustrated here. Among patients suffering from multiple myeloma treated with thalidomide, some experienced improvement in chronic hepatitis C. A closed system approach will involve trying to unravel the potential mechanisms underpinning how thalidomide treats chronic hepatitis C (Weeber et al., 2005).

Although original implementation of LBD involved overwhelming human intervention, current approaches are augmented with automated systems developed using machine learning/statistical models and/or NLP techniques. LBD is defined as computer-aided generation of functionally relevant hypotheses with potential for further scientific exploration. A number of LBD systems have been implemented using co-occurrence-based approaches (Cohen and Hersh, 2005; Weeber et al., 2005). Additionally, LBDs have been utilized to simulate already published or described novel discoveries (Weeber et al., 2000; Weeber et al., 2003; Kwofie et al., 2011). The limitations of co-occurrence-based systems are that they sometimes overlook semantic nature of the inferred relations and may not offer any concrete explanation for the generated hypothesis, and users are required to scrutinize manually the myriad of texts accompanying the suggested hypothesis (Hristovski et al., 2006). The performance of co-occurrence-based approach has been improved in BITOLA (Hristovski et al., 2005) by integrating with NLP derived semantic predications extracted from SemRep (Rindflesch and Fiszman, 2003) and BioMedLee (Lussier et al., 2006). This integrated approach provides details and support explanation for inferred discoveries or relationships. The system exhibited the potential to achieve fewer numbers of false positive results. Another LBD system, EpiphaNet (Cohen et al., 2010a), has been implemented using predication-based semantic indexing (PSI) and reflective random indexing (RRI). RRI is a scalable method for detecting implicit connections (Cohen et al., 2010b). Additionally, RRI methods augment the ability of EpiphaNet to harvest associations between concepts or terms that do not directly co-occur. The predications used in EpiphaNet were also obtained from SemRep (Rindflesch and Fiszman, 2003) by mining abstracts and MEDLINE titles. In simple terms, predications are also referred to as "object relation object triplets", for example, "merlot IS-A wine". PSI enables the quick extraction of strongly associated concepts based on

their distribution across all of the predications present in SemRep (Cohen et al., 2010a). A similar LBD approach which does not depend on direct co-occurrence of concepts within abstracts of journal articles was used to generate a ranked hit list of genes related to periodontitis and atherosclerosis (Hettne et al., 2007). This approach generates concept vectors that represent the textual context surrounding a specific concept. Entities sharing most common concepts have much higher likelihood of being associated with each other and likely to be assigned higher association scores. Association based on overlapping or sharing of concepts has also been implemented in CoPub, a LBD system (Frijters et al., 2010).

The future potential of text mining systems, including LBDs far outweighs their limitations for the simple reason that information retrieval/extraction systems are still considered as potent approaches that can be used to glean knowledge from enormous amounts of published texts. Irrespective of the accuracy or statistical scores assigned to generated hypothesis, interpretive exploration of biomedical literature can still serve to augment the logic underpinning the discovery.
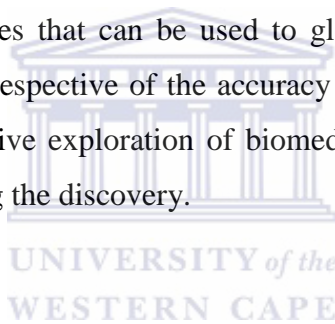
## 1.4 Protein-Protein Interactions

A typical pathway consists of a series of interconnected cellular events between biomolecular entities such as proteins, genes and metabolites (Chowbina et al., 2009). Furthermore, it is estimated that 80% of proteins do not operate alone but in complexes, and protein-protein interactions (PPIs) constitute a part of a bigger cellular network (Berggård et al., 2007). Therefore cataloging PPIs is key for the study of protein interactome to enhance current understanding of regulatory mechanisms.

## 1.4.1 Efforts towards harnessing predicted protein-protein interactions

Section 1.3.5 of this thesis deals comprehensively with text mining approaches and tools used in extracting PPI information from published biomedical texts. Efforts towards developing robust and efficient text mining extraction models were given a boost under BioCreative II (Krallinger et al., 2008), a community platform that dealt with tasks for evaluating algorithms employed in extracting PPI. The task assigned to the teams were (i) identifying protein interaction relevant articles; (ii) harvesting and normalizing protein interaction pairs; (iii) cataloging experimental methods used in detecting interactions; and (iv) retrieving the blocks of text reporting the protein interactions. Some of the difficulties encountered by the task teams were: conversion of full text article formats, detecting, annotating and cross referencing of PPI to external databases. Other limitations were in the field of protein name normalization, restricting PPIs elucidation to co-occurrence techniques and the complexities involved in distinguishing between novel, already known or experimentally determined PPI. Nevertheless, BioCreative II was useful because benchmark datasets for both training and testing of models were made available for comparing the performance of various algorithms and it also provided standardized approaches for comparing text-mining procedures with manual annotation of PPIs (Chatr-aryamontri et al., 2008).

Other computational methods used for modeling and predicting PPI are based on co-evolution (Pazos et al., 2008; Sharon et al., 2009; Lewis et al., 2010), combination of structural information and sequence pattern in protein interfaces (Aytuna et al., 2005; Espadaler et al., 2005; Keskin et al., 2008), and domain-domain interactions (Singhal and Resat, 2007; Yellaboina et al., 2011). To support current usage of machine learning models for predicting PPIs (Chen et al., 2008; Martin et al., 2008; Guo et al., 2010), a number of resources are being developed to provide protein specific features as training datasets. KUPS (The University of Kansas Proteomics Service) provides high quality benchmark data for developing and evaluating PPI prediction algorithms (Chen et al., 2011). KUPS provides lists comprising of interacting protein pairs (IPPs) as positive data and non-interacting protein pairs (NIPs) as negative data, as well their features for each data point (instance). While protein interaction datasets are currently readily available for

downloads from many databases but obtaining non-interaction datasets with computed machine learning features used to be a difficult task. Negatome (Smialowski et al., 2010) and GRIP (Browne et al., 2009) are additional web-based resources providing non-interacting protein pairs for mammalian models (experimentally supported) and *Saccharomyces cerevisiae* (both IPPs and NIPs) respectively. Even though both Negatome and GRIP presently do not provide computed features, they do in fact augment existing PPI resources. Another community platform that provide benchmark dataset for assessing the quality of predicted PPI through structural analysis of protein complexes is the Critical Assessment of PRedicted Interactions (CAPRI) experiment which started in 2000 (Janin, 2005). These efforts have led to the introduction of two scoring functions based on the similarity of the interfaces of the complex, namely interfacial Template Modeling score (iTM-score) and Interface Similarity score (IS-score). The iTM-score calculates the geometric distance between the protein interfaces whiles the IS-score calculates residue-residue contact similarity plus their geometric similarity (Gao and Skolnick, 2011). Plewczyński and Ginalski (2008) have comprehensively reviewed some of the approaches and challenges involved in predicting PPIs, and additionally have supported the consensus approach adopted by Sen et al. (2004). Sen et al. (2004) employed a metadata mining approach by combining support vector models, protein structural threading, phylogenetic analysis to predict conserved residues on protein interfaces, and the Conservatism of Conservatism method of Mirny and Shakhnovich (1999). It was illustrated that the combinatorial approach yielded improved predictions over the individual methods.

A trend has emerged where some databases house both experimental and computationally derived PPI obtained from multiple data sources. HAPPI, the Human Annotated Protein-Protein Interactions contains data integrated from complementary sources consisting of the HPRD, BIND, MINT, STRING, and OPHID databases (Chen et al., 2009). Experimentally derived PPIs data from both low and high-throughput techniques are obtained from BIND and HPRD; literature curated PPI from BIND; text-mined PPI from STRING; and computationally predicted PPI from both STRING and OPHID. HAPPI assigns confidence scores ranging from 1 to 5 to enable users gauge the reliability of the

interactions and also enhance the generation of high confidence interaction networks. The interaction data in HAPPI is also enriched with computed annotations, including biological pathways, gene functions, protein families, protein structures, sequence features, and literature sources.

As mentioned earlier, STRING, a Search Tool for the Retrieval of Interacting Genes also provides comprehensive and reliable PPI data composed of both experimentally determined and predicted interaction information (Szklarczyk et al., 2011). As in the case of HAPPI, STRING also provides confidence scores for judging reliability of interactions and interacting partners are mutually cross-referenced via HTML with a number of popular knowledgebases, including UniProt (UniProt Consortium, 2010), SMART (Letunic et al., 2009), GeneCards (Safran et al., 2010) and SwissModelRepository (Kiefer et al., 2009). In STRING, interactions are not restricted to direct, physical interactions between two proteins but proteins can also be linked when they exhibit a genetic interaction or are reported to catalyze subsequent steps in a regulatory or metabolic pathway. Most associations between proteins inferred using prediction algorithms cannot be defined precisely in the context of their mode of interaction, nor cellular conditions under which they occur, and may be viewed rather as functional associations (Szklarczyk et al., 2011). Other resources providing predicted PPI data are PIPs (McDowall et al., 2009), IntNetDB (Xia et al., 2006), POINT (Huang et al., 2004b), HPID (Han et al., 2004a), PreSPI (Han et al., 2004b), PIPE (Pitre et al., 2006) and OPHID (Brown and Jurisica, 2005).

## 1.4.2 Towards the storage and utilization of curated protein-protein interactions

Experimentally derived PPI data is primarily obtained from: (i) literature-curated interactions and (ii) directly from both low throughput and large scale screening of pair-wise physical interactions (de Chassey et al., 2008). Literature-curated interaction datasets are hypothesis-driven and may allow the determination of biological functions of

the interacting proteins from the actual study while high-throughput datasets are discovery-based but are not geared towards inferring of functions of the interacting proteins (Cusick et al., 2009). High-quality binary interaction information is warehoused in public databases for protein interactome exploration. The notable detection and analysis methods (Phizicky and Fields, 1995) for protein-protein interactions are: (i) physical methods to select and detect proteins that bind another protein (e.g. affinity chromatography, affinity blotting, immunopreciptation and cross-linking); (ii) library-based methods (e.g. protein probing, phage display and two-hybrid system (high-throughput technique)); and (iii) genetic methods (e.g. extragenic suppressors, synthetic lethal effects and overproduction phenotypes). Techniques for verifying PPIs are confocal microscopy for intracellular colocalization of proteins, coimmunoprecipitation, surface plasmon and spectroscopic studies (Berggård et al., 2007). Additionally, 3-dimensional X-ray crystallography and nuclear magnetic resonance structural analysis are used to elucidate PPI. Comprehensive reviews detailing the merits and demerits of the various interaction detection methods are discussed elsewhere (Lalonde et al., 2008).

Since curated interactions and their associated metadata can be obtained from multiple database sources, a concerted effort must be made to provide standardized and comprehensive procedures for judging the quality of the data sources. Metamining databases consist of consolidated PPIs obtained from multiple source databases. Metamining database authors may choose not to integrate some data from their sources based on in-house curation quality checks. Therefore a user queried interaction may fail in the metamining database, even though it is present in the source data (Plewczynski and Klingström, 2011). Furthermore, each source database may assign its own unique identifiers to interaction metadata, necessitating the need to unify multiple database identifiers pertaining to the same query. The "Good Interaction Data Metamining Practice" standard (GIDMP) has been proposed (Plewczynski and Klingström, 2011) and when adopted could provide PPI warehousing community with: (i) a standardized approach to judging the statistics made available by each metamining database, thus enhancing user satisfaction; (ii) a stable contact point for each database, enabling the smooth transition of statistics; and (iii) a fully automated system, enhancing time- and

cost-effectiveness. It is recommended that databases complying with GIDMP standards should provide a page titled Source Databases (SourceDB) composed of a table consisting of columns that capture relevant information and statistics. The table would consist of seven columns arranged in the order: name of source database, link to resource, notes, version, latest version, and coverage. The first five columns are static data provided by the metamining database authors and the last two are obtained from the source database. It is expected that a simple script would be embedded within the www pages of both metamining and source databases to automatically update the last two columns thereby providing synchronized information including statistics.

An empirical approach has been employed to quantitatively evaluate the level of agreement across major public databases concerning curated PPI data (Turinsky et al., 2010). The analysis was carried out on a global landscape of PPI data that was consolidated from nine key databases that are known to be composed primarily of curated experimentally detected physical interactions: BIND, BioGRID, CORUM, DIP, IntAct, HPRD, MINT, MPact and MPPI. The iRefIndex (Razick et al., 2008) and iRefWeb (Turner et al., 2010) were used for both data consolidation and analysis. The analysed data comprised of 271,716 distinct physical interactions involving 7,449 proteins associated with 1324 different organism-taxonomy identifiers and were gleaned from 42,651 publications. Computationally predicted and genetic interactions representing phenotype alterations produced by the mutation/deletion of genes were excluded from the analysis. Annotated PPI data derived from the same journal article but co-cited by different databases were evaluated for agreement using similarity scores. After quantifying agreement between curated interactions from a total of 15,471 publications, the results revealed that on average, two databases fully agree on 42% of the interactions and 62% of the proteins curated from the same publication. It further revealed that a sizeable portion of the measured differences could be attributed divergent annotations of organism or splice isoforms, different organism focus and alternative representations of multi-protein complexes. This study highlights the impact of the divergent in-house curation policies of data providers (Turinsky et al., 2010) and is a motivator for the PPI research community to adopt data standardization practices. To buttress above concerns,

evaluation of existing curations of protein interaction experiments reported in journal articles revealed that curation could be error-prone and that the quality may be of lower standard than what is already assumed (Cusick et al., 2009).

A large amount of PPI data emanating from both small- and large-scale experiments are scattered across databases, web pages or in journal articles prompting the need to unify and standardize details describing such interactions. The Human Proteome Organization Proteomics Standards Initiative Molecular Interactions (HUPO PSI-MI) format XML1.0 was established as a single, unified format by which molecular interactions, particularly PPIs are presented (Hermjakob et al., 2004). Notable interaction data producers and providers, including BIND, DIP, IntAct, MINT, MIPS, Hybrigenics and STRING, jointly formulated this community standard data model for representation and exchanging of PPI data. Included in the PSI-MI schema are the controlled vocabularies (CVs), which are used to standardize the meaning of data objects. The continuous use of CVs prevents data providers from using synonyms or name variants repeatedly and also allow consistent interpretation of terms among multiple data producers and users. The CVs are hierarchically structured and consist of higher and lower level descriptors assigned to parents and child terms. This organization enables data to be annotated to an appropriate level of granularity and enhance easy retrieval of information by search tools via the child and parents terms when applicable. The CVs are maintained in the Open Biological and Biomedical Ontologies web pages (Smith et al., 2007) accessible via http://www.obofoundry.org/ and can also be retrieved using the EBI Ontology Lookup Service (Barsnes et al., 2010) accessible via http://www.ebi.ac.uk/ontology-lookup/. For illustration purposes, a binary protein interaction between two protein molecules inferred using yeast two hybrid as detection method is assigned the PSI-MI term name "two hybrid" and term ID "MI:0018". Currently, PSI-MI has been updated to PSI-MI XML2.5 and MITAB2.5 formats to facilitate data exchange among repositories and users without loss of information. The updated version expanded the scope of the described interactions to cover a broader range of molecular types, including nucleic acids, chemical entities, and molecular complexes. PSI-MI XML2.5 provides detailed description about each molecular interaction, including the biological role of each interacting molecule,

interacting domains, and the kinetic parameters governing the interactions. MITAB2.5 provides only minimal interaction information in a tab-delimited format for easy systems configuration through fast Perl parsing or loading into Microsoft Excel (Kerrien et al., 2007).

It is extremely difficult for curators to extract all the essential information describing PPI and associated metadata from published journal articles when authors of such articles provide sparse details describing the interactions. Even though enormous amount of energy may be spent in trying to guess the missing data in the poorly described interaction, this approach is still fraught with error. According to anecdotal estimates, about 70% of total curation time is spent by curators on disambiguating molecule identifiers (Orchard et al., 2005). The minimum information required for reporting a molecular interaction experiment (MIMIx), a community effort seek to provide a checklist on information to supply when describing experimental molecular interaction data in a journal article, displaying on a web page or incorporating into PPI repositories (Orchard et al., 2005). The MIMIx checklist provides details on: (i) interaction submission profile, including author contact addresses and publication identifier; (ii) experimental setup including host organism taxonomy and interactions in PSI-MI ontology vocabularies; (iii) interaction participant list; and (iv) confidence scores assigned to the interactions.

Although PPI databases constitute a smaller portion of the total number of reported biological databases, it is important that they comply with global approaches aimed at promoting consistency and interoperability among knowledgebases as well as semantic and syntactic standards. BioDBCore, a community-defined, uniform, generic description of the core attributes of biological databases seek to provide guidelines that would help achieve the above (Gaudet et al., 2011). BioDBCore would make available a catalogue of database attributes contained in a checklist to facilitate the sharing of rich information among diverse research communities. Researchers may have easy access to information outside their area of expertise by way of browsing the descriptors for various biological databases. This approach may enable developers to avoid repeating of routine mistakes

committed when implementing new resources. The proposed BioDBCore descriptors are: database name, main resource URL, contact information, date resource established, conditions of use (free, or type of license), scope (data types captured, curation policy, and standards used), Standards (data formats and terminologies), taxonomic coverage, data accessibility/output options, data release frequency, versioning policy and access to historical files, documentation availability, user support options, data submission policy, relevant publications, resource's Wikipedia URL, and tools available. BioDBCore is in line with the MIBBI project, which aims to promote coherent minimum reporting guidelines for biological and biomedical investigations (Taylor et al., 2008). Embracing the BioDBCore project could make PPI databases accessible to the greater biological community and enhance cross integration of PPI data with other high-throughput data for exploration.

## 1.4.3 Available resources and tools for exploring PPI data

Online repositories do not merely warehouse PPI information but also provide means for data exploration, comparison and integration into other secondary databases. Key among them is iRefIndex (Razick et al., 2008) and its update iRefWeb (Turner et al., 2010). Currently, iRefWeb is made up of data consolidated from 10 public PPI databases, comprising of BIND, BioGRID, CORUM, DIP, IntAct, HPRD, MINT, MPact, MPPI and OPHID. Notably, iRefWeb provides global view of PPI exploration and also gives users the option to examine discrepancies among same PPI information provided by multiple databases and be able to choose interactions that closely resemble the reported interaction in the journal article. Additionally, statistical summaries of interactions across data sources are provided and it also enable evaluation of reliability of interactions based on simple modalities such as the number of journal articles reporting interactions and whether the detection methods were low or high throughout techniques. The iRefIndex database originally provided a unifying index for retrieving PPI data and assigned a key for each PPI record and a key for each interacting partner. Using primary sequences of the interacting proteins, their taxonomy identifiers and the Secure Hash Algorithm, users

can generate these keys. The indexing is useful because redundant or equivalent PPIs are assigned as a single group. Additionally, the system enables data integrity to be verified and helps to identify problematic PPI entries across different repositories. Other web servers or software plug-ins that allow retrieval of consolidated interaction datasets and generation of interaction networks are APID (Prieto and De Las Rivas, 2006), APID2NET (Hernandez-Toro et al., 2007), MiMI (Tarcea et al., 2008), UniHI (Chaurasia et al., 2007), DASMI (Blankenburg et al., 2009), ConsensusPathDB (Kamburov et al., 2011), BisoGenet (Martin et al., 2010), PIANA (Aragues et al., 2006), BioNetBuilder (Avila-Campillo et al., 2007), cPath (Cerami et al., 2006) and BioGRID access tools, namely BioGRID REST Service, BiogridPlugin2 and BioGRID WebGraph (Winter et al., 2011).

Users of PPI data aim to identify multiple-drug targets within pathogenic protein-protein interaction networks (PPINs) by way of perturbing known pathways or complexes (Hormozdiari et al., 2010) and also by analyzing protein-protein interaction binding pockets (Meireles et al., 2010). It is also interesting to note that topological and clustering analysis of PPI networks have been employed to predict new interactions (Bu et al., 2003; Sen et al., 2006). Therefore software systems and web servers that enable graph theoric analysis and modeling of both experimentally and predicted interactions are of optimal importance. Some of the available tools for analysis and graph theoric explorations of PPI networks are BiNoM (Zinovyev et al., 2008), NetworkAnalyzer (Assenov et al, 2008), NeAT (Brohée et al., 2008) and frequently used software libraries, including JUNG- the Java Universal Network/Graph Framework (http://jung.sourceforge.net/), LEDA (http://algorithmic-solutions.com/leda/index.htm), NetworkX (https://networkx.lanl.gov/) and yFiles (http://www.yworks.com/en/index.html). PPI visualization tools or platforms are Cytoscape (Smoot et al., 2011), ProViz (Iragne et al., 2005), PIMWalker (Meil et al., 2005), VisANT (Hu et al., 2009), Osprey (Breitkreutz et al., 2003), and PATIKA (Demir et al., 2002) and its web version PATIKAweb (Dogrusoz et al., 2006). A comprehensive review on tools for visualizing biological networks is presented elsewhere (Suderman and Hallett, 2007).

## 1.5 THESIS RATIONALE

From the aforementioned (Section 1.2) with particular emphasis on HCV pathobiology and therapeutic challenges, it is imperative that diverse approaches must be adapted to outwit a "skillful and deceitful medical foe" such as HCV and its accompanying pathobiological ramifications. Provision of resources that separately integrate text mining and protein-protein interactions can go a long way to aid in the search for therapeutic drugs and possibly augment efforts to generate functional hypotheses that can be further investigated for the development of non-invasive diagnostic biomarkers and novel therapeutics. Computational text mining has been highlighted as an important component of the modern-day drug discovery pipelines (Yao et al., 2009, 2010). It has particular utility in the following stages: analysis of disease mechanisms and "disease genes", genomics, structure-based compound design, analysis of mechanisms of drug actions, toxicology, pharmcovigilance, and drug repurposing. Similarly, text mining and PPI feature prominently as part of workflows for the discovery of therapeutic targets or diagnostic biomarkers (Figure 1.13). This thesis aims to develop freely available HCV-focused integrated online resources providing enriched biomedical knowledge as baseline information for further exploration and to also aid in computational prediction of functional hypothesis of therapeutic relevance.
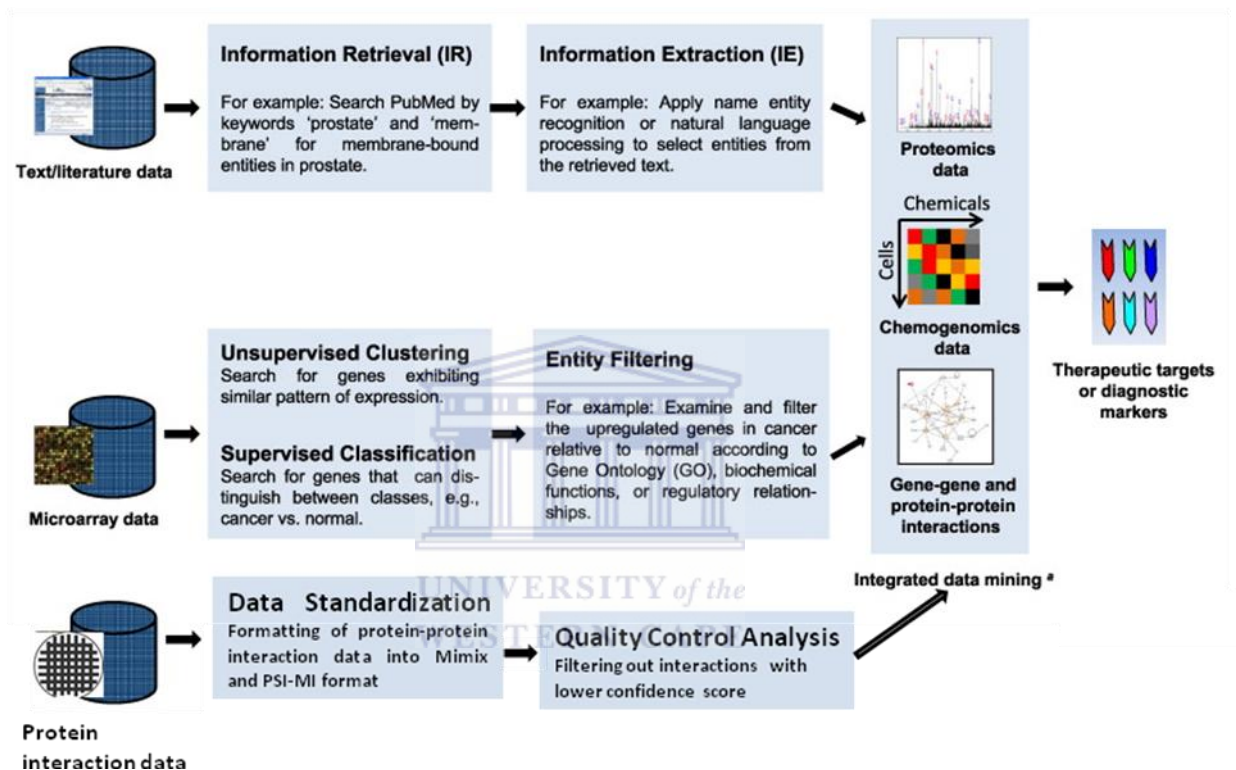
**FIGURE 1.13.** A DRUG DISCOVERY WORKFLOW.
The workflow incorporating text mining and protein interaction data combined with other high-throughput data (e.g. microarray data mining) to aid discovery of therapeutic targets or diagnostic biomarkers (modified from Yang et al., 2009). This workflow supports the data mining approach adopted in this research, which involved harnessing of integrated data consisting of protein-protein interactions and HCV related hepatocellular carcinoma associated genes identified from proteomic and microarray experiments. Text mining approach was utilized in proposing literature-based discoveries of therapeutic relevance.

### 1.5.1 Aims and Objectives

1. Development of HCV-focused web-based text mining resource
2. Development of HCV online protein knowledgebase
3. Harnessing the myriad of data in the developed knowledgebases to facilitate generation of functional hypothesis of therapeutic relevance.

### 1.5.2 Thesis outline

This thesis consists of four chapters:

Chapter 1:

- Gives a brief over view on: (i) HCV biology and therapeutic challenges, (ii) general text mining principles and applications, (iii) extraction and utilization of protein-protein interactions, and (iv) thesis rationale and aims (as outlined in this section 1.5).
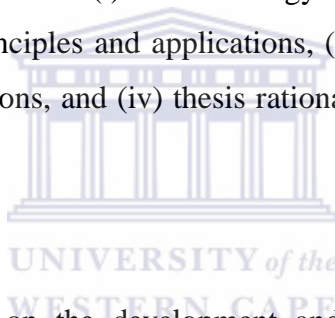
Chapter 2

- Provides an overview on the development and implementation of an HCV-customized web-based text mining resource known as the Dragon Exploratory System on Hepatitis C Virus (DESHCV) (Kwofie et al., 2011). It further describes how DESHCV has been used to reproduce already published literature-based discovery and also generate novel hypothesis relevant to the current efforts geared towards HCV combination therapy search.

Chapter 3

- Describes the development and implementation of HCVpro, Hepatitis C Virus Protein Interaction Database. And demonstrate how the incorporated data in HCVpro could be used to generate baseline hypothesis relevant to unraveling potential diagnostic markers and therapeutic targets.

Chapter 4

- Discusses the relevance of this work towards ongoing HCV research efforts and provides the conclusion of this thesis. Additionally, it discusses the limitations and proposes future work.

## 1.6 References

Agnello, V., Abel, G., Elfahal, M., Knight, G. B., & Zhang, Q. X. (1999). Hepatitis C virus and other flaviviridae viruses enter cells via low density lipoprotein receptor. *Proc Natl Acad Sci U S A, 96*(22), 12766-12771.

Alako, B. T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S., Rullmann, T., et al. (2005). CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics, 6*, 51.

Aphinyanaphongs, Y., & Aliferis, C. F. (2003). Text categorization models for retrieval of high quality articles in internal medicine. *AMIA Annu Symp Proc*, 31-35.

Aphinyanaphongs, Y., Tsamardinos, I., Statnikov, A., Hardin, D., & Aliferis, C. F. (2005). Text categorization models for high-quality article retrieval in internal medicine. *J Am Med Inform Assoc, 12*(2), 207-216.

Aragues, R., Jaeggi, D., & Oliva, B. (2006). PIANA: protein interactions and network analysis. *Bioinformatics, 22*(8), 1015-1017.

Assenov, Y., Ramirez, F., Schelhorn, S. E., Lengauer, T., & Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics, 24*(2), 282-284.

Avila-Campillo, I., Drew, K., Lin, J., Reiss, D. J., & Bonneau, R. (2007). BioNetBuilder: automatic integration of biological networks. *Bioinformatics, 23*(3), 392-393.

Aytuna, A. S., Gursoy, A., & Keskin, O. (2005). Prediction of protein-protein interactions by combining structure and sequence conservation in protein interfaces. *Bioinformatics, 21*(12), 2850-2855.

Baldo, V., Baldovin, T., Trivello, R., & Floreani, A. (2008). Epidemiology of HCV infection. *Curr Pharm Des, 14*(17), 1646-1654.

Barsnes, H., Cote, R. G., Eidhammer, I., & Martens, L. (2010). OLS dialog: an open-source front end to the ontology lookup service. *BMC Bioinformatics, 11*, 34.

Bartenschlager, R., Frese, M., & Pietschmann, T. (2004). Novel insights into hepatitis C virus replication and persistence. *Adv Virus Res, 63*, 71-180.

Barth, H., Schafer, C., Adah, M. I., Zhang, F., Linhardt, R. J., Toyoda, H., et al. (2003). Cellular binding of hepatitis C virus envelope glycoprotein E2 requires cell surface heparan sulfate. *J Biol Chem, 278*(42), 41003-41012.

Bartosch, B., Vitelli, A., Granier, C., Goujon, C., Dubuisson, J., Pascale, S., et al. (2003). Cell entry of hepatitis C virus requires a set of co-receptors that include the CD81 tetraspanin and the SR-B1 scavenger receptor. *J Biol Chem, 278*(43), 41624-41630.

Berggard, T., Linse, S., & James, P. (2007). Methods for the detection and analysis of

protein-protein interactions. *Proteomics, 7*(16), 2833-2842.

Blankenburg, H., Finn, R. D., Prlic, A., Jenkinson, A. M., Ramirez, F., Emig, D., et al. (2009). DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics, 25*(10), 1321-1328.

Blaschke, C., & Valencia, A. (2001). The potential use of SUISEKI as a protein interaction discovery tool. *Genome Inform, 12*, 123-134.

Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O., & Eisenberg, D. (2004). Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol, 5*(5), R35.

Breitkreutz, B. J., Stark, C., & Tyers, M. (2003). Osprey: a network visualization system. *Genome Biol, 4*(3), R22.

Brohee, S., Faust, K., Lima-Mendez, G., Sand, O., Janky, R., Vanderstocken, G., et al. (2008). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res, 36*(Web Server issue), W444-451.

Brown, K. R., & Jurisica, I. (2005). Online predicted human interaction database. *Bioinformatics, 21*(9), 2076-2082.

Browne, F., Wang, H., Zheng, H., & Azuaje, F. (2009). GRIP: A web-based system for constructing Gold Standard datasets for protein-protein interaction prediction. *Source Code Biol Med, 4*, 2.

Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., et al. (2003). Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res, 31*(9), 2443-2450.

Caillot, F., Hiron, M., Goria, O., Gueudin, M., Francois, A., Scotte, M., et al. (2009). Novel serum markers of fibrosis progression for the follow-up of hepatitis C virus-infected patients. *Am J Pathol, 175*(1), 46-53.

Cerami, E. G., Bader, G. D., Gross, B. E., & Sander, C. (2006). cPath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics, 7*, 497.

Chang, J. T., Schutze, H., & Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics, 20*(2), 216-225.

Chatr-aryamontri, A., Kerrien, S., Khadake, J., Orchard, S., Ceol, A., Licata, L., et al. (2008). MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. *Genome Biol, 9 Suppl 2*, S5.

Chaurasia, G., Iqbal, Y., Hanig, C., Herzel, H., Wanker, E. E., & Futschik, M. E. (2007). UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res, 35*(Database issue), D590-594.

Chaussabel, D., & Sher, A. (2002). Mining microarray expression data by literature profiling. *Genome Biol, 3*(10), RESEARCH0055.

Chen, D., Muller, H. M., & Sternberg, P. W. (2006). Automatic document classification of biological literature. *BMC Bioinformatics, 7*, 370.

Chen, J. Y., Mamidipalli, S., & Huan, T. (2009). HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics, 10 Suppl 1*, S16.

Chen, X. W., Han, B., Fang, J., & Haasl, R. J. (2008). Large-scale Protein-Protein Interaction prediction using novel kernel methods. *Int J Data Min Bioinform,*

*2*(2), 145-156.

Chen, X. W., Jeong, J. C., & Dermyer, P. (2011). KUPS: constructing datasets of interacting and non-interacting protein pairs with associated attributions. *Nucleic Acids Res, 39*(Database issue), D750-754.

Chevaliez, S., & Pawlotsky, J. M. (2006). HCV Genome and Life Cycle.

Choo, Q. L., Kuo, G., Weiner, A. J., Overby, L. R., Bradley, D. W., & Houghton, M. (1989). Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. *Science, 244*(4902), 359-362.

Chowbina, S. R., Wu, X., Zhang, F., Li, P. M., Pandey, R., Kasamsetty, H. N., et al. (2009). HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics, 10 Suppl 11*, S5.

Chun, H. W., Tsuruoka, Y., Kim, J. D., Shiba, R., Nagata, N., Hishiki, T., et al. (2006). Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. *Pac Symp Biocomput*, 4-15.

Codran, A., Royer, C., Jaeck, D., Bastien-Valle, M., Baumert, T. F., Kieny, M. P., et al. (2006). Entry of hepatitis C virus pseudotypes into primary human hepatocytes by clathrin-dependent endocytosis. *J Gen Virol, 87*(Pt 9), 2583-2593.

Coelho, L. P., Ahmed, A., Arnold, A., Kangas, J., Sheikh, A. S., Xing, E. P., et al. (2010). Structured Literature Image Finder: Extracting Information from Text and Images in Biomedical Literature. *Lect Notes Comput Sci, 6004*, 23-32.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Brief Bioinform, 6*(1), 57-71.

Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. *PLoS Comput Biol, 4*(1), e20.

Cohen, T., Schvaneveldt, R., & Widdows, D. (2010b). Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections. *J Biomed Inform, 43*(2), 240-256.

Cohen, T., Whitfield, G. K., Schvaneveldt, R. W., Mukund, K., & Rindflesch, T. (2010a). EpiphaNet: An Interactive Tool to Support Biomedical Discoveries. *J Biomed Discov Collab, 5*, 21-49.

FlyBase Consortium (2002). The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res, 30*(1), 106-108.

UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res, 38*(Database issue), D142-148.

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., et al. (2009). Literature-curated protein interaction datasets. *Nat Methods, 6*(1), 39-46.

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., & Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics, 20*(5), 604-611.

de Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugue, S., et al. (2008). Hepatitis C virus infection protein network. *Mol Syst Biol, 4*, 230.

Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R., et al. (2002). PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics, 18*(7), 996-1003.

DiGiacomo, R. A., Kremer, J. M., & Shah, D. M. (1989). Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind,

controlled, prospective study. *Am J Med, 86*(2), 158-164.

Ding, J., Berleant, D., Nettleton, D., & Wurtele, E. (2002). Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*, 326-337.

Dogrusoz, U., Erson, E. Z., Giral, E., Demir, E., Babur, O., Cetintas, A., et al. (2006). PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics, 22*(3), 374-375.

Dubuisson, J. (2007). Hepatitis C virus proteins. *World J Gastroenterol, 13*(17), 2406-2415.

Dubuisson, J., Helle, F., & Cocquerel, L. (2008). Early steps of the hepatitis C virus life cycle. *Cell Microbiol, 10*(4), 821-827.

Durantel, D. (2009). Celgosivir, an alpha-glucosidase I inhibitor for the potential treatment of HCV infection. *Curr Opin Investig Drugs, 10*(8), 860-870.

Egger, D., Wolk, B., Gosert, R., Bianchi, L., Blum, H. E., Moradpour, D., et al. (2002). Expression of hepatitis C virus proteins induces distinct membrane alterations including a candidate viral replication complex. *J Virol, 76*(12), 5974-5984.

Egorov, S., Yuryev, A., & Daraselia, N. (2004). A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J Am Med Inform Assoc, 11*(3), 174-178.

EHRN (2007). *HCV infection in Europe*: Eurasian Harm Reduction Network.

Elazar, M., Liu, P., Rice, C. M., & Glenn, J. S. (2004). An N-terminal amphipathic helix in hepatitis C virus (HCV) NS4B mediates membrane association, correct localization of replication complex proteins, and HCV RNA replication. *J Virol, 78*(20), 11393-11400.

Elbers, K., Tautz, N., Becher, P., Stoll, D., Rumenapf, T., & Thiel, H. J. (1996). Processing in the pestivirus E2-NS2 region: identification of proteins p7 and E2p7. *J Virol, 70*(6), 4131-4135.

Espadaler, J., Romero-Isart, O., Jackson, R. M., & Oliva, B. (2005). Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics, 21*(16), 3360-3368.

Evans, J. A., & Foster, J. G. (2011). Metaknowledge. *Science, 331*(6018), 721-725.

Flint, M., Thomas, J. M., Maidens, C. M., Shotton, C., Levy, S., Barclay, W. S., et al. (1999). Functional analysis of cell surface-expressed hepatitis C virus E2 glycoprotein. *J Virol, 73*(8), 6782-6790.

Fontaine, J. F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M., & Andrade-Navarro, M. A. (2009). MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res, 37*(Web Server issue), W141-146.

Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P., & Coster, J. (2002). Protein names and how to find them. *Int J Med Inform, 67*(1-3), 49-61.

Fried, M. W., Shiffman, M. L., Reddy, K. R., Smith, C., Marinos, G., Goncales, F. L., Jr., et al. (2002). Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. *N Engl J Med, 347*(13), 975-982.

Frijters, R., van Vugt, M., Smeets, R., van Schaik, R., de Vlieg, J., & Alkema, W. (2010). Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput Biol, 6*(9).

Fukuda, K., Tamura, A., Tsunoda, T., & Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. *Pac Symp*

*Biocomput*, 707-718.

Fundel, K., Kuffner, R., & Zimmer, R. (2007). RelEx--relation extraction using dependency parse trees. *Bioinformatics, 23*(3), 365-371.

Gabow, A. P., Leach, S. M., Baumgartner, W. A., Hunter, L. E., & Goldberg, D. S. (2008). Improving protein function prediction methods with integrated literature data. *BMC Bioinformatics, 9*, 198.

Gambarin-Gelwan, M., & Jacobson, I. M. (2008). Optimal dose of peginterferon and ribavirin for treatment of chronic hepatitis C. *J Viral Hepat, 15*(9), 623-633.

Gangadharan, B., Antrobus, R., Dwek, R. A., & Zitzmann, N. (2007). Novel serum biomarker candidates for liver fibrosis in hepatitis C patients. *Clin Chem, 53*(10), 1792-1799.

Gao, M., & Skolnick, J. (2011). New benchmark metrics for protein-protein docking methods. *Proteins, 79*(5), 1623-1634.

Gardelli, C., Attenni, B., Donghi, M., Meppen, M., Pacini, B., Harper, S., et al. (2009). Phosphoramidate prodrugs of 2'-C-methylcytidine for therapy of hepatitis C virus infection. *J Med Chem, 52*(17), 5394-5407.

Gardner, J. P., Durso, R. J., Arrigale, R. R., Donovan, G. P., Maddon, P. J., Dragic, T., et al. (2003). L-SIGN (CD 209L) is a liver-specific capture receptor for hepatitis C virus. *Proc Natl Acad Sci U S A, 100*(8), 4498-4503.

Gaudet, P., Bairoch, A., Field, D., Sansone, S. A., Taylor, C., Attwood, T. K., et al. (2011). Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res, 39*(Database issue), D7-10.

Gouttenoire, J., Penin, F., & Moradpour, D. (2010). Hepatitis C virus nonstructural protein 4B: a journey into unexplored territory. *Rev Med Virol, 20*(2), 117-129.

Guo, Y., Li, M., Pu, X., Li, G., Guang, X., Xiong, W., et al. (2010). PRED_PPI: a server for predicting protein-protein interactions based on sequence data with probability assignment. *BMC Res Notes, 3*, 145.

Han, D. S., Kim, H. S., Jang, W. H., Lee, S. D., & Suh, J. K. (2004b). PreSPI: design and implementation of protein-protein interaction prediction service system. *Genome Inform, 15*(2), 171-180.

Han, K., Park, B., Kim, H., Hong, J., & Park, J. (2004a). HPID: the Human Protein Interaction Database. *Bioinformatics, 20*(15), 2466-2470.

Hanisch, D., Fluck, J., Mevissen, H. T., & Zimmer, R. (2003). Playing biology's name game: identifying protein names in scientific text. *Pac Symp Biocomput*, 403-414.

Hanisch, D., Fundel, K., Mevissen, H. T., Zimmer, R., & Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics, 6 Suppl 1*, S14.

Helle, F., Wychowski, C., Vu-Dac, N., Gustafson, K. R., Voisset, C., & Dubuisson, J. (2006). Cyanovirin-N inhibits hepatitis C virus entry by binding to envelope protein glycans. *J Biol Chem, 281*(35), 25177-25183.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., et al. (2004). The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol, 22*(2), 177-183.

Hernandez-Toro, J., Prieto, C., & De las Rivas, J. (2007). APID2NET: unified interactome graphic analyzer. *Bioinformatics, 23*(18), 2495-2497.

Hettne, K. M., Weeber, M., Laine, M. L., ten Cate, H., Boyer, S., Kors, J. A., et al.

(2007). Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study. *J Clin Periodontol, 34*(12), 1016-1024.

Hirschman, L., Morgan, A. A., & Yeh, A. S. (2002). Rutabaga by any other name: extracting biological names. *J Biomed Inform, 35*(4), 247-259.

Hormozdiari, F., Salari, R., Bafna, V., & Sahinalp, S. C. (2010). Protein-protein interaction network evaluation for identifying potential drug targets. *J Comput Biol, 17*(5), 669-684.

Hou, W., Tian, Q., Zheng, J., & Bonkovsky, H. L. (2010). Zinc mesoporphyrin induces rapid proteasomal degradation of hepatitis C nonstructural 5A protein in human hepatoma cells. *Gastroenterology, 138*(5), 1909-1919.

Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. (2006). Exploiting semantic relations for literature-based discovery. *AMIA Annu Symp Proc*, 349-353.

Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *Int J Med Inform, 74*(2-4), 289-298.

Hu, Z., Hung, J. H., Wang, Y., Chang, Y. C., Huang, C. L., Huyck, M., et al. (2009). VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res, 37*(Web Server issue), W115-121.

Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K., & Li, M. (2004a). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics, 20*(18), 3604-3612.

Huang, M., Zhu, X., & Li, M. (2006). A hybrid method for relation extraction from biomedical literature. *Int J Med Inform, 75*(6), 443-455.

Huang, T. W., Tien, A. C., Huang, W. S., Lee, Y. C., Peng, C. L., Tseng, H. H., et al. (2004b). POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics, 20*(17), 3273-3276.

Hunter, L., & Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed? *Mol Cell, 21*(5), 589-594.

Iragne, F., Nikolski, M., Mathieu, B., Auber, D., & Sherman, D. (2005). ProViz: protein interaction visualization and exploration. *Bioinformatics, 21*(2), 272-274.

Janin, J. (2005). Assessing predictions of protein-protein interaction: the CAPRI experiment. *Protein Sci, 14*(2), 278-283.

Jelier, R., Jenster, G., Dorssers, L. C., van der Eijk, C. C., van Mulligen, E. M., Mons, B., et al. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics, 21*(9), 2049-2058.

Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet, 7*(2), 119-129.

Jenssen, T. K., Laegreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet, 28*(1), 21-28.

Kahn, C. E., Jr., & Thao, C. (2007). GoldMiner: a radiology image search engine. *AJR Am J Roentgenol, 188*(6), 1475-1478.

Kamburov, A., Pentchev, K., Galicka, H., Wierling, C., Lehrach, H., & Herwig, R. (2011). ConsensusPathDB: toward a more complete picture of cell biology.

*Nucleic Acids Res, 39*(Database issue), D712-717.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A. F., Vinod, N., et al. (2007). Broadening the horizon--level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol, 5*, 44.

Keskin, O., Nussinov, R., & Gursoy, A. (2008). PRISM: protein-protein interaction prediction by structural matching. *Methods Mol Biol, 484*, 505-521.

Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res, 37*(Database issue), D387-392.

Kim, W. R., Brown, R. S., Jr., Terrault, N. A., & El-Serag, H. (2002). Burden of liver disease in the United States: summary of a workshop. *Hepatology, 36*(1), 227-242.

Koike, A., Niwa, Y., & Takagi, T. (2005). Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics, 21*(7), 1227-1236.

Kou, Z., Cohen, W. W., & Murphy, R. F. (2005). High-recall protein entity recognition using a dictionary. *Bioinformatics, 21 Suppl 1*, i266-273.

Krallinger, M., Leitner, F., Rodriguez-Penagos, C., & Valencia, A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol, 9 Suppl 2*, S4.

Krekulova, L., Rehak, V., & Riley, L. W. (2006). Structure and functions of hepatitis C virus proteins: 15 years after. *Folia Microbiol (Praha), 51*(6), 665-680.

Kremsdorf, D., & Brezillon, N. (2007). New animal models for hepatitis C viral infection and pathogenesis studies. *World J Gastroenterol, 13*(17), 2427-2435.

Kuiken, C., & Simmonds, P. (2009). Nomenclature and numbering of the hepatitis C virus. *Methods Mol Biol, 510*, 33-53.

Kwofie, S. K., Radovanovic, A., Sundararajan, V. S., Maqungo, M., Christoffels, A., & Bajic, V. B. (2011). Dragon exploratory system on hepatitis C virus (DESHCV). *Infect Genet Evol, 11*(4), 734-739.

Lalonde, S., Ehrhardt, D. W., Loque, D., Chen, J., Rhee, S. Y., & Frommer, W. B. (2008). Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations. *Plant J, 53*(4), 610-635.

Lanford, R. E., Bigger, C., Bassett, S., & Klimpel, G. (2001). The chimpanzee model of hepatitis C virus infections. *ILAR J, 42*(2), 117-126.

Lawitz, E., Godofsky, E., Rouzier, R., Marbury, T., Nguyen, T., Ke, J., et al. (2011). Safety, pharmacokinetics, and antiviral activity of the cyclophilin inhibitor NIM811 alone or in combination with pegylated interferon in HCV-infected patients receiving 14 days of therapy. *Antiviral Res, 89*(3), 238-245.

Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. *Pac Symp Biocomput*, 652-663.

Lee, K. J., Hwang, Y. S., Kim, S., & Rim, H. C. (2004). Biomedical named entity recognition using two-phase model based on SVMs. *J Biomed Inform, 37*(6), 436-447.

Lemm, J. A., O'Boyle, D., 2nd, Liu, M., Nower, P. T., Colonno, R., Deshpande, M. S., et al. (2010). Identification of hepatitis C virus NS5A inhibitors. *J Virol, 84*(1), 482-491.

Leroy, G., & Chen, H. (2002). Filling preposition-based templates to capture information

from medical abstracts. *Pac Symp Biocomput*, 350-361.

Leroy, G., Chen, H., & Martinez, J. D. (2003). A shallow parser based on closed-class words to capture relations in biomedical text. *J Biomed Inform, 36*(3), 145-158.

Letunic, I., Doerks, T., & Bork, P. (2009). SMART 6: recent updates and new developments. *Nucleic Acids Res, 37*(Database issue), D229-232.

Lewis, A. C., Saeed, R., & Deane, C. M. (2010). Predicting protein-protein interactions in the context of protein evolution. *Mol Biosyst, 6*(1), 55-64.

Li, Y., Hu, X., Lin, H., & Yang, Z. (2011). A framework for semisupervised feature generation and its applications in biomedical literature mining. *IEEE/ACM Trans Comput Biol Bioinform, 8*(2), 294-307.

Liu, F., Ackerman, M., & Fontelo, P. (2006). BabelMeSH: development of a cross-language tool for MEDLINE/PubMed. *AMIA Annu Symp Proc*, 1012.

Liu, S., Yang, W., Shen, L., Turner, J. R., Coyne, C. B., & Wang, T. (2009). Tight junction proteins claudin-1 and occludin control hepatitis C virus entry and are downregulated during infection to prevent superinfection. *J Virol, 83*(4), 2011-2014.

Lohmann, V., Korner, F., Koch, J., Herian, U., Theilmann, L., & Bartenschlager, R. (1999). Replication of subgenomic hepatitis C virus RNAs in a hepatoma cell line. *Science, 285*(5424), 110-113.

Long, G., Hiet, M. S., Windisch, M. P., Lee, J. Y., Lohmann, V., & Bartenschlager, R. (2011). Mouse Hepatic Cells Support Assembly of Infectious Hepatitis C Virus Particles. *Gastroenterology*.

Lozach, P. Y., Amara, A., Bartosch, B., Virelizier, J. L., Arenzana-Seisdedos, F., Cosset, F. L., et al. (2004). C-type lectins L-SIGN and DC-SIGN capture and transmit infectious hepatitis C virus pseudotype particles. *J Biol Chem, 279*(31), 32035-32045.

Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. *Database (Oxford), 2011*, baq036.

Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y., & Friedman, C. (2006). PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput*, 64-75.

Ma, S., Boerner, J. E., TiongYip, C., Weidmann, B., Ryder, N. S., Cooreman, M. P., et al. (2006). NIM811, a cyclophilin inhibitor, exhibits potent in vitro activity against hepatitis C virus alone or in combination with alpha interferon. *Antimicrob Agents Chemother, 50*(9), 2976-2982.

Martell, M., Esteban, J. I., Quer, J., Genesca, J., Weiner, A., Esteban, R., et al. (1992). Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. *J Virol, 66*(5), 3225-3229.

Martin, A., Ochagavia, M. E., Rabasa, L. C., Miranda, J., Fernandez-de-Cossio, J., & Bringas, R. (2010). BisoGenet: a new tool for gene network building, visualization and analysis. *BMC Bioinformatics, 11*, 91.

Martin, S., Brown, W. M., & Faulon, J. L. (2008). Using product kernels to predict protein interactions. *Adv Biochem Eng Biotechnol, 110*, 215-245.

Mas, V. R., Maluf, D. G., Archer, K. J., Yanek, K., Bornstein, K., & Fisher, R. A. (2009). Proteomic analysis of HCV cirrhosis and HCV-induced HCC: identifying

biomarkers for monitoring HCV-cirrhotic patients awaiting liver transplantation. *Transplantation, 87*(1), 143-152.

McDowall, M. D., Scott, M. S., & Barton, G. J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res, 37*(Database issue), D651-656.

Meil, A., Durand, P., & Wojcik, J. (2005). PIMWalker: visualising protein interaction networks using the HUPO PSI molecular interaction format. *Appl Bioinformatics, 4*(2), 137-139.

Meireles, L. M., Domling, A. S., & Camacho, C. J. (2010). ANCHOR: a web server and database for analysis of protein-protein interaction binding pockets for drug discovery. *Nucleic Acids Res, 38*(Web Server issue), W407-411.

Mercadal, C. M., Martin, S., & Rouse, B. T. (1991). Apparent requirement for CD4+ T cells in primary anti-herpes simplex virus cytotoxic T-lymphocyte induction can be overcome by optimal antigen presentation. *Viral Immunol, 4*(3), 177-186.

Mika, S., & Rost, B. (2004). NLProt: extracting protein names and sequences from papers. *Nucleic Acids Res, 32*(Web Server issue), W634-637.

Mirny, L. A., & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J Mol Biol, 291*(1), 177-196.

Monazahian, M., Bohme, I., Bonk, S., Koch, A., Scholz, C., Grethe, S., et al. (1999). Low density lipoprotein receptor as a candidate receptor for hepatitis C virus. *J Med Virol, 57*(3), 223-229.

Muller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004a). A review of content-based image retrieval systems in medical applications-clinical benefits and future directions. *Int J Med Inform, 73*(1), 1-23.

Muller, H. M., Kenny, E. E., & Sternberg, P. W. (2004b). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol, 2*(11), e309.

Narayanaswamy, M., Ravikumar, K. E., & Vijay-Shanker, K. (2003). A biological named entity recognizer. *Pac Symp Biocomput*, 427-438.

Neveol, A., Zeng, K., & Bodenreider, O. (2006). Besides precision & recall: exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE. *AMIA Annu Symp Proc*, 589-593.

Ono, T., Hishigaki, H., Tanigami, A., & Takagi, T. (2001). Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics, 17*(2), 155-161.

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stumpflen, V., et al. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol, 25*(8), 894-898.

Outwater, E. K. (2010). Imaging of the liver for hepatocellular cancer. *Cancer Control, 17*(2), 72-82.

Ouzounov, S., Mehta, A., Dwek, R. A., Block, T. M., & Jordan, R. (2002). The combination of interferon alpha-2b and n-butyl deoxynojirimycin has a greater than additive antiviral effect upon production of infectious bovine viral diarrhea virus (BVDV) in vitro: implications for hepatitis C virus (HCV) therapy. *Antiviral Res, 55*(3), 425-435.

Park, J. C., Kim, H. S., & Kim, J. J. (2001). Bidirectional incremental parsing for

automatic pathway identification with combinatory categorial grammar. *Pac Symp Biocomput*, 396-407.

Payer, B. A., Reiberger, T., Rutter, K., Beinhardt, S., Staettermayer, A. F., Peck-Radosavljevic, M., et al. (2010). Successful HCV eradication and inhibition of HIV replication by intravenous silibinin in an HIV-HCV coinfected patient. *J Clin Virol, 49*(2), 131-133.

Pazos, F., Juan, D., Izarzugaza, J. M., Leon, E., & Valencia, A. (2008). Prediction of protein interaction based on similarity of phylogenetic trees. *Methods Mol Biol, 484*, 523-535.

Penin, F., Dubuisson, J., Rey, F. A., Moradpour, D., & Pawlotsky, J. M. (2004). Structural biology of hepatitis C virus. *Hepatology, 39*(1), 5-19.

Petracca, R., Falugi, F., Galli, G., Norais, N., Rosa, D., Campagnoli, S., et al. (2000). Structure-function analysis of hepatitis C virus envelope-CD81 binding. *J Virol, 74*(10), 4824-4830.

Phizicky, E. M., & Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev, 59*(1), 94-123.

Pileri, P., Uematsu, Y., Campagnoli, S., Galli, G., Falugi, F., Petracca, R., et al. (1998). Binding of hepatitis C virus to CD81. *Science, 282*(5390), 938-941.

Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., et al. (2006). PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC Bioinformatics, 7*, 365.

Plewczynski, D., & Ginalski, K. (2009). The interactome: predicting the protein-protein interactions in cells. *Cell Mol Biol Lett, 14*(1), 1-22.

Plewczynski, D., & Klingstrom, T. (2011). GIDMP: Good protein-protein interaction data metamining practice. *Cell Mol Biol Lett, 16*(2), 258-263.

Plikus, M. V., Zhang, Z., & Chuong, C. M. (2006). PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC Bioinformatics, 7*, 424.

Ploss, A., Evans, M. J., Gaysinskaya, V. A., Panis, M., You, H., de Jong, Y. P., et al. (2009). Human occludin is a hepatitis C virus entry factor required for infection of mouse cells. *Nature, 457*(7231), 882-886.

Prieto, C., & De Las Rivas, J. (2006). APID: Agile Protein Interaction DataAnalyzer. *Nucleic Acids Res, 34*(Web Server issue), W298-302.

Proux, D., Rechenmann, F., & Julliard, L. (2000). A pragmatic information extraction strategy for gathering data on genetic interactions. *Proc Int Conf Intell Syst Mol Biol, 8*, 279-285.

Proux, D., Rechenmann, F., Julliard, L., Pillet, V. V., & Jacq, B. (1998). Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome Inform Ser Workshop Genome Inform, 9*, 72-80.

Pustejovsky, J., Castano, J., Zhang, J., Kotecki, M., & Cochran, B. (2002). Robust relational parsing over biomedical literature: extracting inhibit relations. *Pac Symp Biocomput*, 362-373.

Qian, Y., & Murphy, R. F. (2008). Improved recognition of figures containing fluorescence microscope images in online journal articles using graphical models. *Bioinformatics, 24*(4), 569-576.

Razick, S., Magklaras, G., & Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics, 9*, 405.

Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform, 36*(6), 462-477.

Robertson, B., Myers, G., Howard, C., Brettin, T., Bukh, J., Gaschen, B., et al. (1998). Classification, nomenclature, and database development for hepatitis C virus (HCV) and related viruses: proposals for standardization. International Committee on Virus Taxonomy. *Arch Virol, 143*(12), 2493-2503.

Rossignol, J. F., Elfert, A., & Keeffe, E. B. (2010). Treatment of chronic hepatitis C using a 4-week lead-in with nitazoxanide before peginterferon plus nitazoxanide. *J Clin Gastroenterol, 44*(7), 504-509.

Rychlowska, M., & Bienkowska-Szewczyk, K. (2007). Hepatitis C--new developments in the studies of the viral life cycle. *Acta Biochim Pol, 54*(4), 703-715.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., et al. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford), 2010*, baq020.

Sasaki, Y., Tsuruoka, Y., McNaught, J., & Ananiadou, S. (2008). How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics, 9 Suppl 11*, S5.

Saunier, B., Triyatni, M., Ulianich, L., Maruvada, P., Yen, P., & Kohn, L. D. (2003). Role of the asialoglycoprotein receptor in binding and entry of hepatitis C virus structural proteins in cultured human hepatocytes. *J Virol, 77*(1), 546-559.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., et al. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res, 39*(Database issue), D38-51.

Scarselli, E., Ansuini, H., Cerino, R., Roccasecca, R. M., Acali, S., Filocamo, G., et al. (2002). The human scavenger receptor class B type I is a novel candidate receptor for the hepatitis C virus. *EMBO J, 21*(19), 5017-5025.

Sekimizu, T., Park, H. S., & Tsujii, J. (1998). Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. *Genome Inform Ser Workshop Genome Inform, 9*, 62-71.

Sen, T. Z., Kloczkowski, A., & Jernigan, R. L. (2006). Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics, 7*, 355.

Sen, T. Z., Kloczkowski, A., Jernigan, R. L., Yan, C., Honavar, V., Ho, K. M., et al. (2004). Predicting binding sites of hydrolase-inhibitor complexes by combining several methods. *BMC Bioinformatics, 5*, 205.

Sharma, S. D. (2010). Hepatitis C virus: molecular biology & current therapeutic options. *Indian J Med Res, 131*, 17-34.

Sharon, I., Davis, J. V., & Yona, G. (2009). Prediction of protein-protein interactions: a study of the co-evolution model. *Methods Mol Biol, 541*, 61-88.

Shatkay, H., Chen, N., & Blostein, D. (2006). Integrating image data into biomedical text categorization. *Bioinformatics, 22*(14), e446-453.

Sievert, W., Altraif, I., Razavi, H. A., Abdo, A., Ahmed, E. A., Alomair, A., et al. (2011). A systematic review of hepatitis C virus epidemiology in Asia, Australia and Egypt. *Liver Int, 31 Suppl 2*, 61-80.

Simmonds, P., Alberti, A., Alter, H. J., Bonino, F., Bradley, D. W., Brechot, C., et al. (1994). A proposed system for the nomenclature of hepatitis C viral genotypes. *Hepatology, 19*(5), 1321-1324.

Simmonds, P., Bukh, J., Combet, C., Deleage, G., Enomoto, N., Feinstone, S., et al. (2005). Consensus proposals for a unified system of nomenclature of hepatitis C virus genotypes. *Hepatology, 42*(4), 962-973.

Simmonds, P., Holmes, E. C., Cha, T. A., Chan, S. W., McOmish, F., Irvine, B., et al. (1993). Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. *J Gen Virol, 74 ( Pt 11)*, 2391-2399.

Singhal, M., & Resat, H. (2007). A domain-based approach to predict protein-protein interactions. *BMC Bioinformatics, 8*, 199.

Smialowski, P., Pagel, P., Wong, P., Brauner, B., Dunger, I., Fobo, G., et al. (2010). The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res, 38*(Database issue), D540-544.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol, 25*(11), 1251-1255.

Smith, L., Tanabe, L. K., Ando, R. J., Kuo, C. J., Chung, I. F., Hsu, C. N., et al. (2008). Overview of BioCreative II gene mention recognition. *Genome Biol, 9 Suppl 2*, S2.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., & Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics, 27*(3), 431-432.

Song, Y., Kim, E., Lee, G. G., & Yi, B. K. (2005). POSBIOTM-NER: a trainable biomedical named-entity recognition system. *Bioinformatics, 21*(11), 2794-2796.

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: making sense of raw text. *Brief Bioinform, 6*(3), 239-251.

Stapley, B. J., & Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac Symp Biocomput*, 529-540.

Strader, D. B., Wright, T., Thomas, D. L., & Seeff, L. B. (2004). Diagnosis, management, and treatment of hepatitis C. *Hepatology, 39*(4), 1147-1171.

Suderman, M., & Hallett, M. (2007). Tools for visually exploring biological networks. *Bioinformatics, 23*(20), 2651-2659.

Suzuki, R., Suzuki, T., Ishii, K., Matsuura, Y., & Miyamura, T. (1999). Processing and functions of Hepatitis C virus proteins. *Intervirology, 42*(2-3), 145-152.

Suzuki, T., Aizaki, H., Murakami, K., Shoji, I., & Wakita, T. (2007). Molecular biology of hepatitis C virus. *J Gastroenterol, 42*(6), 411-423.

Swanson, D. R. (1991). *Complementary structures in disjoint science literatures.* Paper presented at the Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Swanson, D. R. (1990). Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc, 78*(1), 29-37.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public

knowledge. *Perspect Biol Med, 30*(1), 7-18.

Sy, T., & Jamal, M. M. (2006). Epidemiology of hepatitis C virus (HCV) infection. *Int J Med Sci, 3*(2), 41-46.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res, 39*(Database issue), D561-568.

Tamames, J. (2005). Text detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics, 6 Suppl 1*, S10.

Taylor, C. F., Field, D., Sansone, S. A., Aerts, J., Apweiler, R., Ashburner, M., et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat Biotechnol, 26*(8), 889-896.

Taylor, D. P. (2007). An integrated biomedical knowledge extraction and analysis platform: using federated search and document clustering technology. *Methods Mol Biol, 356*, 293-300.

Tencate, V., Sainz, B., Cotler, S. J., & Uprichard, S. L. (2010). Potential treatment options and future research to increase hepatitis C virus treatment response rate. *Hepat Med, 2010*(2), 125-145.

Torii, M., Hu, Z., Wu, C. H., & Liu, H. (2009). BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assoc, 16*(2), 247-255.

Tsai, R. T., Lai, P. T., Dai, H. J., Huang, C. H., Bow, Y. Y., Chang, Y. C., et al. (2009). HypertenGene: extracting key hypertension genes from biomedical literature with position and automatically-generated template features. *BMC Bioinformatics, 10 Suppl 15*, S9.

Tsuruoka, Y., McNaught, J., Tsujii, J., & Ananiadou, S. (2007). Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics, 23*(20), 2768-2774.

Tsuruoka, Y., & Tsujii, J. (2004). Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform, 37*(6), 461-470.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., & Wodak, S. J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database (Oxford), 2010*, baq026.

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database (Oxford), 2010*, baq023.

Uwimana, E., & Ruiz, M. E. (2008). Integrating an automatic classification method into the medical image retrieval process. *AMIA Annu Symp Proc*, 747-751.

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

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., et al. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res, 33*(Database issue), D433-437.

Wakita, T., Pietschmann, T., Kato, T., Date, T., Miyamoto, M., Zhao, Z., et al. (2005). Production of infectious hepatitis C virus in tissue culture from a cloned viral

genome. *Nat Med, 11*(7), 791-796.

Wang, J., Cetindil, I., Ji, S., Li, C., Xie, X., Li, G., et al. (2010). Interactive and fuzzy search: a dynamic way to explore MEDLINE. *Bioinformatics, 26*(18), 2321-2327.

Weeber, M., Klein, H., De Jong-van den Berg, L. T. W. and Vos, R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *J. Amer. Soc. Inf. Sci. Tech, 52*(7), 254–262.

Weeber, M., Klein, H., Aronson, A. R., Mork, J. G., de Jong-van den Berg, L. T., & Vos, R. (2000). Text-based discovery in biomedicine: the architecture of the DAD-system. *Proc AMIA Symp*, 903-907.

Weeber, M., Kors, J. A., & Mons, B. (2005). Online tools to support literature-based discovery in the life sciences. *Brief Bioinform, 6*(3), 277-286.

Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L. T., Aronson, A. R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J Am Med Inform Assoc, 10*(3), 252-259.

WHO (2009). Hepatitis C Virus. *Viral Cancers- World Health Organization (WHO) Initiative for Vaccine Research (IVR)* Retrieved 11/23/2009, 2009, from http://www.who.int/vaccine_research/diseases/viral_cancers/en/

Winter, A. G., Wildenhain, J., & Tyers, M. (2011). BioGRID REST Service, BiogridPlugin2 and BioGRID WebGraph: new tools for access to interaction data at BioGRID. *Bioinformatics, 27*(7), 1043-1044.

Wren, J. D., & Garner, H. R. (2004). Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics, 20*(2), 191-198.

Xia, K., Dong, D., & Han, J. D. (2006). IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. *BMC Bioinformatics, 7*, 508.

Xu, S., McCusker, J., & Krauthammer, M. (2008). Yale Image Finder (YIF): a new search engine for retrieving biomedical images. *Bioinformatics, 24*(17), 1968-1970.

Xu, Z., Choi, J., Yen, T. S., Lu, W., Strohecker, A., Govindarajan, S., et al. (2001). Synthesis of a novel hepatitis C virus protein by ribosomal frameshift. *EMBO J, 20*(14), 3840-3848.

Yahia, M. (2011). Global health: a uniquely Egyptian epidemic. *Nature, 474*(7350), S12-13.

Yang, W., Qiu, C., Biswas, N., Jin, J., Watkins, S. C., Montelaro, R. C., et al. (2008). Correlation of the tight junction-like distribution of Claudin-1 to the cellular tropism of hepatitis C virus. *J Biol Chem, 283*(13), 8643-8653.

Yang, Y., Adelstein, S. J., & Kassis, A. I. (2009). Target discovery from data mining approaches. *Drug Discov Today, 14*(3-4), 147-154.

Yao, L., Evans, J. A., & Rzhetsky, A. (2010). Novel opportunities for computational biology and sociology in drug discovery. *Trends Biotechnol, 28*(4), 161-170.

Yao, L., Evans, J. A., & Rzhetsky, A. (2009). Novel opportunities for computational biology and sociology in drug discovery. *Trends Biotechnol, 27*(9), 531-540.

Yeh, A., Morgan, A., Colosimo, M., & Hirschman, L. (2005). BioCreAtIvE task 1A:

gene mention finding evaluation. *BMC Bioinformatics, 6 Suppl 1*, S2.

Yeh, A. S., Hirschman, L., & Morgan, A. A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics, 19 Suppl 1*, i331-339.

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., & Jothi, R. (2011). DOMINE: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Res, 39*(Database issue), D730-735.

Yu, H., Kim, T., Oh, J., Ko, I., Kim, S., & Han, W. S. (2010). Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. *BMC Bioinformatics, 11 Suppl 2*, S6.

Yu, N. C., Chaudhari, V., Raman, S. S., Lassman, C., Tong, M. J., Busuttil, R. W., et al. (2011). CT and MRI improve detection of hepatocellular carcinoma, compared with ultrasound alone, in patients with cirrhosis. *Clin Gastroenterol Hepatol, 9*(2), 161-167.

Zender, L., & Kubicka, S. (2008). Molecular pathogenesis and targeted therapy of hepatocellular carcinoma. *Onkologie, 31*(10), 550-555.

Zhou, G., Zhang, J., Su, J., Shen, D., & Tan, C. (2004). Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics, 20*(7), 1178-1190.

Zinovyev, A., Viara, E., Calzone, L., & Barillot, E. (2008). BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics, 24*(6), 876-877.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., & Cohen, K. B. (2007). Frontiers of biomedical text mining: current progress. *Brief Bioinform, 8*(5), 358-375.

# CHAPTER 2: Dragon Exploratory System on Hepatitis C Virus (DESHCV)
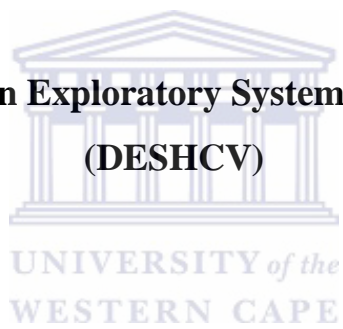
## 2.1 Abstract

Even though Hepatitis C Virus (HCV) cDNA was characterized about 20 years ago, there is insufficient understanding of the molecular etiology underlying HCV infections. Current global rates of infection and its increasingly chronic character are causes of concern for health policy experts. Vast amount of data accumulated from biochemical, genomic, proteomic, and other biological analyses allows for novel insights into the HCV viral structure, life cycle and functions of its proteins. Biomedical text-mining is a useful approach for analyzing the increasing corpus of published scientific literature on HCV. This is the first reported comprehensive HCV customized biomedical text-mining based online web resource, dragon exploratory system on Hepatitis C Virus (DESHCV), a biomedical text-mining and relationship exploring knowledgebase was developed by exploring literature on HCV. The pre-compiled dictionaries existing in the dragon exploratory system (DES) were enriched with biomedical concepts pertaining to HCV proteins, their name variants and symbols to make it suitable for targeted information exploration and knowledge extraction as focused on HCV. A list of 32,895 abstracts retrieved via the PubMed database using specific keywords searches related to HCV were processed based on concept recognition of terms from several dictionaries. The web query interface enables retrieval of information using specified concepts, keywords and phrases, generating text-derived association networks and hypotheses, which could be tested to identify potentially novel relationships between different concepts. Such an approach could also augment efforts in the search for diagnostic or even therapeutic targets. DESHCV thus represents an online literature-based discovery resource freely accessible for academic and non-profit users via http://apps.sanbi.ac.za/DESHCV/ and its mirror site http://cbrc.kaust.edu.sa/deshcv/.

## 2.2 Introduction

The skyrocketing chronicity and global infection rate of Hepatitis C Virus (HCV) necessitate the need to unlock the molecular etiology underlying the pathophysiology of HCV related diseases such as liver cancer. The plethora of essential molecular data in the corpus of published biomedical literature could be leveraged to augment efforts towards

discovery of novel anti-viral drugs, cellular receptors and appropriate predictive biomarkers. Most of the data derived from high throughput and "omics" experiments exist in variety of formats, thereby making cross data integration difficult. The development of HCV specific databases as repositories of information utilizable in cross discipline biology research is therefore vital. The Los Alamos Hepatitis C Virus sequence database (http://hcv.lanl.gov) offers annotated sequences and analysis tools (Kuiken et al., 2005). The Los Alamos hepatitis C immunology database (http://hcv.lanl.gov/content/immuno/immuno-main.html) is a repository of biocurated immunological epitopes integrated with retrieval and analysis tools (Yusim et al., 2005). The Japanese HCV database integrated in the HVDB (http://s2as02.genes.nig.ac.jp) comprises data on phylogenetic and provides java embedded viewers for visualizing phylogenetic trees and the HCV genome. The European Hepatitis C Virus database (euHCVdb, http://euhcvdb.ibcp.fr) provides annotated sequences and tools for analysis, and information on protein structure and function (Combet et al., 2007). Hepatitis C Virus sequence and immunology database and analytical applications (HCVdb, http://www.hcvdb.org/index.asp?bhcp=1) offers data on analyzed protein sequence and features, epitopes, and curated knowledge on protein interactions and function. Binding site finder (BSFINDER, http://wilab.inha.ac.kr/bsfinder) enables prediction of HCV binding site residues and potential interacting protein partners using support vector machine (Chen and Han, 2009). A comprehensive review of selected HCV related database has highlighted the useful capabilities, utilities and applications of these resources (Kuiken et al., 2006). Hepatitis C Virus-specific databases store useful information on molecular biology, sequences, immunology, protein structure and function, viral evolution and genetics. Nevertheless, there is no resource that allows for the exploration of potential links (associations) between different biomedical concepts of relevance to HCV. One such resource is the dragon exploratory system on Hepatitis C Virus (DESHCV, http://apps.sanbi.ac.za/DESHCV/ and its mirror site http://cbrc.kaust.edu.sa/deshcv/), based on text-mining approach to complement the existing HCV resources and to enable different insights into the molecular context of HCV functioning.

As reported by Cohen and Hersh (2005), the biomedical knowledgebase is growing at an increasing rate. PubMed database is currently a repository of about 20 million citations for biomedical articles from MEDLINE and life science related journals. MeSH indexers index about 500,000 journal articles annually for PubMed/MEDLINE (Mitchell et al., 2003). A search by publication dates in PubMed shows over 20,000 HCV related records published over the last decade. At the time when this study was conducted a total of 32,895 HCV related documents were available that makes it virtually impossible for a single researcher or a research group to process in any reasonable time. However, a biomedical text-mining approach could be utilized to analyze this large volume of scientific data and reports published on HCV. Biomedical text-mining employs different techniques to extract and summarize information from text (Cohen and Hunter, 2008). Its algorithms may derive putative relationships between disjunct sets of concepts to unravel potentially new associations and hypotheses for possible novel discovery. The co-occurrence of the concepts of interests, either in a portion of text (say abstract) or in a sentence, is identified computationally, and provides useful clues on the potential associations between these concepts, some of which may be completely new. For example, using text-mining techniques fish oil was proposed to have a potential therapeutic effect on Raynaud's disease (Swanson, 1986). The relationship between fish oil and Raynaud's disease was hypothesized via physiological concepts such as high blood pressure and platelet aggregation and subsequently this relationship was confirmed. As another example, literature based discovery was previously used to predict thalidomide as a possible therapeutic drug for HCV infection after detecting implicit associations in biomedical text (Weeber et al., 2003). A text-mining approach was also employed in identifying some of the hepatocellular proteins used in generating the human HCV interactome (de Chassey et al., 2008).

The essential features and characteristics of some of the available text-mining tools have been discussed elsewhere (Bajic et al., 2005; Weeber et al., 2005). Shi and Campagne (2005) have described in detail the various concepts, principles, challenges and algorithms behind development of biomedical text-mining tools during the building of protein catalogue and implementation of the Textractor Framework

(http://icb.med.cornell.edu/crt/textractor/index.xml).                Anni                2.0
(http://biosemantics.org/anni/), a web-based biomedical text-mining tool offers an
ontology-based interface to MEDLINE and enables retrieval of documents and possible
association amongst biomedical concepts (Jelier et al., 2008). PolySearch
(http://wishart.biology.ualberta.ca/polysearch), a web-based text-mining system allows
the retrieval of relationships between human diseases, genes, mutations, drugs and
metabolites (Cheng et al., 2008). Nowadays, customized knowledgebases or topic-
specific text-mining resources are increasingly becoming popular. Typical examples are
the dragon exploratory system (DES) based resources: dragon TF association miner
(DTFAM, http://research.i2r.a-star.edu.sg/DRAGON/TFAM), useful for exploring
functional association amongst transcription factors (Pan et al., 2004); dragon database
for exploration of sodium channels in human (DDESC, http://apps.sanbi.ac.za/ddesc),
which provides comprehensive text-mining information related to sodium channels
(Sagar et al., 2008). DES and its previous versions were successfully used in compilation
of several other resources such as in database on ovarian cancer (Kaur et al., 2009) and
esophageal cancer (Essack et al., 2009), as well as in studies on prioritizing disease genes
(Tiffin et al., 2005; Lombard et al., 2007).

At the time of initiating this thesis project, there was no single database solely focused on
HCV research that allowed for the comprehensive exploration of the association between
the biomedical concepts related to HCV. This thesis describes dragon exploratory system
on Hepatitis C Virus (DESHCV), the previous version of it being reported earlier
(Kwofie et al., 2009). DESHCV is developed using DES. The HCV proteins and their
name variants have been integrated into the pre-compiled dictionaries of biological
concepts present in DES. These concepts are cross-referenced to database such as gene
ontologies (GO), UNIPROT, KEGG Pathway, REACTOME and Entrez Gene. A list of
abstracts was retrieved via PubMed database using keywords related to HCV. These
abstracts were analyzed using concepts in the following dictionaries: "human genes and
proteins", "metabolites and enzymes", "pathways", "chemicals with pharmacological
effects", "Hepatitis C Virus concepts", and "disease concepts".

The user-friendly online interface allows concepts, keywords and phrases searches. A concept query could generate networks and hypothesis. The computationally suggested associations between genes and other concepts such as diseases may assist experimental biologist to explore which genes amongst a pool of genes need to be characterized for further molecular analysis. Such an approach could in principle also lead to possible discovery of new vaccines; and enhance the development of appropriate diagnostic method. The user has the possibility to inspect the post-processed PubMed abstracts with colour-coded tagged concepts from the used dictionaries as found in the text. The downloadable concept lists sheet could be a primary source of data for biocuration. The paired concept list spreadsheet can serve as the essential preliminary data for exploring associations between concepts and can be converted easily into simple interaction file format (SIF) compatible with some of the interaction visualization and analysis tools such as Cytoscape (Killcoyne et al., 2009). Researchers with minimal or no knowledge on text-mining can explore DESHCV with ease via system's simplified user query interface. The integrated downloadable tutorial manual and frequently asked questions (FAQ) information further aids easy use of the system (appendix I and II). DESHCV is an online text-mining developed knowledgebase freely available for non-commercial use via http://apps.sanbi.ac.za/DESHCV and http://cbrc.kaust.edu.sa/deshcv.

## 2.3 Construction and content

## 2.3.1 Implementation

A list of 32,895 MEDLINE abstracts was collected via PubMed interface using the following keywords query: HCV OR "Hepatitis C Virus". The PubMed textual data downloaded in the extensible markup language (XML) format allow for easy data integration into DES for semantic processing and analysis. The DESHCV data files were generated by DES, a proprietary biomedical text-mining tool of OrionCell (http://www.orioncell.org). The HCV proteins symbols and name variants were added to the "human protein and genes" dictionaries, and were mapped onto external annotation database. The HCV name variants have been disambiguated and integrated in the database. For example, a concept query with the word "core" retrieves "core protein" and

not the words "score" or "core". The word core apart from being part of the morphological features of HCV core protein has different meanings in various English dictionaries.

Post processed PubMed abstracts containing found concepts from used dictionaries constitute part of the database files. The DESHCV precompiled dictionary data files systems are composed of categories of terms such as names of "genes", "proteins", "metabolites", enzymes", "pathways", "pharmacological chemicals", and "diseases". The categories of terms consist of biological entities that are referred to as concepts. Therefore for the purpose of this work, each dictionary consists of a catalogue of related concepts. The DESHCV system is organized into 6 distinct dictionaries: "human genes and proteins", "metabolites and enzymes", "pathways", "chemicals with pharmacological effects", "Hepatitis C Virus concepts", and "disease concepts". Hierarchically, each dictionary is considered as a parent with no subcategories. For example, both human gene "interferon alpha 2" and human protein "alanine aminotransferase" are concepts belonging to the same parent "human genes and proteins" dictionary and not separate subcategories. In some instances certain concepts are assigned to one dictionary, even though they could belong to another. For example the concept name ribavirin, a therapeutic drug for hepatitis C infection has been assigned to the "metabolites and enzymes" dictionary, even though it could also belong to "chemicals with pharmacological effects" dictionary. The decision is influenced by the belief that most literature reports ribavirin as a metabolite.

For the purpose of hypothesis generation, the DES association module generates association maps in the form of graphs. The nodes of these graphs represent various concepts and the edges linking the nodes are weighted with frequencies of occurrence in the PubMed abstracts. Hypotheses are generated if for example, concepts X and Y are correlated, as well as concepts Y and Z, but no correlation is found between X and Z. Then the hypotheses generator module will suggest a potential link between concepts X and Z.

## 2.3.2 Database architecture

The database comprises of an Apache HTTP web server integrated with a back-end MySQL server whilst HTML/CSS and JavaScript scripts constitute the front-end (Figure 2.1). The data files consist of precompiled concept dictionaries, post processed PubMed abstracts, and cross-referenced annotations. The logic systems consist of Perl and PHP modules.
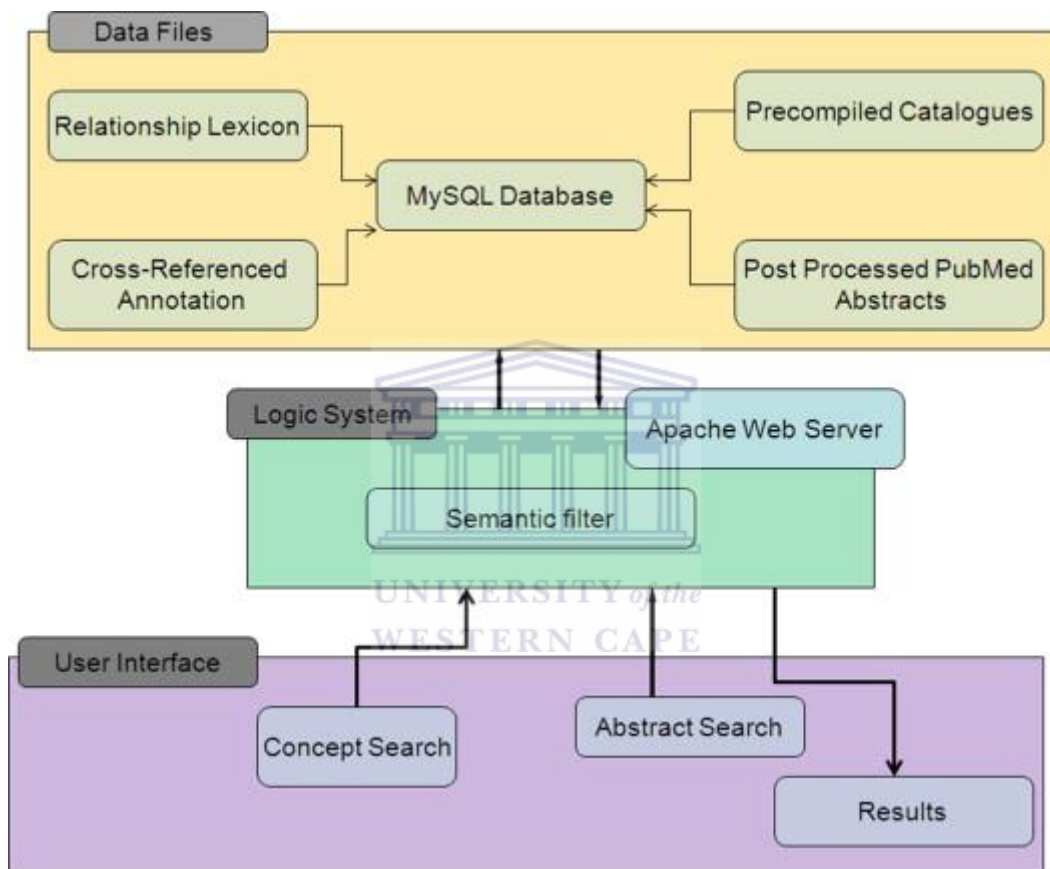
**FIGURE 2.1.** SCHEMATIC DIAGRAM OF THE INTEGRATED DATABASE OF DESHCV.
A layout of the DESHCV database architecture showing the relationship between the incorporated data files, logic systems and user query interface.

### 2.3.3 The database interfaces

DESHCV is based on a client-server model and can be accessed by any user with a standard web browser. The user interfaces in DESHCV allow easy navigation, query, inspection, and retrieval of data. It comprises of concept and abstract search menus (Figure 2.2). The concept search menu allows users to search the database using specified concepts. The hypothesis generator is an integral functional component of the concept search menu.
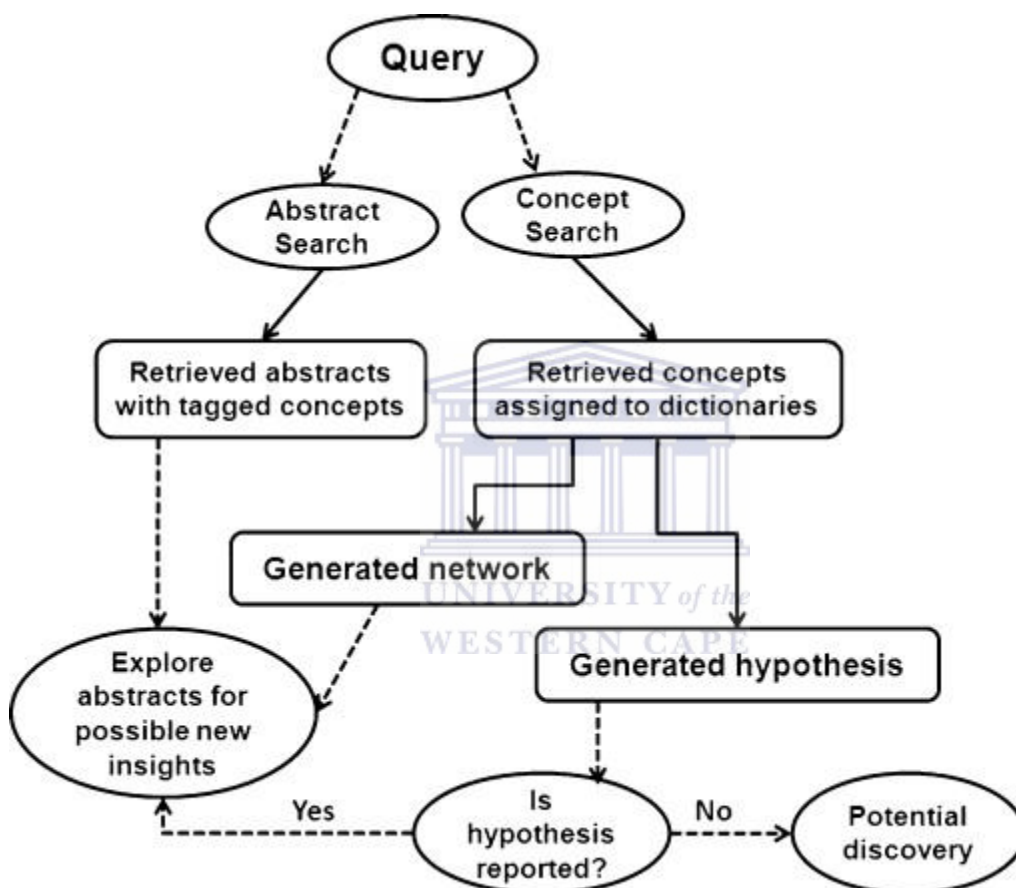
**FIGURE 2.2.** DESHCV DATA FLOW SCHEMA DIAGRAM.
A structured workflow outlining the various steps and decision-making processes involved in retrieving enriched biological data from DESHCV.

## 2.4 Results and discussions

### 2.4.1 Concepts queries

DESHCV is the first text-mining based web accessible resource developed using published abstracts of scientific reports on HCV as referenced in PubMed. DESHCV provides comprehensive information on HCV and enables users to gain easy insights through exploration of potential associations between the concepts of interest. Users can query the database using concepts such as genes, proteins, metabolites, enzymes, pathway, disease concepts, and pharmacological chemicals to retrieve useful associations that may suggest new insights into the problem. DESHCV provides users the chance to query the compiled abstracts within the framework of concept-based retrieval and extraction system.

For example, a concept query "thalidomide" retrieves all associated concepts from their respective dictionaries found in the analyzed text. These could be displayed either in a graphical or tabular format. These identified concepts co-occur with thalidomide in the PubMed abstracts. The frequency of occurrence of the concepts is shown and the link can be clicked to view the abstracts. The user has the option of either viewing the abstract with or without tagged concepts. All disease concepts associated with thalidomide are retrieved including chronic hepatitis C, which co-occurs with thalidomide within three abstracts. This result can be displayed in the form of an association map by using the "draw network" generator. The association map is a graph consisting of interacting network of nodes representing thalidomide and its associated concepts. The edges linking the nodes shows the relationship between the concepts and are weighted with the frequency of co-occurrence. To ensure effective interpretation and evaluation of results, users have the option of limiting the number of interactions for display by ignoring links with fewer frequencies. The association map can be resized from A0 to A5 and the detail slider can be used to alter viewing capabilities. For the purpose of this discussion, concepts with less than three links were ignored to obtain high degree of details without obscuring vital information. The association established between thalidomide and chronic hepatitis C is a loose relationship and no inference may be deduced until the literature is manually verified to either accept or reject the relationship (Figure 2.3). The linking

abstracts were manually inspected to ascertain the proposed relationship. Accordingly, thalidomide is a promising novel compound for chronic hepatitis C therapy since thalidomide decreased liver enzymes in six out of eight patients suffering from chronic hepatitis C (Milazzo et al., 2006).
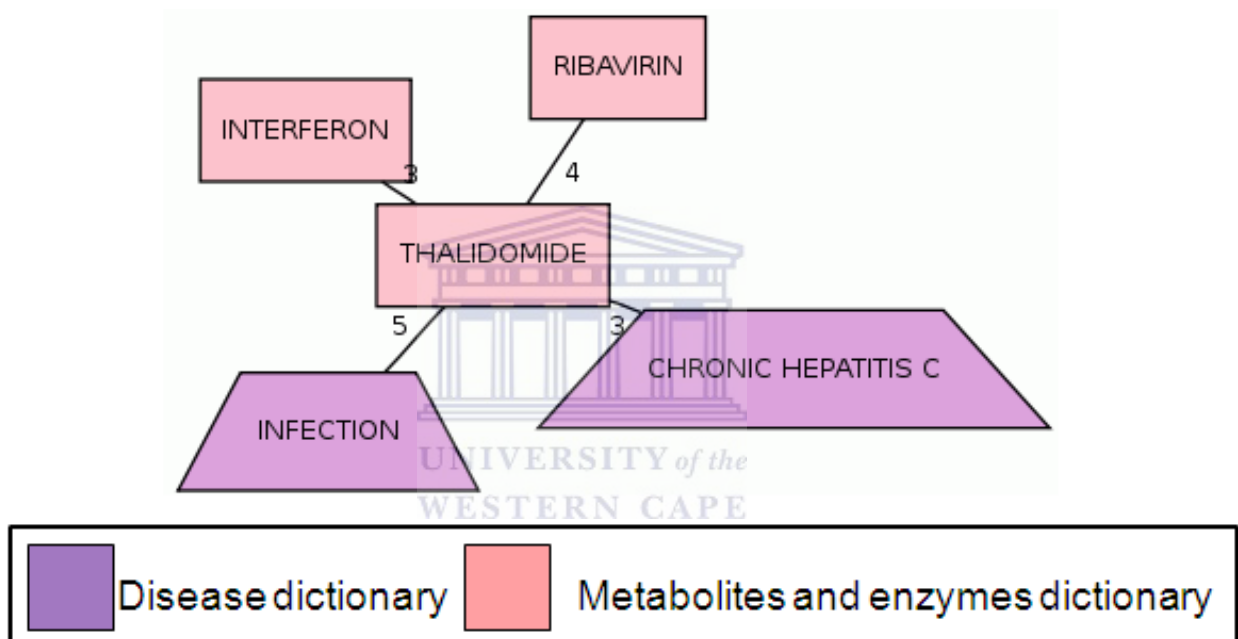
**FIGURE 2.3.** A DIAGRAM DISPLAYING THALIDOMIDE CONCEPT QUERY OUTPUT.
This shows the association map comprising of a network of interaction nodes consisting
of concepts assigned to various biomedical dictionaries.

## 2.4.2 Abstract queries

The abstract search menu allows users to do keywords searches, use quotes for "phrase search", and Boolean logical operators such as "OR", "AND" or "NOT". An abstract search with the keyword "hypervariability" therefore returns a list of abstracts containing the queried keyword "hypervariability". The retrieved abstracts also contain tagged colour-coded concepts assigned to their respective dictionaries. The abstract search results can serve as rich source of data useful for curation since it contains automatically identified biomedical concepts. Even though manual curation using human expertise could yield high quality results (Muller et al., 2004), it is sometimes fraught with errors since the human eyes may unintentionally overlook useful data. In addition, the rapidly increasing volume of published literature sometimes could render it herculean for researchers to manually inspect and efficiently locate or identify biomedical concepts of interest embedded in literature. This feature of DESHCV is aimed at complimenting the already existing information search, retrieval and extraction resources.
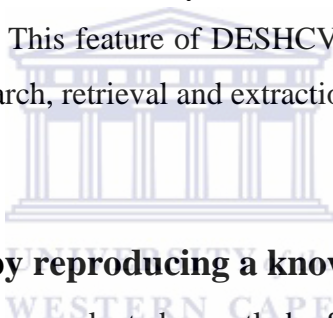
## 2.4.3 Evaluation of DES by reproducing a known hypothesis

The performance of DES has been evaluated recently by Sagar et al. (2008) in the context of sodium channels in human. Since it was not possible to evaluate all the concepts embedded in the 5243 documents used in Sagar's analysis, SCN1A was chosen as a reference gene for the analysis. DES accurately identified most of the concepts present in the 131 abstracts associated with SCN1A. The analysis showed both precision and recall for identified concepts from different dictionaries that ranged between 81% and 100% whilst the *F*-measure was between 87% and 100%.

In this report, a different approach was used to evaluate the performance of DES by simulating an already confirmed scientific discovery. Such approach has been used during the implementation of the DAD-system, Swanson's Raynaud's disease-fish oil discovery was simulated to test the performance of the system (Weeber et al., 2000). Swanson's discovery of a hidden connection between disjunct literatures on magnesium and migraine was successfully re-implemented in the LitLinker systems (Meredith et al.,

2005). Anni biomedical text-mining tool was used to reproduce a previously published thalidomide-chronic hepatitis C discovery (Jelier et al., 2008). Here, DESHCV is used successfully to simulate the thalidomide-chronic hepatitis C association (Weeber et al., 2003). By clicking on the hypothesis generator, the user can retrieve hypothesis to reveal potential relationships existing between concepts. The hypothesis generator query menu allows users to automatically or manually select categories with which to generate hypothesis. The open discovery approach previously described by Swanson and adapted by Meredith has been employed here. Thalidomide (from "metabolites and enzymes" dictionaries) was used as the starting term to retrieve all linking concepts within the human proteins and genes category. The tumor necrosis factor (TNF) was manually selected as the linking term whilst concepts within the diseases catalogues were defined as the target (Figure 2.4). The system successfully generated hypothesis to infer potential relationships between thalidomide and the following disease concepts: chronic liver disease, chronic viral hepatitis, chronic persistent hepatitis and chronic renal failure. The system used disjuncted literature between the different concepts to predict implicit relationship amongst them. Chronic liver disease, chronic viral hepatitis and chronic persistent hepatitis are possible name variants and these conditions are implicated in liver failure. By clicking on the test button linking the chronic viral hepatitis and thalidomide retrieved an abstract on a case report concerning thalidomide-associated hepatitis where the patient had medical history including chronic hepatitis C (Fowler and Imrie, 2001). The hypothesis generated by DESHCV was reasonable since the verification of the literature supported the "discovery".
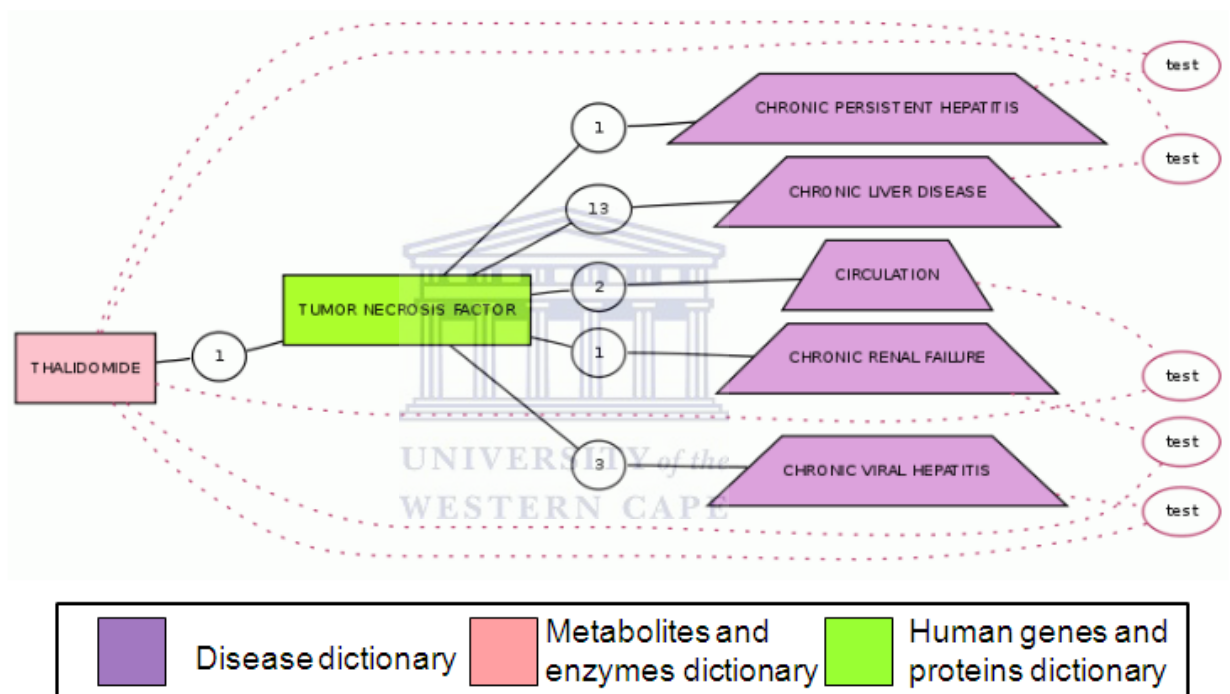
**FIGURE 2.4.** A DIAGRAM DISPLAYING THALIDOMIDE-CHRONIC HEPATITIS C HYPOTHESIS. This shows an implicit relationship between thalidomide and chronic hepatitis C inferring potentially new hypotheses. The biomedical concepts "thalidomide", "tumor necrosis factor" and "chronic hepatitis C" belongs to the "metabolites and enzymes", "human genes and proteins", and 'disease concepts" dictionaries respectively.

## 2.4.4 Generation of thalidomide-amantadine association

DESHCV has also been used to generate potentially new hypotheses, which propose relationships between thalidomide and amantadine as possible combination therapy for chronic viral hepatitis C. All disease concepts associated with thalidomide were retrieved and chronic hepatitis C was selected as linking term with concepts in metabolites and enzymes dictionaries considered as target terms. This hypothesis was tested automatically by checking the auto test radio button. This inferred an implicit relationship between thalidomide and amantadine but no PubMed abstract was retrieved implying a potentially new discovery (Figure 2.5). This means that these two concepts do not co-occur in any of the PubMed abstracts. This potential discovery is based on textual analysis of abstracts and not full text papers. The standard treatment for chronic hepatitis C is a combination therapy of pegylated IFN-alpha to elicit immune response and antiviral effect of ribavirin. Patients infected with certain genotypes of HCV do not respond to this treatment, necessitating the need for enhanced combination therapy. Reports on available data concerning triple therapy comprising of pegylated IFN-alpha, ribavirin and protease inhibitors targeting NS3-4SA protease looks promising, and this could become standard treatment feature in the near future (Zeuzem, 2008; Flisiak and Parfieniuk, 2010). The usage of thalidomide in the treatment of chronic hepatitis C unresponsive to alpha-interferon and ribavirin has been investigated (Caseiro, 2006; Milazzo et al., 2006), whilst amantadine has been combined with interferon-alpha plus ribavirin (Chrissafidou and Musch, 2009; von Wagner et al., 2008).

The possibility of a triple therapy for effective management of chronic hepatitis C should prompt researchers on the need to investigate combining any one of the available treatment drugs (IFN-alpha and ribavirin) or possibly protease inhibitors to thalidomide and amantadine as implied by the generated hypotheses. Although the beneficial effect of amantadine and other antiviral drug for the treatment of chronic hepatitis C is controversial and sometimes contentious, potent analogues with less toxicity, augmented pharmacophore and enhanced pharmacokinetics could be explored further. The possibility of augmenting the above therapy with hepatoprotective drug such as silibinin could be explored. Any discovery proposed using biomedical text-mining approach

should undergo the rigour of laboratory evaluations and ethical consideration and possibly must conform to existing legislation before usage.

UNIVERSITY *of the*
WESTERN CAPE
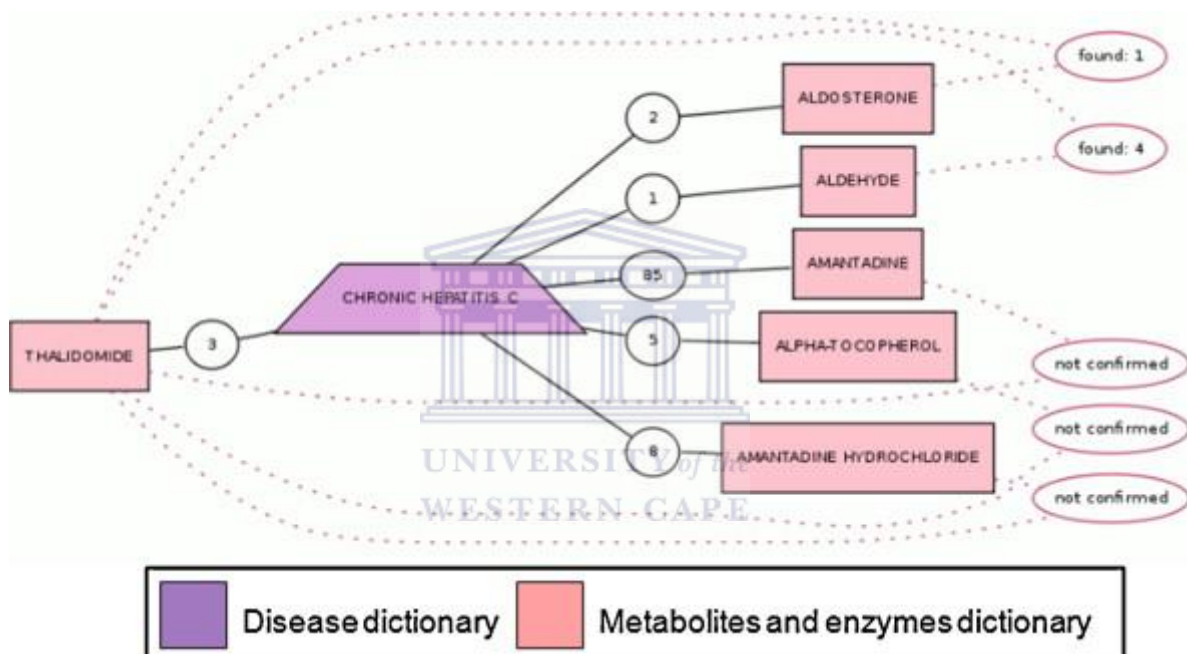
**FIGURE 2.5.** A DIAGRAM DISPLAYING THALIDOMIDE-AMANTADINE HYPOTHESIS.
This shows an implicit relationship between thalidomide and amantadine inferring potentially new hypotheses. The biomedical concepts "thalidomide" and "amantadine" belong to the "metabolites and enzymes" dictionaries whilst "chronic hepatitis C" concept belongs to the "disease" dictionary.

## 2.4.5 Systems reports

Comprehensive summary reports generated from DESHCV are made available on the left menu of the user interface. The "concept list report" groups concepts by dictionaries and frequency of appearance inside documents. The frequency of document report lists abstract with their total number of embedded tagged concepts. The frequencies of pairs report group pairs of concepts that co-occur in documents. Another useful feature is the document clustering report which displays clusters of concepts sorted by frequency of appearance and is compiled using artificial neural networking algorithm. Its usefulness lies in the fact that similar documents tend to cluster together, thereby allowing biologist to harness potential information amongst clustered abstracts. Recommended readings displays the link to top 10 documents with most concepts, though not manually generated they could give an overview of HCV research to a new researcher in virology. For example, the most recommended reading is an abstract on a review describing in detail the current antiviral drugs in clinical usage (De Clercq, 2004).
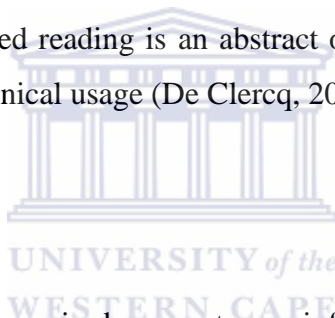
## 2.5 Limitations

Associations generated between paired concepts are inferred from co-occurrences and may not necessarily relate to any molecular functionality. Textual data is obtained from abstracts, which are easy to index. Some details of research are present in full body text and as such vital information may not be reported in abstract. Full text documents were not analyzed in this study.

## 2.6 Future directions

Integrating blast and identifier queries to enhance querying capabilities of DESHCV is currently being investigated. The possibility of integrating full text document is currently being explored and could be added to the database as a separate feature. The database would be updated every six months to meet the demands of ever increasing PubMed records related to HCV.

## 2.7 Conclusions

This chapter described the implementation of Hepatitis C Virus customized web-based text-mining resource that allows researchers to intuitively use the system to get insight into possible novel associations between concepts. This system has been used successfully to reproduce already published thalidomide-chronic hepatitis C biomedical text-mining discovery (Weeber et al., 2003). DESHCV database is free to use for academic and non-profit purposes.

## 2.8 References

Bajic, V. B., Veronika, M., Veladandi, P. S., Meka, A., Heng, M. W., Rajaraman, K., et al. (2005). Dragon Plant Biology Explorer. A text-mining tool for integrating associations between genetic and biochemical entities with genome annotation and biochemical terms lists. *Plant Physiol, 138*(4), 1914-1925.

Caseiro, M. M. (2006). Treatment of chronic hepatitis C in non-responsive patients with pegylated interferon associated with ribavirin and thalidomide: report of six cases of total remission. *Rev Inst Med Trop Sao Paulo, 48*(2), 109-112.

Chen, Y., & Han, K. (2009). BSFINDER: finding binding sites of HCV proteins using a support vector machine. *Protein Pept Lett, 16*(4), 373-382.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., & Wishart, D. S. (2008). PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res, 36*(Web Server issue), W399-405.

Chrissafidou, A., & Musch, E. (2009). [Peripheral polyneuropathy and bilateral optic neuropathy during treatment of chronic hepatitis C]. *Dtsch Med Wochenschr, 134*(18), 927-930.

Cohen, A. M., & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Brief Bioinform, 6*(1), 57-71.

Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. *PLoS Comput Biol, 4*(1), e20.

Combet, C., Garnier, N., Charavay, C., Grando, D., Crisan, D., Lopez, J., et al. (2007). euHCVdb: the European hepatitis C virus database. *Nucleic Acids Res, 35*(Database issue), D363-366.

de Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugue, S., et al. (2008). Hepatitis C virus infection protein network. *Mol Syst Biol, 4*, 230.

De Clercq, E. (2004). Antiviral drugs in current clinical use. *J Clin Virol, 30*(2), 115-133.

Essack, M., Radovanovic, A., Schaefer, U., Schmeier, S., Seshadri, S. V., Christoffels, A., et al. (2009). DDEC: Dragon database of genes implicated in esophageal cancer. *BMC Cancer, 9*, 219.

Flisiak, R., & Parfieniuk, A. (2010). Investigational drugs for hepatitis C. *Expert Opin Investig Drugs, 19*(1), 63-75.

Fowler, R., & Imrie, K. (2001). Thalidomide-associated hepatitis: a case report. *Am J Hematol, 66*(4), 300-302.

Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G., & Kors, J. A. (2008). Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol, 9*(6), R96.

Kaur, M., Radovanovic, A., Essack, M., Schaefer, U., Maqungo, M., Kibler, T., et al. (2009). Database for exploration of functional context of genes implicated in ovarian cancer. *Nucleic Acids Res, 37*(Database issue), D820-823.

Killcoyne, S., Carter, G. W., Smith, J., & Boyle, J. (2009). Cytoscape: a community-based framework for network modeling. *Methods Mol Biol, 563*, 219-239.

Kuiken, C., Mizokami, M., Deleage, G., Yusim, K., Penin, F., Shin, I. T., et al. (2006). Hepatitis C databases, principles and utility to researchers. *Hepatology, 43*(5), 1157-1165.

Kuiken, C., Yusim, K., Boykin, L., & Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics, 21*(3), 379-384.

Kwofie, S. K., Radovanovic, A., Maqungo, M., Bajic, V.B., Christoffels, A. (2009, December 01–03, 2009). *Hepatitis C Virus discovery database (HCVdd): a biomedical text mining and relationship exploring knowledgebase.* Abstract presented at the ISCB Africa ASBCB Joint Conference on Bioinformatics of Infectious Diseases. Bamako, Mali.

Lombard, Z., Tiffin, N., Hofmann, O., Bajic, V. B., Hide, W., & Ramsay, M. (2007). Computational selection and prioritization of candidate genes for fetal alcohol syndrome. *BMC Genomics, 8*, 389.

Meredith, M., Skeels, H., Kiera, Y.Y., Meliha, Wanda, P. (2005, April 02–07, 2005). *Interaction design for literature-based discovery.* Paper presented at the CHI '05 Extended Abstracts on Human factors in Computing Systems Portland, USA

Milazzo, L., Biasin, M., Gatti, N., Piacentini, L., Niero, F., Zanone Poma, B., et al. (2006). Thalidomide in the treatment of chronic hepatitis C unresponsive to alfa-interferon and ribavirin. *Am J Gastroenterol, 101*(2), 399-402.

Mitchell, J. A., Aronson, A. R., Mork, J. G., Folk, L. C., Humphrey, S. M., & Ward, J. M. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu Symp Proc*, 460-464.

Muller, H. M., Kenny, E. E., & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol, 2*(11), e309.

Pan, H., Zuo, L., Choudhary, V., Zhang, Z., Leow, S. H., Chong, F. T., et al. (2004). Dragon TF Association Miner: a system for exploring transcription factor associations through text-mining. *Nucleic Acids Res, 32*(Web Server issue), W230-234.

Sagar, S., Kaur, M., Dawe, A., Seshadri, S. V., Christoffels, A., Schaefer, U., et al. (2008). DDESC: Dragon database for exploration of sodium channels in human. *BMC Genomics, 9*, 622.

Shi, L., & Campagne, F. (2005). Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics, 6*, 88.

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., & Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease

gene candidates. *Nucleic Acids Res, 33*(5), 1544-1552.

von Wagner, M., Hofmann, W. P., Teuber, G., Berg, T., Goeser, T., Spengler, U., et al. (2008). Placebo-controlled trial of 400 mg amantadine combined with peginterferon alfa-2a and ribavirin for 48 weeks in chronic hepatitis C virus-1 infection. *Hepatology, 48*(5), 1404-1411.

Weeber, M., Klein, H., Aronson, A. R., Mork, J. G., de Jong-van den Berg, L. T., & Vos, R. (2000). Text-based discovery in biomedicine: the architecture of the DAD-system. *Proc AMIA Symp*, 903-907.

Weeber, M., Kors, J. A., & Mons, B. (2005). Online tools to support literature-based discovery in the life sciences. *Brief Bioinform, 6*(3), 277-286.

Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L. T., Aronson, A. R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J Am Med Inform Assoc, 10*(3), 252-259.

Yusim, K., Richardson, R., Tao, N., Dalwani, A., Agrawal, A., Szinger, J., et al. (2005). Los alamos hepatitis C immunology database. *Appl Bioinformatics, 4*(4), 217-225.

Zeuzem, S. (2008). Interferon-based therapy for chronic hepatitis C: current and future perspectives. *Nat Clin Pract Gastroenterol Hepatol, 5*(11), 610-622.

Zeuzem, S. (2008). Interferon-based therapy for chronic hepatitis C: current and future perspectives. *Nat Clin Pract Gastroenterol Hepatol, 5*(11), 610-622.
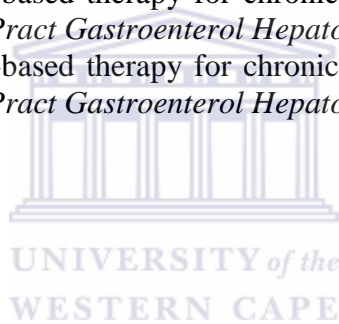
**Chapter 3: Hepatitis C Virus Protein Interaction Database (HCVpro)**
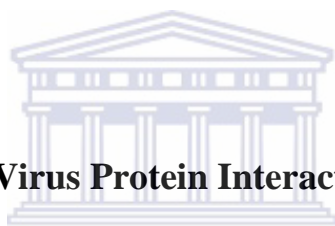
## 3.1 ABSTRACT

It is essential to catalog characterized Hepatitis C Virus (HCV) protein-protein interaction (PPI) data and the associated plethora of vital functional information to augment the search for therapies, vaccines and diagnostic biomarkers. Hepatitis C Virus Protein Interaction Database (HCVpro) has been developed in furtherance of these goals. HCVpro contains manually verified hepatitis C virus-virus and virus-human protein interactions curated from literature and databases. HCVpro is a comprehensive and integrated HCV-specific knowledgebase housing consolidated information on PPIs, functional genomics and molecular data obtained a from variety of virus databases (VirHostNet, VirusMint, HCVdb and euHCVdb), and from BIND and other relevant biology repositories. HCVpro is further populated with information on hepatocellular carcinoma (HCC) related genes that are mapped onto their encoded cellular proteins. Incorporated proteins have been mapped onto Gene Ontologies, canonical pathways, Online Mendelian Inheritance in Man (OMIM) and extensively cross-referenced to other essential annotations. The database is enriched with exhaustive reviews on structure and functions of HCV proteins, current state of drug and vaccine development, and links to recommended journal articles. Users can query the database using specific protein identifiers (IDs), chromosomal locations of a gene, interaction detection methods, indexed PubMed sources as well as HCVpro, BIND and VirusMint IDs. The use of HCVpro is free and the resource can be accessed via http://apps.sanbi.ac.za/hcvpro/ or http://cbrc.kaust.edu.sa/hcvpro/.

## 3.2 Introduction

Protein-protein interactions (PPIs) are vital in orchestrating molecular cellular events and constitute the basis for multitudes of signal transduction pathways and transcriptional regulatory networks (Raman, 2010). Efforts towards elucidating HCV protein interactions through the use of proteome-wide mapping techniques and the computational construction of the entire HCV-human protein interactome map have been reported (Flajolet et al., 2000; de Chassey et al., 2008). Additionally, several other HCV research

groups have published large amounts of information on HCV PPIs and related enriched functionally analysed data vital for the understanding of HCV infection mechanisms. This progress has been facilitated by the advent of highly innovative low and high throughput techniques essential for the elucidation and characterization of protein interactions. Various experimental methods have been widely used to either screen or confirm HCV protein-protein interactions, including yeast two-hybrid (Dimitrova et al., 2003; Ahn et al., 2004; Huang et al., 2005), confocal microscopy (Tong et al., 2002; Lan et al., 2003), coimmunoprecipitation (Dimitrova et al., 2003; Wang et al., 2005), surface plasmon resonance studies (Sabile et al., 1999; Jennings et al., 2008), spectroscopy (Kang et al., 2005; Masumi et al., 2005), immunoblotting (Tan et al., 1999), cross-linking (Wang et al., 2002), affinity chromatography (Masumi et al., 2005) and 3-dimensional protein structures (Op De Beeck et al., 2000). The use of several types of methods in elucidating PPIs has led to the availability of interaction data in heterogeneous formats thus requiring standardization. In response, numerous efforts have focused on formatting the ever-increasing protein-protein interactions and associated meta-data as well as integrating them into public online repositories. The proteomics standards initiative-molecular interaction (PSI-MI) format (Hermjakob et al., 2004), a community standard data model for the representation and exchange of protein interaction data, and MIMIx, the minimum information required for reporting a molecular interaction experiment guidelines (Orchard et al., 2007) are platforms for harnessing information from structured PPI data.

Towards consolidating viral PPI data, VirusMint (Chatr-aryamontri et al., 2009) and VirHostNet (Navratil et al., 2009) databases have been developed to house viral PPI data. VirHostNet represents more than 180 distinct viral species with unique taxonomy identifications as well as a broad diversity spanning about 36 distinct viral families. VirusMint represents more than 110 different viral strains and provides PPI information on viruses involved in human infections and cancer, including adenovirus, Simian virus 40 (SV40), human papilloma viruses, Epstein–Barr virus (EBV), hepatitis B virus (HBV), hepatitis C virus (HCV), herpes viruses, influenza A virus, vaccinia virus and human immunodeficiency virus (HIV). Primarily, both VirusMint and VirHostNet are

geared towards online network visualization and analysis of integrated human viral interactome maps. A trend is emerging where databases solely dedicated to single organisms are constructed instead of multi-organism PPI databases, and a typical example is HIV-1 Human Protein Interactions Database (Fu et al., 2009). It is believed that a knowledgebase solely focused or customized to answer pertinent biological questions related to HCV has the potential to provide detailed and exhaustive information suitable to generate functional hypotheses *in-silico*. In this regard, this thesis describes the development of a new online hepatitis C virus protein interaction database (HCVpro), an integrated and comprehensive warehouse for storing and managing manually verified HCV virus-virus and virus-human protein interactions curated from literature and databases. HCVpro stores interaction information on human and viral interacting partners as well as meta-data, including interaction detection methods and journal articles reporting the interactions. Additionally, HCVpro contains useful contextual data and is distinct from other curated viral PPI databases because it provides: (1) consolidated PPI data from BIND (Alfarano et al., 2005) and key viral databases comprising of VirusMint and VirHostNet, (2) an integrated query platform enabling users to retrieve interaction data using HCVpro and other PPI database identifiers (BIND and MINT IDs), experiment types and journal article sources, and other genomic qualifiers, (3) useful genomic and functional information from notable HCV knowledgebases, including VBRC HCVdb (Greene et al., 2007), and euHCVdb (Combet et al., 2007), (4) compiled current knowledge on the molecular functions and biochemical features of HCV proteins, including post-translational modifications, subunit structure, domains, subcellular localization, diseases and links to essential PubMed articles, (5) a detailed review on viral drug development and challenges particularly concerning current therapy options and state of antiviral discovery, (6) genes transcriptionally regulated by NetPath curated signaling pathways (Kandasamy et al., 2010) as well as information concerning canonical pathways comprising of NCI-Nature curated pathways (Schaefer et al., 2009), KEGG (Kanehisa et al., 2010), Reactome (Matthews et al., 2009) and HPD (Chowbina et al., 2009), (7) links to external databases such as HPRD (Keshava Prasad et al., 2009), OMIM (Amberger et al., 2009), PharmGKB (Hodge et al., 2007), PDB (Berman et al., 2002), Gene Ontologies (Gene Ontology Consortium, 2010) and other essential

biological knowledgebases, and (8) comprehensive integration of high-confidence HCV-related hepatocellular carcinoma (HCC) genes deciphered from gene expression and proteomics studies, HCC gene expression pattern from EHCO (Hsu et al., 2007) and links to other major HCC databases consisting of HCCNet (He et al., 2010) and OncoDB.HCC (Su et al., 2007).

The diverse integrated information in HCVpro can accelerate the understanding of HCV mediated molecular mechanisms and also enable postulation of appropriate hypothesis to guide discovery of new knowledge of potential use in drug discovery. Additionally, HCVpro can augment efforts towards uncovering novel HCV-related tumorigenic biomarkers with therapeutic potential by harnessing integrated data composed of PPIs, canonical pathways, and HCC-related gene expression information on the genes encoding the host cellular proteins.

## 3.3 Construction and content

## 3.3.1 Protein interaction data curation and integration

Hepatitis C Virus-virus and virus-human protein interactions were manually curated from published peer-reviewed journal articles and database sources including BIND, VirusMint and VirHostNet. Protein interactions were annotated with associated meta-data such as experiment types used in inferring interactions and PubMed identification numbers (PMIDs) of journal articles reporting the interactions. The Entrez gene symbols of the human genes encoding the interacting cellular proteins were used as protein identifiers as well as their assigned non-redundant gene ID. The interacting viral proteins consist of 11 distinct proteins encoded by the HCV genome comprising core, E1, E2, p7, NS2, NS3, NS4A, NS4B, NS5A, NS5B and F proteins. It must be noted that existing curation protocols were followed with particularly emphasis on that of VirHostNet (Navratil et al., 2009). Interactions curated from literature must meet the following baseline criteria to be considered: (i) physical binary interactions between HCV-HCV proteins or HCV-human proteins only; (ii) the interactions must not be predicted (iii) interactions merely mentioned in the review section of a journal article are rejected; and

(iv) there should be sufficient information in the manuscript to support the reported interaction, for example the names of the interacting protein partners and the detection methods used to infer the interactions. Interactions obtained from database sources must meet the following extra criteria: (i) the source database must be published in a peer-reviewed journal and details of the curation procedures must be accessible; (ii) each provided interaction must be assigned a unique interaction ID; (iii) interaction must be annotated with the minimum information comprising the journal article reporting the interactions, experimental methods used for detecting the interaction and protein names as well as their IDs. Interactions that do not meet the above criteria are not considered for incorporation into HCVpro.

For the purpose of HCVpro construction, each annotated protein-protein interaction curated from a specific journal article was assigned a unique HCVpro ID (appendix III). Therefore, a single type of interaction being reported by many journal sources was assigned multiple HCVpro IDs. For example, the binary interactions between HCV core and LTBR proteins were assigned two separate HCVpro IDs hcv0001 and hcv0004 since they have been reported by two different journal articles (Matsumoto et al., 1997) and (Chen et al., 1997), respectively. The hcv0001 has been annotated with interaction meta-data consisting of PubMed ID 8995654, BIND ID 133606, and experiment types comprising of yeast two-hybrid, far-western blot and affinity chromatography. Similar to aforementioned convention, hcv0004 has been annotated with interaction meta-data consisting of PubMed ID 9371602, BIND IDs 181950 and 182177, and experiment types comprising of yeast two-hybrid, affinity chromatography and coimmunoprecipitation. As evident, BIND database appears to use a similar convention for annotation as core-LTBR interactions reported from separate journal articles were assigned different BIND IDs. In addition, different types of interactions being reported in a single journal article are assigned different HCVpro IDs. For example, a total of 307 unique viral-human protein interactions were obtained from proteome-wide analysis studies (de Chassey et al., 2008), although these interactions were curated from a single journal source (PubMed ID 18985028), each was assigned a unique HCVpro ID. The unique HCVpro IDs assigned to multiple protein interactions, number of experiment types confirming the interactions and

multiple journal articles reporting the interactions could serve as parameters to evaluate the confidence of HCV molecular interaction data. These parameters could, for example, serve as weights for scoring binary interactions between two proteins and may be used to filter out interactions with a lower confidence threshold. Comprehensive reviews of PPI detection methods and challenges involved in curation have been presented elsewhere (Phizicky and Fields, 1995; Alfarano et al., 2005; Mathivanan et al., 2006; Berggard et al., 2007; Cusick et al., 2009; Turinsky et al., 2010).

The curated experimental methods describing PPIs were converted into PSI-MI 2.5 controlled vocabularies and assigned PSI-MI 2.5 Term IDs (Hermjakob et al., 2004) using the EMBL-EBI Ontology Lookup Service (http://www.ebi.ac.uk/ontology-lookup/) (Barsnes et al., 2010). Similar to the approach adopted by VirHostNet, HCVpro provide information concerning the two interacting protein partners, experiment type PSI-MI Term IDs and PubMed IDs of articles reporting the interactions. The PPI data can be downloaded as a tab-delimited file enabling users to convert to any machine-readable format of their choice with ease. The downloadable data could be harnessed for functional enrichment analysis and the interacting proteins could be leveraged as seed nodes for modeling protein network and pathways using third-party software such as Cytoscape and its associated plug-ins (Smoot et al., 2010). Furthermore, the data could augment efforts towards an integrated bioinformatics approach where the manually verified interactions could be employed as datasets for training machine learning models to predict protein-protein interactions as well as used in combination with a plethora of genomic data to prioritize and identify potential candidate hepatocellular carcinoma related genes, therapeutic targets and diagnostic markers associated with HCV infections. Due to HCV genetic heterogeneity, this repository provides another downloadable dataset supplemented with information on HCV genotypes/strains from which the HCV proteins were characterized and obtained from published literature, BIND and VirHostNet (appendix IV). This data can be utilized in modeling of genotype- or strain-specific interaction networks and could potentially aid in simulating the effect of genetic variability on pathobiology of HCV associated diseases.

## 3.3.2 HCV and human protein contextual data sources

There is an emerging consensus where HCV strain 1a (Isolate H77) has been adopted as the reference HCV (Kuiken and Simmonds, 2009), and therefore the HCV protein knowledgebase, sequences and annotations housed in VBRC HCVdb are overwhelmingly based on this strain. This approach was adopted in the development of HCVpro where HCV protein data integrated from Uniprot (UniProt Consortium, 2010), euHCVdb (Combet et al., 2007), VBRC HCVdb (Greene et al., 2007), GenBank (Benson, et al., 2009) and Entrez gene (Maglott et al., 2007) are based on this strain. Other biologically relevant HCV data were obtained from OMIM (Amberger et al., 2009), PDB (Berman et al., 2002), Gene Ontology (Gene Ontology Consortium, 2010) and published literature reviews (Penin et al., 2004; Krekulova et al., 2006; Dubuisson, 2007; Sharma, 2010). Additionally, enriched human protein functional data was obtained from Entrez gene (Maglott et al., 2007), HPRD (Keshava Prasad et al., 2009), Ensembl (Flicek et al., 2008), GenBank (Benson et al., 2009), Uniprot (Uniprot Consortium, 2010), Gene Ontology (Gene Ontology Consortium, 2010), GeneCards (Safran et al., 2010), Gene ID converter (Alibes et al., 2007), Netpath (Kandasamy et al., 2010) and Nature curated pathways (Schaefer et al., 2009). Hepatocellular carcinoma associated genes were consolidated from the following databases (appendix V): HCCNet (He et al., 2010), OncoDB.HCC (Su et al., 2007) and EHCO (Hsu et al., 2007),

In total, HCVpro consist of 621 unique HCVpro IDs comprising of 549 HCV viral-human protein and 72 viral-viral protein interactions. The detection method and number of method distributions are shown in Figures 3.1 and 3.2 respectively. Comprehensive information on HCVpro data statistics is also provided on the HCVpro website under the menu item 'Statistics' (appendix VI).
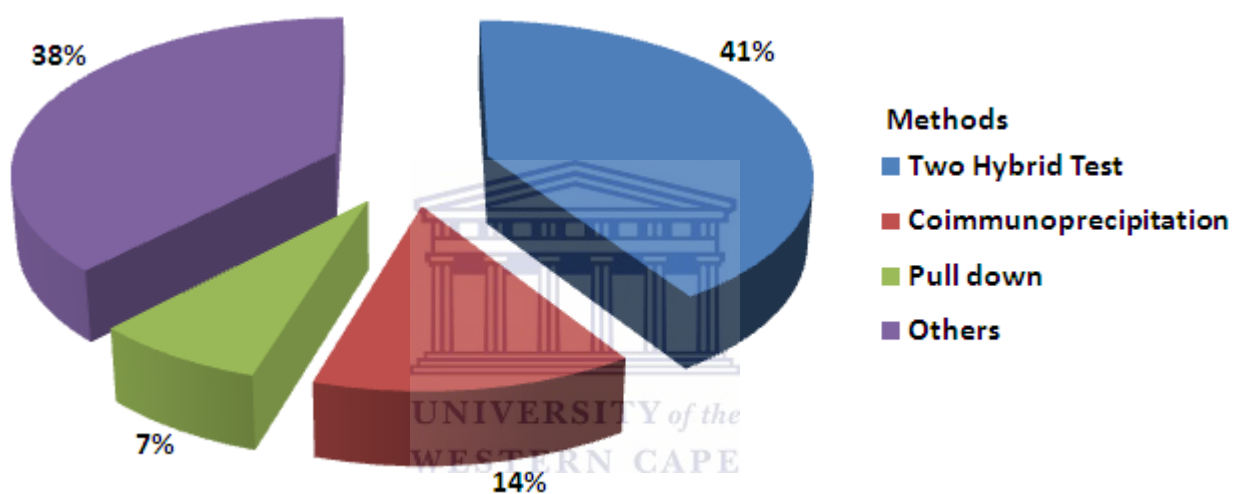
**FIGURE 3.1.** DISTRIBUTION OF EXPERIMENTAL METHODS TO VERIFY PPIs.
A pie chart showing the detailed breakdown in percentage of the experiment types used in verifying the protein-protein interactions. The majority of interactions were detected using two-hybrid test (41%). Coimmunoprecipitation, pull down and other experimental techniques constituted 14%, 7% and 38% respectively.

**FIGURE 3.2.** PROTEIN-PROTEIN INTERACTION VERIFICATION METHOD DISTRIBUTION.
A pie chart showing the detailed breakdown in percentage of the number of methods used jointly for verifying the protein-protein interactions. 64% of detected interactions were inferred by single experimental method, 16% were inferred by two methods whiles 20% were inferred by 3 or more methods.

### 3.3.3 Database architecture

HCVpro is a web-based resource running on platform independent Apache HTTP server with MySQL Server as the back-end and open source application programs such as PHP, HTML, CSS and JavaScript as the front-end. The database contents are stored in MySQL relational database tables and the data fields constituting the tables are linked by unique or non-redundant keys to enable cross table integration, multithreading, multipurpose and multiuser querying environments.

## 3.4 Utility and discussion

### 3.4.1 Database usage

This chapter describes the implementation of a biological repository known as HCV protein interaction database (HCVpro) that integrates a multitude of enriched data from a variety of protein and genomic platforms relevant to HCV research. This knowledgebase is intended to serve as a "one stop shop" for retrieval of comprehensive data on HCV-HCV and HCV-human protein interactions, and associated vital information relevant for functional characterization experiments. The search menu on the index page of HCVpro enables users to navigate through the enhanced or simplified user query interfaces. The user query interfaces enable users to perform searches via the HCVpro by browsing three sub query menu tabs: (i) protein select, (ii) evidence select, and (iii) identifier search. Additionally, HCVpro enables retrieval of data via string searches using gene IDs, names and symbols (Figure 3.3).
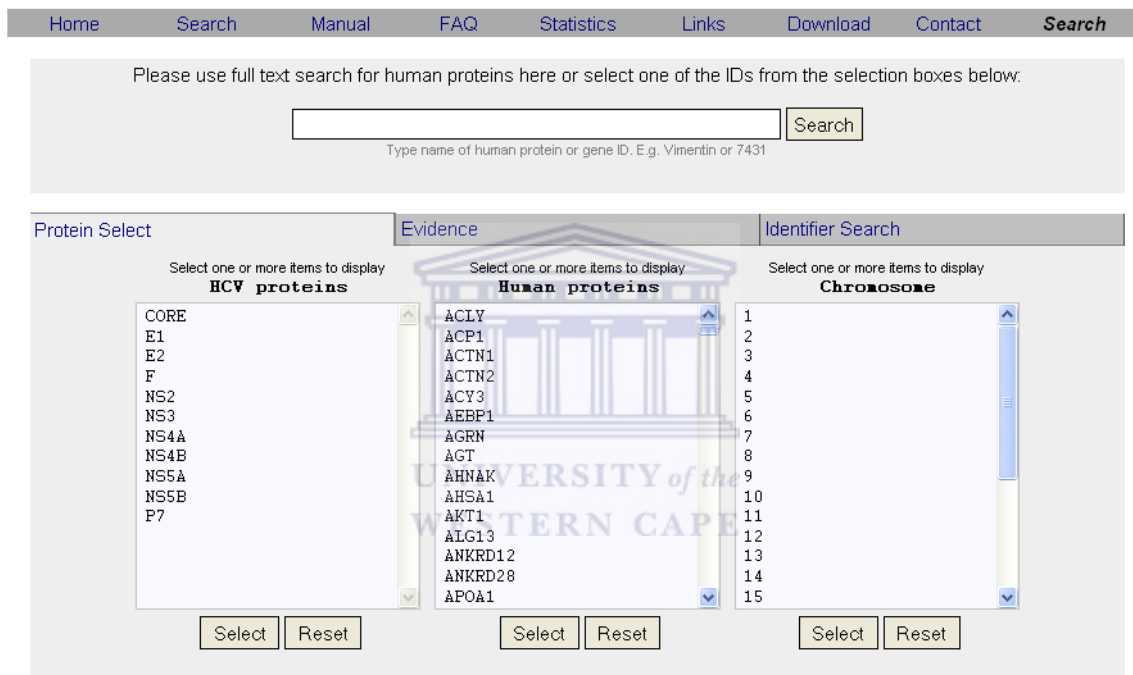
**FIGURE 3.3.** A DIAGRAM DISPLAYING THE HCVPRO USER QUERY INTERFACE.
This depicts a screenshot of the integrated query platform enabling users to retrieve interaction data using database identifiers, genomic qualifiers and information on experiment types as well as PubMed IDs of journal articles reporting the interactions.

### 3.4.1.1 Protein Select

A search via the "protein select" sub-menu by browsing any of the HCV proteins returns a list of entries comprising of HCV-HCV and HCV-human protein interaction data as well as relevant contextual data associated with both HCV and human proteins participating in the binary interactions. Users can also perform searches by using gene names of proteins to retrieve a list of viral-human protein interaction data. For example, a query with the protein GRB2 returned an entry result composed of an interaction between the oncogenic HCV protein NS5A and cellular protein GRB2 with cross-referenced links to external protein interaction databases consisting of BIND and VirusMint. The binary interaction between the two proteins was confirmed using affinity chromatography, yeast-two hybrid, coimmunoprecipitation, and immunoblotting experiments (Tan et al., 1999). An additional search option via "chromosome search" allows users to retrieve a list of genes located on a particular chromosome, which interact with HCV proteins.
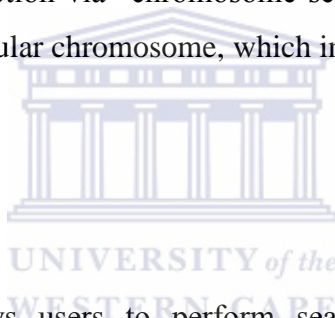
### 3.4.1.2 Evidence Search

The "evidence search" allows users to perform searches via HCVpro using the experiment types used in inferring the interactions and the PubMed sources. By performing a search using the experimental evidence, for example coimmunoprecipitation, HCVpro returns a list of interactions that were confirmed by using coimmunoprecipitation as the experimental technique. This search menu is particularly useful during characterization of protein interactions. At a glance experimental biologists would narrow down the lists of techniques that may be used to elucidate interaction with any of the HCV proteins. For example, knowing the experimental evidence for a gene may aid in choosing an appropriate experiment type for orthologous genes or genes sharing a variety of genomic features with the retrieved genes. Users can also query HCVpro by clicking on a PubMed ID to retrieve all protein-protein interactions reported in the corresponding journal article present in HCVpro.

### 3.4.1.3 Identifier Search

The "identifier search" option enables users to search HCVpro using HCVpro, VirusMint and BIND IDs. This search option therefore returns the list of interactions annotated with HCVpro ID and also integrated from both MINT and BIND databases. Allowing users to query HCVpro with other protein interaction database identifiers represents an integrated platform for data retrieval. For example, querying HCVpro using VirusMint ID MINT-16518 returned a binary interaction between NS5A and NS4A proteins confirmed using coimmunoprecipitation. This entry is designated as hcv0437 with linked BIND ID 130456. Users may click the links to freely access other relevant interaction meta-data provided by the external databases even though resources such as BIND require user registration.

### 3.4.2 Value of the database to prediction of diagnostic biomarkers

The usefulness of the integrated canonical pathways, Gene Ontologies and microarray expression data for exploration of HCV infection associated biomarkers having therapeutic potential has been demonstrated using the protein vimentin (for more details refer to Section C of the online manual and Figure 3.4). Vimentin was chosen as an example query since its potential as a tumor biomarker for HCV-associated hepatocellular carcinoma is widely studied. Vimentin-overexpressing and vimentin-knocked-down experiments showed that cellular vimentin expression enhanced the proteasomal degradation of HCV core protein and eventually restricted HCV particle production (Nitahara-Kasahara et al., 2009). Performing a string search with the term "vimentin" via HCVpro retrieves interaction of vimentin with key oncogenic proteins HCV NS3 and core proteins. The interaction with NS3 protein was established using high throughput yeast two-hybrid screening, whilst immunoblotting and immunofluorescence microscopy were employed with other techniques to confirm interactions with HCV core protein. EHCO microarray expression data incorporated in HCVpro indicated vimentin to be significantly upregulated in HCV associated hepatocellular carcinoma (HCV-HCC) samples in 3 independent datasets (PubMed, mRNA, Protein). Clicking the HCCNet and

OncoDB.HCC links in the user interface of HCVpro retrieved information corroborating the expression levels of vimentin in HCC samples. Additionally, pathway examination pinpointed vimentin as one of the genes transcriptionally upregulated by the NetPath curated immune related IL-2 pathway (Kovanen et al., 2003). The IL-2 pathway is associated with HCV-host immunologic mechanisms (Morshed et al., 1993; Serti et al., 2010). Furthermore, IL-2 is involved in the activation of the Ras/MAPK, JAK/Stat and PI 3-kinase/Akt signaling modules. Interestingly, the JAK/Stat signaling module is a well characterized HCV infection pathway (Basu et al., 2001; Luquin et al., 2007). Vimentin is also involved in HCV infection associated TLR4 signaling (Agaugue et al., 2007) and innate immunity signaling pathways with Reactome IDs 166016 and 168249 respectively. Vimentin has been mapped onto Gene Ontology terms comprising of structural constituents of cytoskeleton, cellular component movement and protein binding. Particularly noteworthy are the cytoskeleton constituents comprising of microtubules and filaments which are implicated in the provision of tracks for the movement of HCV replication complexes (Lai et al., 2008). The significant gene expression levels in HCC patient samples coupled with binary protein interaction with notable oncogenic HCV proteins and transcriptional regulation by IL-2 signaling pathway, provide additional hints that can entice researchers to explore further the potential of vimentin as a plausible diagnostic biomarker or therapeutic target for chronic hepatitis C related HCC.

To further investigate the relevance of HCVpro to candidate disease biomarker discovery, the diagnostic utilities of the other HCVpro integrated cellular proteins were accessed. Even though changes in gene expression levels sometimes may not necessarily correlate well with protein levels in hepatic tissues or serum (Caillot et al., 2009), it could serve as a potentially useful diagnostic indicator. HCC related genes reported in all the three major HCC databases (OncoDB.HCC, HCCNet and EHCO) were selected. In all, 23 genes passed this stringent filter and literature was explored to determine if the genes have been reported previously as potential biomarkers. Six genes comprising of *APOE* (Yokoyama et al., 2006), *CD5L* (Gangadharan et al., 2007), *CTGF* (Kovalenko et al., 2009), *FAS* (El Bassiouny et al., 2008), *TP53* (Jeng et al., 2000; Peng et al., 2004; El

Bassiouny et al., 2008) and *CD81* (Schoniger-Hekele et al., 2005) were suggested or exhibited potential as biomarkers earlier, whilst nine consisting of *APOA1* (Yokoyama et al., 2006), *CDKN1A* (Kasprzak et al., 2009), *CTSB* (Lee et al., 2009), *FN1* (Yoon et al., 2006), *HSPD1* (Looi et al., 2008), *KRT19* (Chang et al., 2009), *MCL1* (Sieghart et al., 2006), *SERPINF2* (Caillot et al., 2009), and *TF* (Mas et al., 2009) have been explored in biomarker research. It appears no article has reported the diagnostic exploration or potential of the remaining eight genes comprising of *ACLY*, *AZGP1*, *DDX3X*, *FGG*, *H19*, *SIAH1*, *SERPING1* and *THBS1*, and these can serve as a rich source of data for further validation. Additionally, some of the 23 genes are associated with cancer related pathways and Gene Ontology terms. Although the approach adopted here is purely qualitative, the above discourses have amply demonstrated how an integrated repository such as HCVpro could be used to harness already existing data to elucidate predictive biomarkers with plausible therapeutic potential.

| | |
|---|---|
| HCCNet Link | http://www.megabionet.org/hcc/detail.php?symbol=VIM |
| EHCO Dataset | PubMed, mRNA, Protein |
| EHCO UP regulated | 3 |
| EHCO Down regulated | 0 |
| Oncodb.hcc link | http://oncodb.hcc.ibms.sinica.edu.tw/hcc/display_by_gene.cgi?stable_id=ENSG00000026025&cut_off=1&option=simple |
| KEGG | |
| Reactome Pathway ID | 166016, 166054, 166058, 168138, 168142, 168176, 168179, 168181, 168188, 168249, 168256, 168898, 181438 |
| Nature Pathway | |
| Nature Pathway ID | |
| NCI Nature curated pathways links | |
| HPD link | |
| Netpath ID | NetPath_14 |
| Gene transcriptionally regulated by Netpath Pathway | IL-2 Signaling Pathway |
| Netpath Description | Interleukin-2 belongs to a family of cytokines, which includes IL-4, IL-7, IL-9, IL-15 and IL-21. IL-2 signals through a receptor complex consisting of IL-2 specific IL-2 receptor alpha (CD25), IL-2 receptor beta (CD122) and a common gamma chain (?c), which is shared by all members of this family of cytokines. Binding of IL-2 activates the Ras/MAPK, JAK/Stat and PI 3-kinase/Akt signaling modules. |
| Netpath Links | http://www.netpath.org/pathways?path_id=NetPath_14 |
| Gene Ontology | GO:0005198, GO:0005200, GO:0005515, GO:0005737, GO:0005856, GO:0005882, GO:0006928, GO:0045103 |

**FIGURE 3.4.** A SCREENSHOT OUTPUT AFTER DATABASE SEARCH WITH VIMENTIN.
A diagram displaying a portion of information on vimentin as well as links to integrated canonical pathways, Gene Ontologies and hepatocellular carcinoma databases.

### 3.4.3 Additional system features

For ease of usage and understanding of the various data fields incorporated in HCVpro, a user manual and frequently asked questions (FAQ) have been provided via the corresponding menu items on the HCVpro website (appendix VII and VIII). Users may send feedback by writing to the authors via the information given under 'Contact' and also submit novel interactions either before or after publications in MIMIx format (Orchard et al., 2007). In addition, links to biologically relevant HCV related and other protein interactions resources are provided to facilitate retrieval and harnessing of enriched biological data. The links are comprised of VirusMint (Chatr-aryamontri et al., 2009), VirHostNet (Navratil et al., 2009), BSFINDER (Y. Chen and Han, 2009), euHCVdb (Combet et al., 2007), LANL HCV-db (Kuiken et al., 2005), HVDB (Shin et al., 2008), VBRC HCVdb (Greene et al., 2007), BIND (Alfarano et al., 2005), DIP (Salwinski et al., 2004), HPRD (Keshava Prasad et al., 2009), INTACT (Kerrien et al., 2007), MINT (Ceol et al., 2010) and DESHCV (Kwofie et al., 2011).
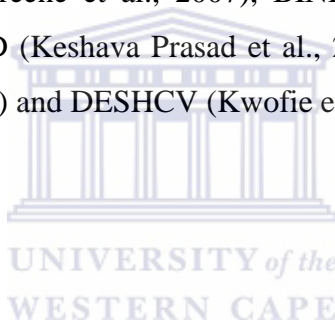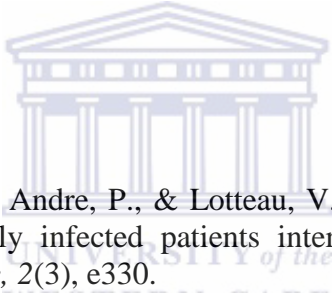
### 3.5 Future prospects

A content curator and database administrator will constantly be improving the database based on user comments and updates will be released twice every year to accommodate the ever-growing published protein interactions. Future update will explore the possibility of incorporating features that compute the druggability of the integrated interactions, and also allow the online screening of potential drugs and cellular drug targets of interest in HCV research into HCVpro. Furthermore, data on interaction details such as domains, motifs and residues involved in the interaction will be incorporated. HCVpro will be integrated with Java web applets that support the visualization and analyze interactions such as the Java Universal Network/Graph Framework (http://jung.sourceforge.net/).

## 3.6 Conclusions

Herein, this report described the development of HCV protein interaction knowledgebase (HCVpro), a relational database dedicated to HCV protein interactions. It contains manually verified hepatitis C virus-virus and virus-human host cellular protein interactions obtained from other databases and curated literature. HCVpro provides multitudes of contextualized and functional genomic data pertaining to human and HCV proteins and cross-referenced links to enriched biological databases. From the aforementioned, the reported resource is therefore not an alternative to the existing panoply of protein interaction databases but could augment currents efforts aimed at elucidating the complex molecular interplay between viral and host cellular proteins after HCV infection.

## 3.7 References

Agaugue, S., Perrin-Cocon, L., Andre, P., & Lotteau, V. (2007). Hepatitis C lipo-Viro-particle from chronically infected patients interferes with TLR4 signaling in dendritic cell. *PLoS One, 2*(3), e330.

Ahn, J., Chung, K. S., Kim, D. U., Won, M., Kim, L., Kim, K. S., et al. (2004). Systematic identification of hepatocellular proteins interacting with NS5A of the hepatitis C virus. *J Biochem Mol Biol, 37*(6), 741-748.

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., et al. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res, 33*(Database issue), D418-424.

Alibes, A., Yankilevich, P., Canada, A., & Diaz-Uriarte, R. (2007). IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC Bioinformatics, 8*, 9.

Amberger, J., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res, 37*(Database issue), D793-796.

Barsnes, H., Cote, R. G., Eidhammer, I., & Martens, L. (2010). OLS dialog: an open-source front end to the ontology lookup service. *BMC Bioinformatics, 11*, 34.

Basu, A., Meyer, K., Ray, R. B., & Ray, R. (2001). Hepatitis C virus core protein modulates the interferon-induced transacting factors of Jak/Stat signaling pathway but does not affect the activation of downstream IRF-1 or 561 gene. *Virology, 288*(2), 379-390.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2009). GenBank. *Nucleic Acids Res, 37*(Database issue), D26-31.

Berggard, T., Linse, S., & James, P. (2007). Methods for the detection and analysis of

protein-protein interactions. *Proteomics, 7*(16), 2833-2842.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., et al. (2002). The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr, 58*(Pt 6 No 1), 899-907.

Caillot, F., Hiron, M., Goria, O., Gueudin, M., Francois, A., Scotte, M., et al. (2009). Novel serum markers of fibrosis progression for the follow-up of hepatitis C virus-infected patients. *Am J Pathol, 175*(1), 46-53.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., et al. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res, 38*(Database issue), D532-539.

Chang, Q., Chen, J., Beezhold, K. J., Castranova, V., Shi, X., & Chen, F. (2009). JNK1 activation predicts the prognostic outcome of the human hepatocellular carcinoma. *Mol Cancer, 8*, 64.

Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S., Sacco, F., et al. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Res, 37*(Database issue), D669-673.

Chen, C. M., You, L. R., Hwang, L. H., & Lee, Y. H. (1997). Direct interaction of hepatitis C virus core protein with the cellular lymphotoxin-beta receptor modulates the signal pathway of the lymphotoxin-beta receptor. *J Virol, 71*(12), 9417-9426.

Chen, Y., & Han, K. (2009). BSFINDER: finding binding sites of HCV proteins using a support vector machine. *Protein Pept Lett, 16*(4), 373-382.

Chowbina, S. R., Wu, X., Zhang, F., Li, P. M., Pandey, R., Kasamsetty, H. N., et al. (2009). HPD: an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics, 10 Suppl 11*, S5.

Combet, C., Garnier, N., Charavay, C., Grando, D., Crisan, D., Lopez, J., et al. (2007). euHCVdb: the European hepatitis C virus database. *Nucleic Acids Res, 35*(Database issue), D363-366.

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., et al. (2009). Literature-curated protein interaction datasets. *Nat Methods, 6*(1), 39-46.

de Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugue, S., et al. (2008). Hepatitis C virus infection protein network. *Mol Syst Biol, 4*, 230.

Dimitrova, M., Imbert, I., Kieny, M. P., & Schuster, C. (2003). Protein-protein interactions between hepatitis C virus nonstructural proteins. *J Virol, 77*(9), 5401-5414.

Dubuisson, J. (2007). Hepatitis C virus proteins. *World J Gastroenterol, 13*(17), 2406-2415.

El Bassiouny, A. E., El-Bassiouni, N. E., Nosseir, M. M., Zoheiry, M. M., El-Ahwany, E. G., Salah, F., et al. (2008). Circulating and hepatic Fas expression in HCV-induced chronic liver disease and hepatocellular carcinoma. *Medscape J Med, 10*(6), 130.

Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., Tiollais, P., et al. (2000). A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene, 242*(1-2), 369-379.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., et al. (2008). Ensembl 2008. *Nucleic Acids Res, 36*(Database issue), D707-714.

Fu, W., Sanders-Beer, B. E., Katz, K. S., Maglott, D. R., Pruitt, K. D., & Ptak, R. G. (2009). Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res, 37*(Database issue), D417-422.

Gangadharan, B., Antrobus, R., Dwek, R. A., & Zitzmann, N. (2007). Novel serum biomarker candidates for liver fibrosis in hepatitis C patients. *Clin Chem, 53*(10), 1792-1799.

Gene Ontology Consortium (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res, 38*(Database issue), D331-335.

Greene, J. M., Collins, F., Lefkowitz, E. J., Roos, D., Scheuermann, R. H., Sobral, B., et al. (2007). National Institute of Allergy and Infectious Diseases bioinformatics resource centers: new assets for pathogen informatics. *Infect Immun, 75*(7), 3212-3219.

He, B., Qiu, X., Li, P., Wang, L., Lv, Q., & Shi, T. (2010). HCCNet: an integrated network database of hepatocellular carcinoma. *Cell Res, 20*(6), 732-734.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., et al. (2004). The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol, 22*(2), 177-183.

Hodge, A. E., Altman, R. B., & Klein, T. E. (2007). The PharmGKB: integration, aggregation, and annotation of pharmacogenomic data and knowledge. *Clin Pharmacol Ther, 81*(1), 21-24.

Hsu, C. N., Lai, J. M., Liu, C. H., Tseng, H. H., Lin, C. Y., Lin, K. T., et al. (2007). Detection of the inferred interaction network in hepatocellular carcinoma from EHCO (Encyclopedia of Hepatocellular Carcinoma genes Online). *BMC Bioinformatics, 8*, 66.

Huang, Y. P., Zhang, S. L., Cheng, J., Wang, L., Guo, J., Liu, Y., et al. (2005). Screening of genes of proteins interacting with p7 protein of hepatitis C virus from human liver cDNA library by yeast two-hybrid system. *World J Gastroenterol, 11*(30), 4709-4714.

Jeng, K. S., Sheen, I. S., Chen, B. F., & Wu, J. Y. (2000). Is the p53 gene mutation of prognostic value in hepatocellular carcinoma after resection? *Arch Surg, 135*(11), 1329-1333.

Jennings, T. A., Chen, Y., Sikora, D., Harrison, M. K., Sikora, B., Huang, L., et al. (2008). RNA unwinding activity of the hepatitis C virus NS3 helicase is modulated by the NS5B polymerase. *Biochemistry, 47*(4), 1126-1135.

Kandasamy, K., Mohan, S. S., Raju, R., Keerthikumar, S., Kumar, G. S., Venugopal, A. K., et al. (2010). NetPath: a public resource of curated signal transduction pathways. *Genome Biol, 11*(1), R3.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res, 38*(Database issue), D355-360.

Kang, S. M., Shin, M. J., Kim, J. H., & Oh, J. W. (2005). Proteomic profiling of cellular proteins interacting with the hepatitis C virus core protein. *Proteomics, 5*(8), 2227-2237.

Kasprzak, A., Adamek, A., Przybyszewska, W., Olejniczak, K., Biczysko, W., Mozer-Lisewska, I., et al. (2009). p21/Wafl/Cipl cellular expression in chronic long-lasting hepatitis C: correlation with HCV proteins (C, NS3, NS5A), other cell-

cycle related proteins and selected clinical data. *Folia Histochem Cytobiol, 47*(3), 385-394.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., et al. (2007). IntAct--open source resource for molecular interaction data. *Nucleic Acids Res, 35*(Database issue), D561-565.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Res, 37*(Database issue), D767-772.

Kovalenko, E., Tacke, F., Gressner, O. A., Zimmermann, H. W., Lahme, B., Janetzko, A., et al. (2009). Validation of connective tissue growth factor (CTGF/CCN2) and its gene polymorphisms as noninvasive biomarkers for the assessment of liver fibrosis. *J Viral Hepat, 16*(9), 612-620.

Kovanen, P. E., Rosenwald, A., Fu, J., Hurt, E. M., Lam, L. T., Giltnane, J. M., et al. (2003). Analysis of gamma c-family cytokine target genes. Identification of dual-specificity phosphatase 5 (DUSP5) as a regulator of mitogen-activated protein kinase activity in interleukin-2 signaling. *J Biol Chem, 278*(7), 5205-5213.

Krekulova, L., Rehak, V., & Riley, L. W. (2006). Structure and functions of hepatitis C virus proteins: 15 years after. *Folia Microbiol (Praha), 51*(6), 665-680.

Kuiken, C., & Simmonds, P. (2009). Nomenclature and numbering of the hepatitis C virus. *Methods Mol Biol, 510*, 33-53.

Kuiken, C., Yusim, K., Boykin, L., & Richardson, R. (2005). The Los Alamos hepatitis C sequence database. *Bioinformatics, 21*(3), 379-384.

Kwofie, S. K., Radovanovic, A., Sundararajan, V. S., Maqungo, M., Christoffels, A., & Bajic, V. B. (2011). Dragon exploratory system on hepatitis C virus (DESHCV). *Infect Genet Evol, 11*(4), 734-739.

Lai, C. K., Jeng, K. S., Machida, K., & Lai, M. M. (2008). Association of hepatitis C virus replication complexes with microtubules and actin filaments is dependent on the interaction of NS3 and NS5A. *J Virol, 82*(17), 8838-8848.

Lan, S., Wang, H., Jiang, H., Mao, H., Liu, X., Zhang, X., et al. (2003). Direct interaction between alpha-actinin and hepatitis C virus NS5B. *FEBS Lett, 554*(3), 289-294.

Lee, N. P., Chen, L., Lin, M. C., Tsang, F. H., Yeung, C., Poon, R. T., et al. (2009). Proteomic expression signature distinguishes cancerous and nonmalignant tissues in hepatocellular carcinoma. *J Proteome Res, 8*(3), 1293-1303.

Looi, K. S., Nakayasu, E. S., Diaz, R. A., Tan, E. M., Almeida, I. C., & Zhang, J. Y. (2008). Using proteomic approach to identify tumor-associated antigens as markers in hepatocellular carcinoma. *J Proteome Res, 7*(9), 4004-4012.

Luquin, E., Larrea, E., Civeira, M. P., Prieto, J., & Aldabe, R. (2007). HCV structural proteins interfere with interferon-alpha Jak/STAT signalling pathway. *Antiviral Res, 76*(2), 194-197.

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res, 35*(Database issue), D26-31.

Mas, V. R., Maluf, D. G., Archer, K. J., Yanek, K., Bornstein, K., & Fisher, R. A. (2009). Proteomic analysis of HCV cirrhosis and HCV-induced HCC: identifying biomarkers for monitoring HCV-cirrhotic patients awaiting liver transplantation. *Transplantation, 87*(1), 143-152.

Masumi, A., Aizaki, H., Suzuki, T., DuHadaway, J. B., Prendergast, G. C., Komuro, K.,

et al. (2005). Reduction of hepatitis C virus NS5A phosphorylation through its interaction with amphiphysin II. *Biochem Biophys Res Commun, 336*(2), 572-578.

Mathivanan, S., Periaswamy, B., Gandhi, T. K., Kandasamy, K., Suresh, S., Mohmood, R., et al. (2006). An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics, 7 Suppl 5*, S19.

Matsumoto, M., Hsieh, T. Y., Zhu, N., VanArsdale, T., Hwang, S. B., Jeng, K. S., et al. (1997). Hepatitis C virus core protein interacts with the cytoplasmic tail of lymphotoxin-beta receptor. *J Virol, 71*(2), 1301-1309.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., et al. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res, 37*(Database issue), D619-622.

Morshed, S. A., Fukuma, H., Kimura, Y., Watanabe, S., & Nishioka, M. (1993). Interferon-gamma, interleukin (IL)-2 and IL-2 receptor expressions in hepatitis C virus-infected liver. *Gastroenterol Jpn, 28 Suppl 5*, 59-66.

Navratil, V., de Chassey, B., Meyniel, L., Delmotte, S., Gautier, C., Andre, P., et al. (2009). VirHostNet: a knowledge base for the management and the analysis of proteome-wide virus-host interaction networks. *Nucleic Acids Res, 37*(Database issue), D661-668.

Nitahara-Kasahara, Y., Fukasawa, M., Shinkai-Ouchi, F., Sato, S., Suzuki, T., Murakami, K., et al. (2009). Cellular vimentin content regulates the protein level of hepatitis C virus core protein and the hepatitis C virus production in cultured cells. *Virology, 383*(2), 319-327.

Op De Beeck, A., Montserret, R., Duvet, S., Cocquerel, L., Cacan, R., Barberot, B., et al. (2000). The transmembrane domains of hepatitis C virus envelope glycoproteins E1 and E2 play a major role in heterodimerization. *J Biol Chem, 275*(40), 31428-31437.

Orchard, S., Salwinski, L., Kerrien, S., Montecchi-Palazzi, L., Oesterheld, M., Stumpflen, V., et al. (2007). The minimum information required for reporting a molecular interaction experiment (MIMIx). *Nat Biotechnol, 25*(8), 894-898.

Peng, S. Y., Chen, W. J., Lai, P. L., Jeng, Y. M., Sheu, J. C., & Hsu, H. C. (2004). High alpha-fetoprotein level correlates with high stage, early recurrence and poor prognosis of hepatocellular carcinoma: significance of hepatitis virus infection, age, p53 and beta-catenin mutations. *Int J Cancer, 112*(1), 44-50.

Penin, F., Dubuisson, J., Rey, F. A., Moradpour, D., & Pawlotsky, J. M. (2004). Structural biology of hepatitis C virus. *Hepatology, 39*(1), 5-19.

Phizicky, E. M., & Fields, S. (1995). Protein-protein interactions: methods for detection and analysis. *Microbiol Rev, 59*(1), 94-123.

Raman, K. (2010). Construction and analysis of protein-protein interaction networks. *Autom Exp, 2*(1), 2.

Sabile, A., Perlemuter, G., Bono, F., Kohara, K., Demaugre, F., Kohara, M., et al. (1999). Hepatitis C virus core protein binds to apolipoprotein AII and its secretion is modulated by fibrates. *Hepatology, 30*(4), 1064-1076.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., et al. (2010). GeneCards Version 3: the human gene integrator. *Database (Oxford), 2010*, baq020.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., & Eisenberg, D.

(2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res, 32*(Database issue), D449-451.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res, 37*(Database issue), D674-679.

Schoniger-Hekele, M., Hanel, S., Wrba, F., & Muller, C. (2005). Hepatocellular carcinoma--survival and clinical characteristics in relation to various histologic molecular markers in Western patients. *Liver Int, 25*(1), 62-69.

Serti, E., Doumba, P. P., Thyphronitis, G., Tsitoura, P., Katsarou, K., Foka, P., et al. (2010). Modulation of IL-2 expression after uptake of hepatitis C virus non-enveloped capsid-like particles: the role of p38 kinase. *Cell Mol Life Sci, 68*(3), 505-522.

Sharma, S. D. (2010). Hepatitis C virus: molecular biology & current therapeutic options. *Indian J Med Res, 131*, 17-34.

Shin, I. T., Tanaka, Y., Tateno, Y., & Mizokami, M. (2008). Development and public release of a comprehensive hepatitis virus database. *Hepatol Res, 38*(3), 234-243.

Sieghart, W., Losert, D., Strommer, S., Cejka, D., Schmid, K., Rasoul-Rockenschaub, S., et al. (2006). Mcl-1 overexpression in hepatocellular carcinoma: a potential target for antisense therapy. *J Hepatol, 44*(1), 151-157.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., & Ideker, T. (2010). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics, 27*(3), 431-432.

Su, W. H., Chao, C. C., Yeh, S. H., Chen, D. S., Chen, P. J., & Jou, Y. S. (2007). OncoDB.HCC: an integrated oncogenomic database of hepatocellular carcinoma revealed aberrant cancer target genes and loci. *Nucleic Acids Res, 35*(Database issue), D727-731.

Tan, S. L., Nakao, H., He, Y., Vijaysri, S., Neddermann, P., Jacobs, B. L., et al. (1999). NS5A, a nonstructural protein of hepatitis C virus, binds growth factor receptor-bound protein 2 adaptor protein in a Src homology 3 domain/ligand-dependent manner and perturbs mitogenic signaling. *Proc Natl Acad Sci U S A, 96*(10), 5533-5538.

Tong, W. Y., Nagano-Fujii, M., Hidajat, R., Deng, L., Takigawa, Y., & Hotta, H. (2002). Physical interaction between hepatitis C virus NS4B protein and CREB-RP/ATF6beta. *Biochem Biophys Res Commun, 299*(3), 366-372.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., & Wodak, S. J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. *Database (Oxford), 2010*, baq026.

UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res, 38*(Database issue), D142-148.

Wang, C., Gale, M., Jr., Keller, B. C., Huang, H., Brown, M. S., Goldstein, J. L., et al. (2005). Identification of FBL2 as a geranylgeranylated cellular protein required for hepatitis C virus RNA replication. *Mol Cell, 18*(4), 425-434.

Wang, Q. M., Hockman, M. A., Staschke, K., Johnson, R. B., Case, K. A., Lu, J., et al. (2002). Oligomerization and cooperative RNA synthesis activity of hepatitis C virus RNA-dependent RNA polymerase. *J Virol, 76*(8), 3865-3872.

Yokoyama, Y., Kuramitsu, Y., Takashima, M., Iizuka, N., Terai, S., Oka, M., et al.

(2006). Protein level of apolipoprotein E increased in human hepatocellular carcinoma. *Int J Oncol, 28*(3), 625-631.

Yoon, S. Y., Kim, J. M., Oh, J. H., Jeon, Y. J., Lee, D. S., Kim, J. H., et al. (2006). Gene expression profiling of human HBV- and/or HCV-associated hepatocellular carcinoma cells using expressed sequence tags. *Int J Oncol, 29*(2), 315-327.
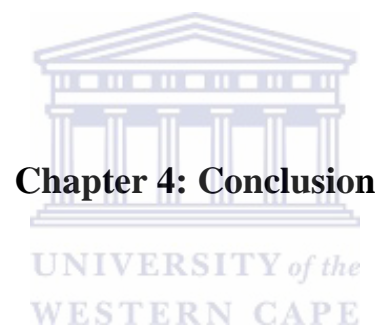
**Chapter 4: Conclusion**

## 4.1 Research aims revisited

This section presents solutions to the research aims posed in this thesis.

### 4.1.1 Aim 1

Development of Hepatitis C Virus (HCV)-focused web-based text mining resource.

### Solution

Chapter 2 of this thesis describes the development of Hepatitis C Virus (HCV) customized online text mining resource, the Dragon Exploratory System on Hepatitis C Virus (DESHCV) (Kwofie et al., 2011). DESHCV is a dictionary lookup system that employs named concept recognition to locate concepts present in a list of HCV-related PubMed abstracts. Two concepts are suggested to be associated when both concepts co-occur in a particular PubMed abstract. As part of the DESHCV implementation, the existing data files in the dragon exploratory system were enriched with HCV-related biomedical concepts, including their name variants and symbols. The concepts consist of the names of biomedical entities that were assigned to variety of dictionaries, including "human genes and proteins", "metabolites and enzymes", "pathways", "chemicals with pharmacological effects", "Hepatitis C Virus concepts", and "disease concepts". DESHCV is therefore a literature-based discovery resource and enables users to query the system using specified keywords, phrases and concept names to retrieve text-derived association networks, hypothesis and a list of PubMed abstracts containing colour-coded tagged concepts.

## 4.1.2 Research aim 2

Development of HCV online protein knowledgebase.

## Solution

Chapter 2 of this thesis described the development of an online knowledgebase, known as the Hepatitis C Virus Protein Interaction Database (HCVpro). HCVpro is a relational database that houses manually verified HCV-HCV and HCV-human protein interactions obtained from external databases and curated from published biomedical literature as well as their associated metadata. A user query via HCVpro retrieves comprehensive information on protein-protein interactions (PPIs), as well as rich data on functional annotations, hepatocellular carcinoma (HCC) related genes, drug development, canonical pathways, and cross-referenced links to other essential biological databases.
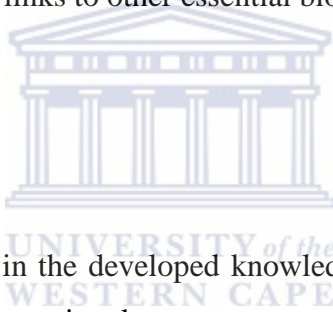
## 4.1.3 Research aim 3

Harnessing the myriad of data in the developed knowledgebases to facilitate generation of functional hypothesis of therapeutic relevance.
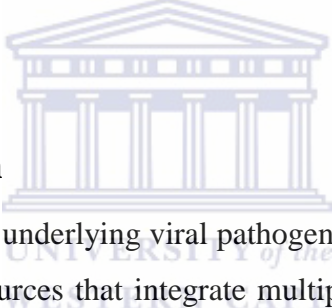
## Solution

DESHCV is a text mining system that computes possible association between concepts and additionally generate hypothesis that can lead to potentially novel discovery. DESHCV has been used to reproduce the already confirmed thalidomide-chronic hepatitis C hypothesis (Weeber et al., 2003). This system has also been employed to suggest a novel hypothesis composed of thalidomide-amantadine association. The generated hypotheses were influenced by Swanson's ABC model and the open discovery approach (Swanson, 1991; Weeber et al., 2001).

HCVpro is an integrated database composed of myriads of data that can be harnessed to generate potentially testable baseline hypothesis or obtain clues that may pinpoint to a

novel research paradigm to aid in circumventing HCV therapeutic challenges. For instance, by combining information on canonical pathways, interactions with key oncogenic HCV proteins and gene expression levels in HCC samples, a clue was obtained which corroborate the need to further support the ongoing exploration of vimentin as a potential therapeutic target or diagnostic biomarker. Additionally, the diagnostic potential of a list of 23 HCC associated genes obtained from database sources were investigated using literature exploration. Some of the 23 investigated genes were related to cancer associated Gene Ontology terms and pathways. Apparently, no literature was found to have provided results on the diagnostic potential of the following eight genes: *ACLY*, *AZGP1*, *DDX3X*, *FGG*, *H19*, *SIAH1*, *SERPING1* and *THBS1*. Using this simple interpretive approach, it is therefore tempting to suggest that the above-mentioned eight genes can be investigated to unearth any diagnostic or therapeutic role.

## 4.2 Research contribution

Understanding the mechanisms underlying viral pathogenesis requires robust approaches, including the provision of resources that integrate multiple biologically relevant data to augment HCV research. The text mining system and HCV protein knowledgebase developed in pursuance of the research aims reported here provide potential platforms that can aid researchers with harvested data with which to generate biologically relevant hypotheses or formulate plausible research questions worthy of investigation. A significant amount of biological data originating from both small- and large-scale experiments are reported in published biomedical text. Therefore, harvesting these data in a structured format for usage by researchers is critical. The primary data source of DESHCV is published biomedical text. HCVpro harvests data from both literature and external databases. The availability of both DESHCV and HCVpro is very timely because these resources provide enriched consolidated HCV-specific information.

Currently, patients infected with certain HCV genotypes do not derive maximum therapeutic benefits from the standard regimen composed of interferon alpha and

ribavirin. A number of pharmacological targets are being investigated for HCV therapy, including cellular receptors mediating HCV entry, factors facilitating HCV replication and assembly, and intracellular pathways (Sato et al., 2008). The essential benefits of computational techniques to support investigation of HCV proteins as targets for drug development have been highlighted (Lahm et al., 2002). DESHCV utilizes biomedical entities such as HCV proteins, cellular receptors and pathways as query concepts to generate functional hypothesis. By adding user intuition, the search for therapeutic targets with potential to improve sustained virologic response rates, and reduce undesirable side effects can be enhanced.

Patients suffering from chronic hepatitis C infection may eventually develop HCV-induced HCC. HCVpro is populated with protein interaction data consolidated with information on HCC genes and this can be harnessed to leverage the current search for plausible non-invasive diagnostic markers as well as therapeutic targets. Computed topological features obtained from protein networks generated using the interaction datasets provided by HCVpro can be combined with the integrated genomic data to prioritize candidate HCC related genes.
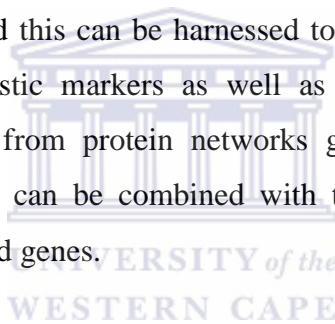
## 4.3 Limitations and future work

## 4.3.1 Limitations and future work on DESHCV

The heuristic filtering modules of DESHCV will be augmented with simple handcrafted rules to enhance the ability of the system to recognize named concepts with poorly defined morphological features. Associations suggested by DESHCV must be confirmed by manual verification of accompanying literature and also via laboratory experimentation when necessary. The updated version of DESHCV will utilize pattern-matching based approaches to predict the types of association that exist between entities within suggested hypothesis or relationships. Full text articles were not used as source of data but rather abstracts. Unfortunately, abstracts provide only a bird's eye view of the summarized results. Currently, the possibility of combining both abstracts and full text journal articles as sources of text corpora in DESHCV is being explored. Additionally,

robust systems will be designed to circumvent challenges that arise as a result of tokenization and indexing of full text documents.

## 4.3.2 Limitations and future work on HCVpro

HCVpro does not provide interaction reliability scores nor support online exploitation of the incorporated data. With possible increase in interaction metadata in the near future, protein interactions will be assigned with reliability scores. Subsequent versions of HCVpro will accommodate the online visualization and manipulation of protein interactions via the use of embedded Java web applets. HCVpro will be integrated with separate modules that allow online screening of drugs and drug targets, as well as computing of the druggability of incorporated interactions.

## 4.4 References

Kwofie, S. K., Radovanovic, A., Sundararajan, V. S., Maqungo, M., Christoffels, A., & Bajic, V. B. (2011). Dragon exploratory system on hepatitis C virus (DESHCV). *Infect Genet Evol, 11*(4), 734-739.

Lahm, A., Yagnik, A., Tramontano, A., & Koch, U. (2002). Hepatitis C virus proteins as targets for drug development: the role of bioinformatics and modelling. *Curr Drug Targets, 3*(4), 281-296.

Sato, K., Takagi, H., Ichikawa, T., Kakizaki, S., & Mori, M. (2008). Emerging therapeutic strategies for hepatitis C virus infection. *Curr Mol Pharmacol, 1*(2), 130-150.

Swanson, D. R. (1991). *Complementary structures in disjoint science literatures.* Paper presented at the Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Weeber, M., Klein, H., De Jong-van den Berg, L. T. W. and Vos, R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *J. Amer. Soc. Inf. Sci. Tech, 52*(7), 254–262.

Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L. T., Aronson, A. R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *J Am Med Inform Assoc, 10*(3), 252-259.
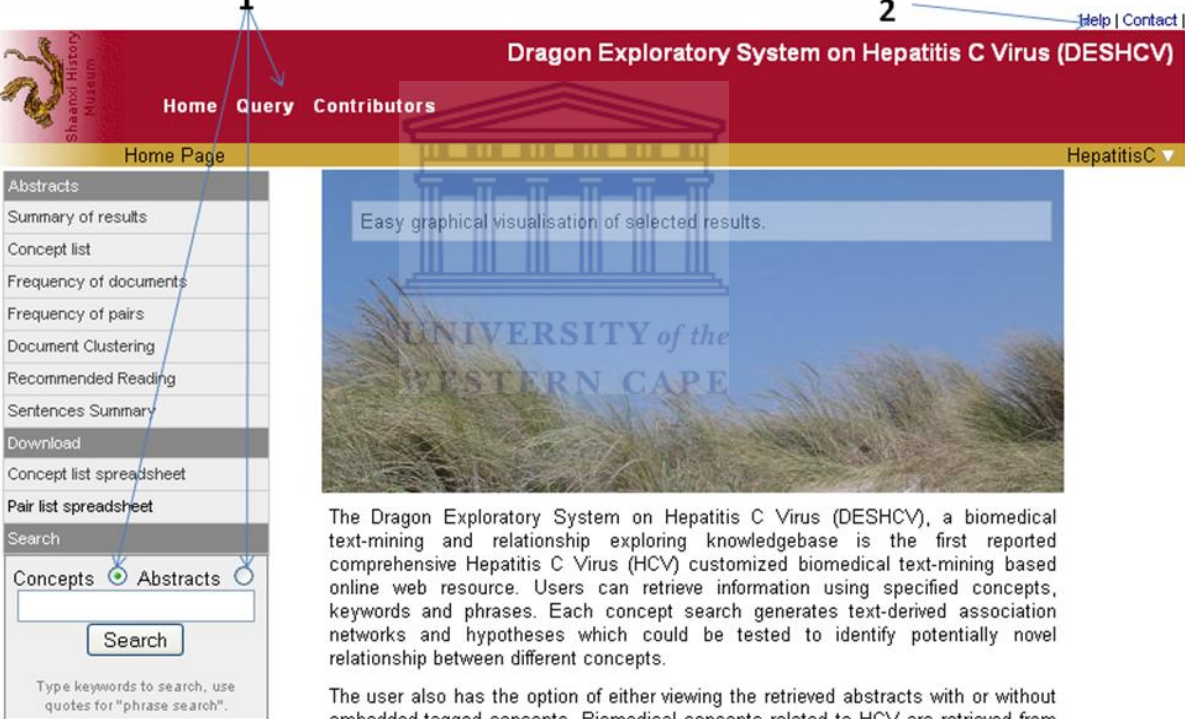
**APPENDIX**

# User Manual for Dragon Exploratory System on Hepatitis C Virus (DESHCV)

## Section A

A typical user query interface showing the various utility components. DESHCV is free for academic and non-profit users via http://apps.sanbi.ac.za/DESHCV/ and its mirror site http://cbrc.kaust.edu.sa/deshcv/



**Figure 1.** DESHCV user Interface

**1.** For Abstract and Concept queries

[For further details consult sections B, C and D of this manual]

**2.** Help

Click on help menu for detailed explanations of concepts and help manual on DESHCV usage.

**Section B: Abstract query**



**Figure 2.** Displays the simplified abstract query interface

**A typical Analysis Flow**

Type any keyword for example: hypervariability -> check abstract radio button -> click on search -> a display of PubMed abstracts with tagged colour-coded concepts

**Section C: Concept query**



**Figure 3.** Displays the simplified concept query interface

**A typical Analysis Flow**

Type any concept name for example: "F protein" -> check concept radio button -> click on search -> a graphical display of colour-coded concept

**Concept Search Output and Analysis Flow**



**Figure 4.** Textual display of concept search output

**1.** Concept query term "F protein"

**2.** Type: F protein -> click on either display as dictionary or table -> click on concepts search button ->concept lists output

Note: do not re-type F protein if it has been queried already (this is an illustration)

**3.** Lists of the various color-coded concepts assigned to their respective dictionaries. Users can retrieve cross-referenced annotations by clicking on any of the concepts.

**Green:** Human proteins and genes

**Light pink:** Metabolites and enzymes

**Blue:** Pathways

**Yellow:** Chemicals with Pharmacological effects

**Purple:** Diseases concepts

**4.** Visualization

**Network generator:** click on draw network -> generates a network of interconnected concepts

**Hypothesis Generator:** click on show hypothesis -> generates correlated concepts in a graphical format as hypothesis

**Download:** click on table download -> retrieves a list of concepts in a CSV format

**Interactive Association Map**
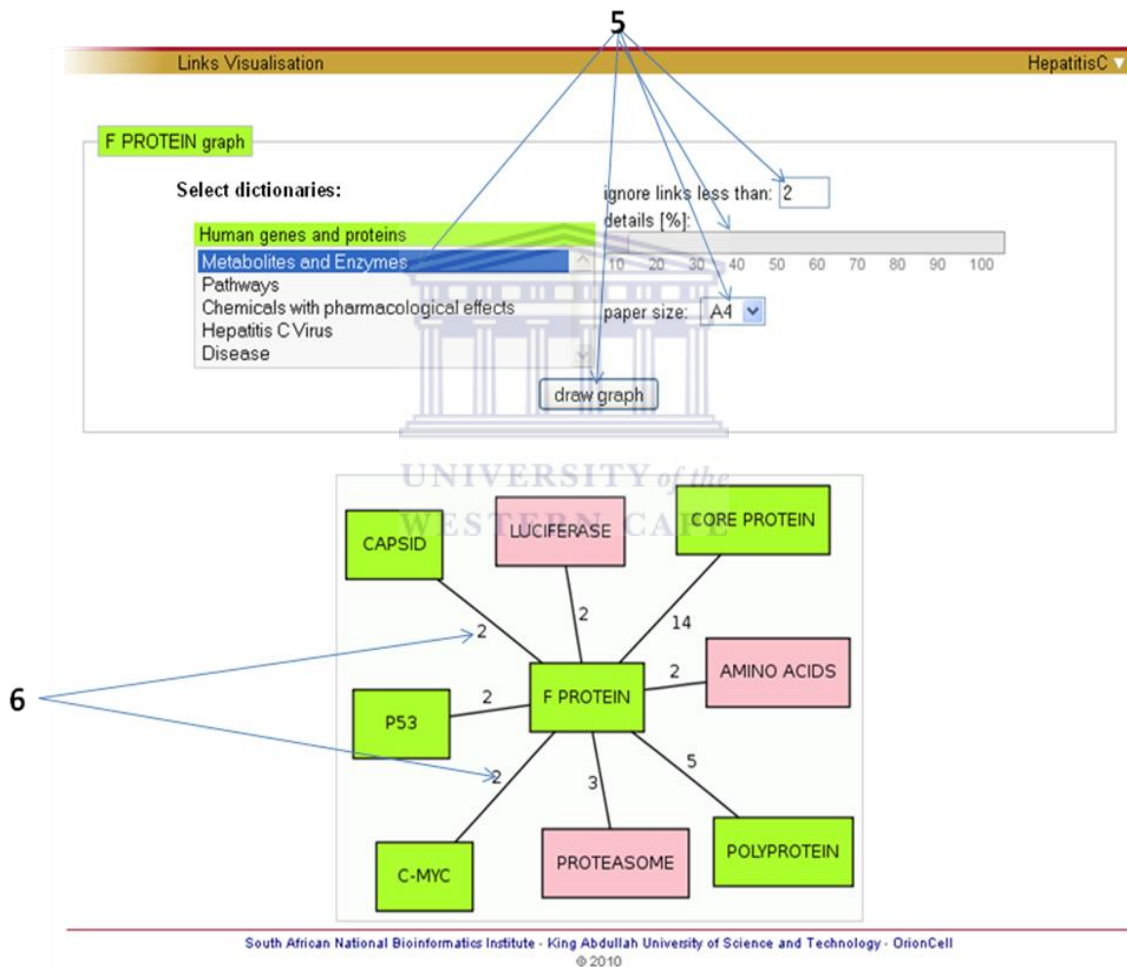


**Figure 5.** Graphical display of interacting concepts

## 5. Association Map Analysis flow

Click on any of the specified dictionaries (For example: metabolites and enzymes) -> click on draw graph -> redraws the graph with the newly defined concepts.

### Buttons

**Ignore links less than:** enable the user to display the network according to the number of links.

**Paper size:** enables the user to resize the map by choosing from a range of A0 to A5.

**Details [%]:** enable the user to enhance the visualization clarity by varying the detail slider.

**6.** Indicates the number of occurrences of the selected concept in PubMed abstracts. Clicking on the link retrieves the related PubMed abstracts.

**Note:** The network consist of proposed correlations and the user is required to manually verify the abstract to either reject or accept the suggested association
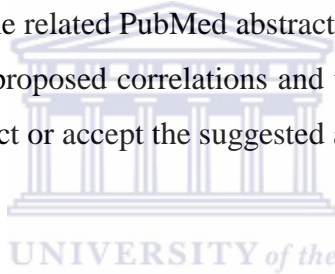
### Hypothesis Generator

Generate implicit relationships between disjunct literatures associated with concepts. This proposed hypothesis must be manually verified by consulting the linked PubMed abstracts
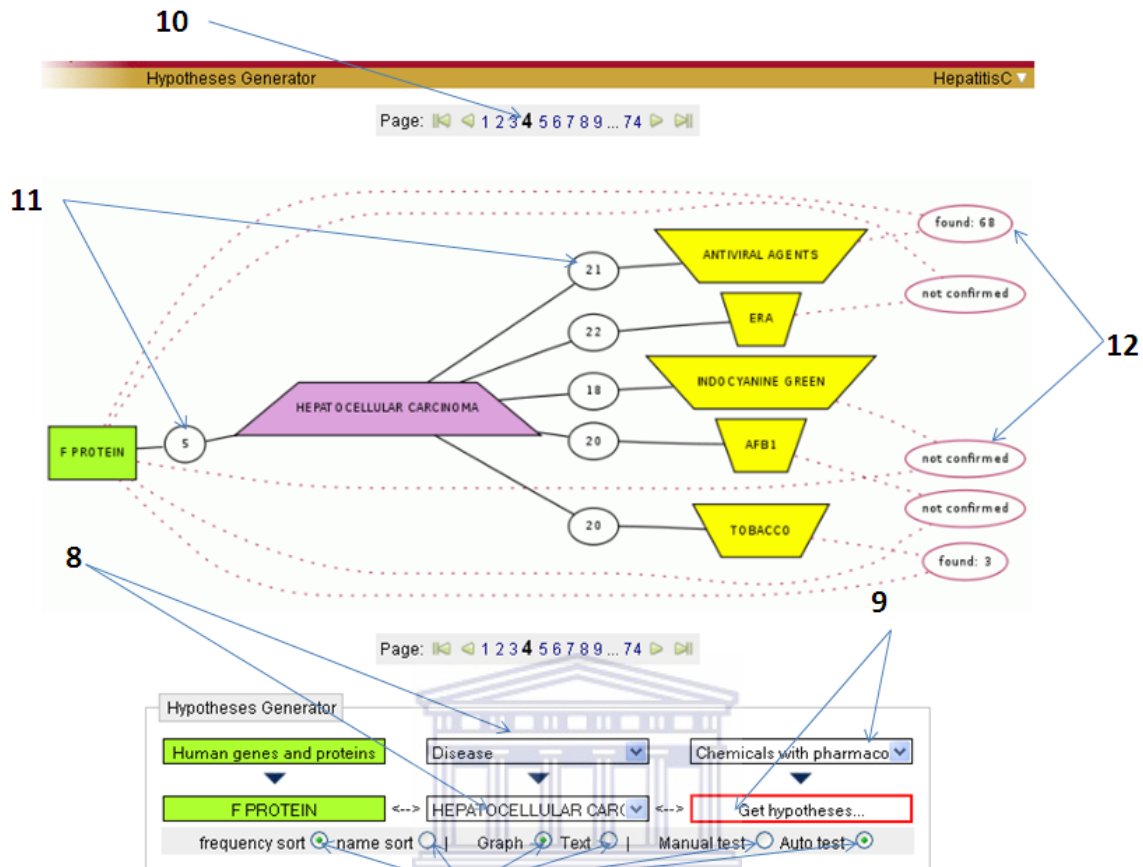
**Figure 6.** Graphical display of generated hypothesis

**7.** Buttons:

**Frequency Sort:** enable users to sort the generated hypothesis according to the frequency of occurrence of the concepts.

**Name Sort:** enable users to retrieve the generated hypothesis in an alphabetical order.

**Graph:** enables users to visualize the generated hypothesis in an interactive map

**Text:** displays the hypothesis in a textual format

**Manual Test:** by checking on this button, users are required to manually verify the abstract to ascertain the veracity of the proposed hypothesis.

**Auto test:** automatically verifies the abstracts to ascertain whether the proposed correlation suggested via the hypothesis has been reported in PubMed abstracts.

128

**8.** This drop down menu enables users to select dictionaries and associated concepts with which to generate the hypothesis. This concept is known as the linking term since it links the starting and the targets terms respectively.

According to Figure 6:

**Starting term:** F protein belongs to the human proteins and genes dictionary

**Linking term:** Hepatocellular Carcinoma belongs to the disease dictionary

**Target term:** Any Concept in the pharmacological chemicals dictionary

**9.** Click on Get hypothesis -> generates hypothesis according to specified dictionary concepts

**10.** Allow browsing of hypothesis by page

**11.** Displays the number of PubMed abstracts reporting the association

**12.** Indicate whether the displayed correlation has been reported

**Not confirmed:** Hypothesis has not been reported which could mean a novel discovery

**Found: 68**, means 68 abstracts have reported the hypothesis

**Section D: Simulated concept query examples**

**1.** Concept query: "Thalidomide"



| Human genes and proteins | Metabolites and Enzymes | Chemicals with pharmacological effects | Hepatitis C Virus | Disease |
|---|---|---|---|---|
| CHC [1] | ADEFOVIR [1] | EMTRICITABINE [1] | HCV [4] | BACK PAIN [1] |
| GAMMA-GLUTAMYLTRANSFERASE [1] | ADRIAMYCIN [1] | ENTECAVIR [1] | HEPATITIS C VIRUS [4] | BEHÇET'S SYNDROME [1] |
| IFN [1] | AZATHIOPRINE [1] | MRNA [1] | | CANCER [1] |
| IFN-ALPHA [1] | CYCLOPHOSPHAMIDE [1] | MYCOPHENOLIC ACID [1] | | CHRONIC HEPATITIS [1] |
| INTERFERON GAMMA [1] | CYCLOSPORIN A [1] | NUCLEOSIDE [1] | | CHURG-STRAUSS SYNDROME [1] |
| PERFORIN [1] | DEXAMETHASONE [1] | PHARMACEUTICAL [1] | | COMPRESSION FRACTURE [1] |
| TUMOR NECROSIS FACTOR [1] | GCA [1] | POLYPHENOL [1] | | DEGENERATIVE CHANGE [1] |
| TUMOR NECROSIS FACTOR ALPHA [1] | HISTAMINE [1] | RNA [1] | | ESSENTIAL MIXED CRYOGLOBULINAEMIA [1] |
| | INTERFERON ALPHA [1] | VENA [1] | | ESSENTIAL MIXED CRYOGLOBULINEMIA [1] |
| | INTERFERON-ALPHA [1] | | | FATIGUE [1] |
| | LEFLUNOMIDE [1] | | | GCS [1] |
| | METHOTREXATE [1] | | | GIANT CELL ARTERITIS [1] |
| | MYCOPHENOLATE MOFETIL [1] | | | HBV [1] |
| | VINCRISTINE [1] | | | HEPATITIS [1] |
| | INTERFERON [3] | | | HEPATITIS B [1] |
| | RIBAVIRIN [4] | | | HEPATITIS C [1] |
| | | | | LESION [1] |
| | | | | LYMPHOPROLIFERATIVE DISORDERS [1] |
| | | | | METASTASIS [1] |
| | | | | MULTIPLE MYELOMA [1] |
| | | | | NON-HODGKIN LYMPHOMA [1] |
| | | | | PLASMA CELL INFILTRATION [1] |
| | | | | REMISSION [1] |
| | | | | SPINALIS [1] |
| | | | | SYSTEMIC VASCULITIS [1] |
| | | | | TOXICITY [1] |
| | | | | TUMOR [1] |
| | | | | TUMOR ANGIOGENESIS [1] |
| | | | | TUMOR PROGRESSION [1] |
| | | | | ULCER [1] |
| | | | | UVEITIS [1] |
| | | | | VASCULITIDES [1] |
| | | | | VIRAL HEPATITIS [1] |
| | | | | HCC [2] |
| | | | | HEPATOCELLULAR CARCINOMA [2] |
| | | | | CHRONIC HEPATITIS C [3] |
| | | | | INFECTION [5] |
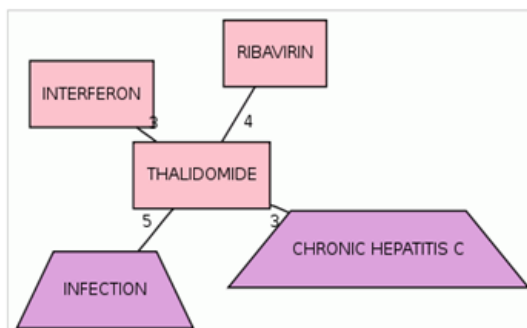
**Figure 7.** A screenshot montage of thalidomide concept query. This displays both the tabular and graphical output of query. Each concept is assigned a color according to its dictionary. Concepts belonging to the same dictionary are grouped together in the tabular format. The association map displays a network of interaction nodes consisting of concepts.

**2.** Thalidomide-Chronic Hepatitis C Association

Generate concept query for "thalidomide" -> Click on Get hypothesis -> generates hypothesis -> Specify the dictionaries as indicated in figure 8 and re-generate hypothesis. -> Click on the linked abstract to verify the hypothesis. -> Explore results for potential novel insights.



**Figure 8.** A screenshot montage showing thalidomide-chronic hepatitis C hypothesis. This displays the implicit relationship existing between the two concepts. The red arrow linking the hypothesis and PubMed abstract displays the verification of the hypothesis. The yellow dash lines and the white question mark sign displays the implicit relationship existing between the two concepts.

**3.** Thalidomide-Amantadine Association

Generate concept query for "thalidomide" -> Click on Get hypothesis -> generates hypothesis -> Specify the dictionaries as indicated in figure 9 and re-generate hypothesis. -> Explore results for potential novel insights.
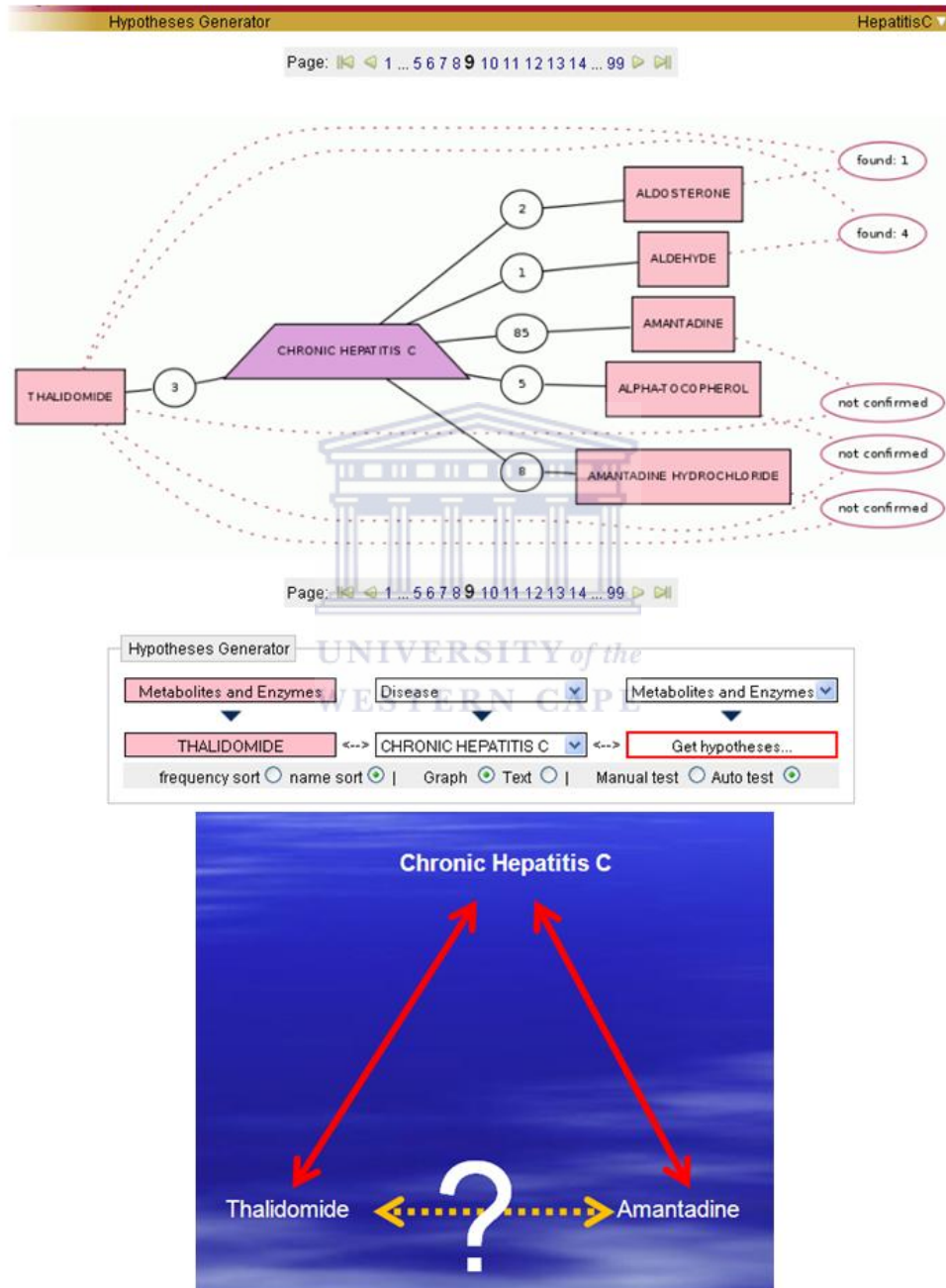


**Figure 9.** A screenshot montage displaying thalidomide-amantadine hypothesis. The yellow dash lines and the white question mark sign displays the implicit relationship existing between the two concepts.

# Appendix II (Chapter 2)

**Frequently asked questions (FAQ)**

**1. What is DESHCV?**

Dragon Exploratory System on Hepatitis C Virus (DESHCV) is the first reported comprehensive Hepatitis C Virus (HCV) customized biomedical text-mining based online web resource. A list of abstracts retrieved via PubMed database using specific keywords searches related to HCV were processed based on concept recognition of concepts from several dictionaries. It is a tool with the potential to assist in discovering associations between biomedical concepts and could lead to new research focus and possible novel discoveries.

**2. What are the benefits of using DESHCV?**

The web query interface enables retrieval of information using specified concepts, keywords and phrases, generating text-derived association networks and hypotheses which could be tested to identify potentially novel relationship between different concepts.

**3. Which dictionaries can be retrieved from DESHCV?**

Biomedical concepts related to HCV are retrieved from the following dictionaries: "Human genes and proteins", "Metabolites and Enzymes", "Pathways", "Chemicals with pharmacological effects", and "Disease concepts".

**4. What makes DESHCV unique?**

DESHCV contains pre-compiled dictionaries enriched with biomedical concepts pertaining to HCV proteins, their name variants and symbols to make it suitable for targeted information exploration and knowledge extraction as focused on HCV.

**5. Can I download the query results retrieved from the DESHCV?**

Yes, you can download the concept list and pair list spreadsheet. And also the user has the option of downloading the retrieved table of concepts for further analysis. The user may right-click on the images or figures generated to download.

**6. Are there any limitations in the usage of DESHCV?**

Yes, associations generated between paired concepts are inferred from co-occurrences and may not necessarily relate to any molecular functionality. Users are required to manually confirm the text accompanying the hypothesis to either accept or reject the suggested hypothesis.

**7. Are there any future updates or modification in the pipeline?**

Yes, we would like to integrate Blast and identifier queries to enhance querying capabilities of DESHCV. The possibility of integrating full text document is currently being explored and could be added to the database as a separate feature. The database would be updated every six months to meet the demands of ever increasing PubMed records related to HCV

**8. Is there a help manual on DESHCV?**

Yes, a help manual has been developed to ease the exploration of this database and can be accessed by clicking the help menu on index page (http://apps.sanbi.ac.za/DESHCV/) or via http://apps.sanbi.ac.za/DESHCV/help.pdf/.

**9. To whom can I report a bug or discrepancy?**

Please refer all queries via:

http://apps.sanbi.ac.za/DESHCV/Contact.php

# Appendix III (Chapter 3)

Annotated protein-protein interactions assigned with HCVpro IDs and experimental evidence mapped onto PSI-MI term IDs.

| HCVpro ID | Molecule A | Molecule B Gene Symbols | Molecule B Gene ID | PMID | Experimental Evidence (PSI-MI Term ID) |
|---|---|---|---|---|---|
| hcv0001 | CORE | LTBR | 4055 | 8995654 | Two Hybrid Test (MI:0018); Far-Western blot (MI:0047); Affinity Chromatography (MI:0004) |
| hcv0002 | CORE | APOA2 | 336 | 9037030 | Confocal microscopy (MI:0663) |
| hcv0003 | CORE | TP53 | 7157 | 9110985 | Mutational analysis (MI:0074) |
| hcv0004 | CORE | LTBR | 4055 | 9371602 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0005 | CORE | HNRNPK | 3190 | 9651361 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059); Far-Western blot (MI:0047); Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0006 | CORE | CORE | 951475 | 9851697 | Two Hybrid Test (MI:0018); Mass spectrometry (MI:0943); Gel filtration (MI:0071) |
| hcv0007 | CORE | DDX3X | 1654 | 10074132 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |
| hcv0008 | CORE | DDX3X | 1654 | 10329544 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059); Affinity Chromatography (MI:0004) |
| hcv0009 | CORE | DDX3X | 1654 | 10336476 | Two Hybrid Test (MI:0018); Affinity Chromatography (MI:0004); Immunostaining (MI:0022) |
| hcv0010 | CORE | APOA2 | 336 | 10498661 | Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); surface plasmon resonance (MI:0107); Colocalization (MI:0403) |
| hcv0011 | CORE | TP53 | 7157 | 10544138 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059); Far-Western blot (MI:0047); Mutational analysis (MI:0074); Affinity Chromatography (MI:0004); Colocalization (MI:0403) |
| hcv0012 | CORE | YWHAB | 7529 | 10644344 | Two Hybrid Test (MI:0018); Colocalization (MI:0403); Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0013 | CORE | YWHAZ | 7534 | 10644344 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Western blot (MI:0113); Affinity Chromatography (MI:0004) |
| hcv0014 | CORE | CREB3 | 10488 | 10675342 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |
| hcv0015 | CORE | CORE | 951475 | 10721731 | Two Hybrid Test (MI:0018) |
| hcv0016 | CORE | CDKN1A | 1026 | 10873631 | GST pull-down (MI:0059); Mutational analysis (MI:0074); Affinity Chromatography (MI:0004) |
| hcv0017 | CORE | TAF11 | 6882 | 10924497 | GST pull-down (MI:0059) |
| hcv0018 | CORE | TP53 | 7157 | 10924497 | Western blot (MI:0113); GST pull-down (MI:0059) |
| hcv0019 | CORE | C1QBP | 708 | 11086025 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059) |
| hcv0020 | CORE | TNFRSF1A | 7132 | 11226577 | GST pull-down (MI:0059); Kinase assay (MI:0424) (MI:0424); Affinity Chromatography (MI:0004) |
| hcv0021 | CORE | FADD | 8772 | 11336543 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |

| hcv0022 | CORE | TNF | 7124 | 11374864 | Western blot (MI:0113); GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019) |
|---------|------|-----|------|----------|-----------------------------------------------------------------------------------|
| hcv0023 | CORE | TRADD | 8717 | 11374864 | Western blot (MI:0113); GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019) |
| hcv0024 | CORE | TRAF2 | 7186 | 11374864 | Western blot (MI:0113); GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019) |
| hcv0025 | CORE | RXRA | 6256 | 11915042 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019) |
| hcv0026 | CORE | STAT3 | 6774 | 12208879 | Coimmunoprecipitation (MI:0019); Western blot (MI:0113); GST pull-down (MI:0059) |
| hcv0027 | CORE | TSN | 7247 | 12532453 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019) |
| hcv0028 | CORE | FUNDC2 | 65991 | 12665903 | Two Hybrid Test (MI:0018) |
| hcv0029 | CORE | TP73 | 7161 | 12730672 | Coimmunoprecipitation (MI:0019); Mutational analysis (MI:0074); Immunostaining (MI:0022) |
| hcv0030 | CORE | TP53 | 7157 | 12730672 | Coimmunoprecipitation (MI:0019) |
| hcv0031 | CORE | JAK1 | 3716 | 12764155 | Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0032 | CORE | JAK2 | 3717 | 12764155 | Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0033 | CORE | FAS | 355 | 12919737 | Immunoblotting (MI:0113); Coimmunoprecipitation (MI:0019) |
| hcv0034 | CORE | CCNH | 902 | 14711830 | Immunofluorescence (MI:0022); Confocal microscopy (MI:0663); GST pull-down (MI:0059); Immunoblotting (MI:0113); Coimmunoprecipitation (MI:0019); Western blot (MI:0113); Immunostaining (MI:0022) |
| hcv0035 | CORE | TBP | 6908 | 14730212 | Affinity Chromatography (MI:0004) |
| hcv0036 | CORE | TP53BP2 | 7159 | 14985081 | Two Hybrid Test (MI:0018); Affinity Chromatography (MI:0004); Coimmunoprecipitation (MI:0019); Immunostaining (MI:0022) |
| hcv0037 | CORE | SMAD3 | 4088 | 15334054 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059); Affinity Chromatography (MI:0004) |
| hcv0038 | CORE | EP300 | 2033 | 15380363 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059); Western blot (MI:0113); Affinity Chromatography (MI:0004) |
| hcv0039 | CORE | CREBBP | 1387 | 15380363 | GST pull-down (MI:0059); Western blot (MI:0113); Affinity Chromatography (MI:0004) |
| hcv0040 | CORE | TLR2 | 7097 | 15521019 | Enzyme-linked immunosorbent assay (MI:0411); Electrophoretic mobility shift assay (MI:0413); Western blot (MI:0113); Kinase assay (MI:0424) (MI:0424); Confocal microscopy (MI:0663); Flow cytometry (MI:0054); Colocalization (MI:0403) |
| hcv0041 | CORE | HLA-A | 3105 | 15681828 | Fluorescence-activated cell sorting (MI:0054); Immunostaining (MI:0022) |
| hcv0042 | CORE | HLA-E | 3133 | 15681828 | Fluorescence-activated cell sorting (MI:0054); Immunostaining (MI:0022) |
| hcv0043 | CORE | STAT1 | 6772 | 15825084 | Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0044 | CORE | ACP1 | 52 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0045 | CORE | CFL1 | 1072 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0046 | CORE | FKBP7 | 51661 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0047 | CORE | HSPD1 | 3329 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0048 | CORE | KRT18 | 3875 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |

| | | | | | |
|---|---|---|---|---|---|
| hcv0049 | CORE | KRT19 | 3880 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0050 | CORE | KRT8 | 3856 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0051 | CORE | SLC22A7 | 10864 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0052 | CORE | TATDN1 | 83940 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0053 | CORE | GLRX3 | 10539 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0054 | CORE | VIM | 7431 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943); Immunoblotting (MI:0113); Immunofluorescence (MI:0022) |
| hcv0055 | CORE | NPM1 | 4869 | 16170350 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |
| hcv0056 | CORE | YY1 | 7528 | 16170350 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |
| hcv0057 | CORE | PML | 5371 | 16322229 | Western blot (MI:0113); Coimmunoprecipitation (MI:0019); Colocalization (MI:0403) |
| hcv0058 | CORE | DICER1 | 23405 | 16530526 | Confocal microscopy (MI:0663); Immunofluorescence (MI:0022); Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0059 | CORE | PSME3 | 10197 | 16611896 | Western blot (MI:0113) |
| hcv0060 | CORE | DICER1 | 23405 | 18325616 | Western blot (MI:0113); Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022); Immunostaining (MI:0022); Mutational analysis (MI:0074) |
| hcv0061 | CORE | AGRN | 375790 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0062 | CORE | BCAR1 | 9564 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0063 | CORE | CD68 | 968 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0064 | CORE | COL4A2 | 1284 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0065 | CORE | DDX3Y | 8653 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0066 | CORE | EGFL7 | 51162 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0067 | CORE | FBLN2 | 2199 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0068 | CORE | FBLN5 | 10516 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0069 | CORE | GAPDH | 2597 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0070 | CORE | GRN | 2896 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0071 | CORE | HIVEP2 | 3097 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0072 | CORE | HOXD8 | 3234 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0073 | CORE | LPXN | 9404 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0074 | CORE | LRRTM1 | 347730 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0075 | CORE | LTBP4 | 8425 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0076 | CORE | MAGED1 | 9500 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0077 | CORE | MEGF6 | 1953 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0078 | CORE | MMRN2 | 79812 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0079 | CORE | NR4A1 | 3164 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0080 | CORE | PABPN1 | 8106 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0081 | CORE | PAK4 | 10298 | 18985028 | Two Hybrid Test (MI:0018) |

| hcv0082 | CORE | PLSCR1 | 5359 | 18985028 | Two Hybrid Test (MI:0018) |
|---------|------|--------|------|----------|---------------------------|
| hcv0083 | CORE | RNF31 | 55072 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0084 | CORE | SETD2 | 29072 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0085 | CORE | SLC31A2 | 1318 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0086 | CORE | VWF | 7450 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0087 | CORE | ZNF271 | 10778 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0088 | CORE | YWHAE | 7531 | 10644344 | Affinity Chromatography (MI:0004); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018) |
| hcv0089 | CORE | NS5A | 951475 | 11883187 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Colocalization (MI:0403); Affinity Chromatography (MI:0004) |
| hcv0090 | CORE | NS5B | 951475 | 11929715 | Immunofluorescence (MI:0022); Coimmunoprecipitation (MI:0019) |
| hcv0091 | CORE | RSF1 | 51773 | 12401801 | GST pull-down (MI:0059) |
| hcv0092 | CORE | E1 | 951475 | 12466485 | GST pull-down (MI:0059); Affinity Chromatography (MI:0004); Coimmunoprecipitation (MI:0019) |
| hcv0093 | CORE | SP110 | 3431 | 14559998 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |
| hcv0094 | CORE | KPNA1 | 3836 | 15613354 | GST pull-down (MI:0059) |
| hcv0095 | CORE | DDX5 | 1655 | 15846844 | 2-DE (MI:0982); Mass spectrometry (MI:0943) |
| hcv0096 | CORE | NS5A | 951475 | 16166788 | Two Hybrid Test (MI:0018) |
| hcv0097 | CORE | EIF2AK2 | 5610 | 17267064 | Two Hybrid Test (MI:0018); Colocalization (MI:0403); Coimmunoprecipitation (MI:0019) |
| hcv0098 | CORE | PPARA | 5465 | 17764115 | Immunoblotting (MI:0113); Affinity Chromatography (MI:0004) |
| hcv0099 | CORE | ACY3 | 91703 | 18158989 | Two Hybrid Test (MI:0018) |
| hcv0100 | CORE | MCL1 | 4170 | 19605477 | Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0101 | CORE | CORE | 951475 | 8615040 | Two Hybrid Test (MI:0018); Far-Western blot (MI:0047); Mutational analysis (MI:0074); GST pull-down (MI:0059); Affinity Chromatography (MI:0004); Far-Western blot (MI:0047); Cross Linking (MI:0030) |
| hcv0102 | CORE | CORE | 951475 | 9191926 | Affinity Chromatography (MI:0004); Two Hybrid Test (MI:0018) |
| hcv0103 | E1 | E2 | 951475 | 8083956 | Coimmunoprecipitation (MI:0019); Colocalization (MI:0403); sedimentation (MI:0029) |
| hcv0104 | E1 | CORE | 951475 | 8764026 | Mutational analysis (MI:0074); Coimmunoprecipitation (MI:0019) |
| hcv0105 | E1 | LTF | 4057 | 9223490 | Far-Western blot (MI:0047); Coimmunoprecipitation (MI:0019); Pull-down (MI:0096) |
| hcv0106 | E1 | CALR | 811 | 9557669 | Metabolic labeling (MI:2131); Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0107 | E1 | CANX | 821 | 9557669 | Metabolic labeling (MI:2131); Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0108 | E1 | HSPA5 | 3309 | 9557669 | Metabolic labeling (MI:2131); Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0109 | E1 | E2 | 951475 | 10807921 | Coimmunoprecipitation (MI:0019); Mutational analysis (MI:0074); circular dichroism spectroscopy (MI:0016); Three dimensional structure (MI:0105) |
| hcv0110 | E1 | E2 | 951475 | 12502883 | Coimmunoprecipitation (MI:0019) |
| hcv0111 | E1 | CD209 | 30835 | 12634366 | Enzyme-linked immunosorbent assay (MI:0411) |

| hcv0112 | E1 | CLEC4M | 10332 | 12634366 | Enzyme-linked immunosorbent assay (MI:0411) |
|---------|----|--------|-------|----------|---------------------------------------------|
| hcv0113 | E1 | E2 | 951475 | 15136562 | Mutational analysis (MI:0074); Coimmunoprecipitation (MI:0019) |
| hcv0114 | E1 | CD209 | 30835 | 15254204 | Mutational analysis (MI:0074); Enzyme-linked immunosorbent assay (MI:0411) |
| hcv0115 | E1 | JUN | 3725 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0116 | E1 | NR4A1 | 3164 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0117 | E1 | PFN1 | 5216 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0118 | E1 | SETD2 | 29072 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0119 | E1 | TMSB4X | 7114 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0120 | E1 | NS5A | 951475 | 10721731 | Two Hybrid Test (MI:0018) |
| hcv0121 | E1 | E2 | 951475 | 8212557 | Coimmunoprecipitation (MI:0019) |
| hcv0122 | E2 | LTF | 4057 | 9223490 | Far-Western blot (MI:0047); Coimmunoprecipitation (MI:0019); Pull-down (MI:0096) |
| hcv0123 | E2 | CALR | 811 | 9557669 | Metabolic labeling (MI:2131); Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0124 | E2 | CANX | 821 | 9557669 | Metabolic labeling (MI:2131); Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0125 | E2 | HSPA5 | 3309 | 9557669 | Metabolic labeling (MI:2131); Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0126 | E2 | EIF2AK2 | 5610 | 10390359 | Kinase assay (MI:0424) (MI:0424); His pull-down (MI:0061) |
| hcv0127 | E2 | NS3 | 951475 | 10721731 | Two Hybrid Test (MI:0018) |
| hcv0128 | E2 | CD81 | 975 | 10846074 | Competition Binding Experiment (MI:0405); Enzyme-linked immunosorbent assay (MI:0411) |
| hcv0129 | E2 | CD81 | 975 | 11080483 | Enzyme-linked immunosorbent assay (MI:0411); Immunofluorescence (MI:0022); Competition Binding Experiment (MI:0405) |
| hcv0130 | E2 | SCARB1 | 949 | 12356718 | Coimmunoprecipitation (MI:0019); Western blot (MI:0113) |
| hcv0131 | E2 | CD81 | 975 | 12522210 | Far-Western blot (MI:0047) |
| hcv0132 | E2 | LTF | 4057 | 12522210 | Far-Western blot (MI:0047) |
| hcv0133 | E2 | TF | 7018 | 12522210 | Far-Western blot (MI:0047) |
| hcv0134 | E2 | CD81 | 975 | 12604806 | Affinity Chromatography (MI:0004) |
| hcv0135 | E2 | EIF2AK3 | 9451 | 12610133 | Coimmunoprecipitation (MI:0019); Western blot (MI:0113); Kinase assay (MI:0424) (MI:0424) |
| hcv0136 | E2 | CD209 | 30835 | 12634366 | Fluorescence-activated cell sorting (MI:0054); Enzyme-linked immunosorbent assay (MI:0411); Flow cytometry (MI:0054); Western blot (MI:0113); Immunostaining (MI:0022) |
| hcv0137 | E2 | CLEC4M | 10332 | 12634366 | Fluorescence-activated cell sorting (MI:0054); Enzyme-linked immunosorbent assay (MI:0411); Immunostaining (MI:0022) |
| hcv0138 | E2 | SDC2 | 6383 | 12867431 | Enzyme-linked immunosorbent assay (MI:0411); Flow cytometry (MI:0054) |
| hcv0139 | E2 | CD209 | 30835 | 15254204 | Mutational analysis (MI:0074); Enzyme-linked immunosorbent assay (MI:0411) |
| hcv0140 | E2 | HOXD8 | 3234 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0141 | E2 | ITGB1 | 3688 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0142 | E2 | FAM135A | 57579 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0143 | E2 | NR4A1 | 3164 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0144 | E2 | PSMA6 | 5687 | 18985028 | Two Hybrid Test (MI:0018) |

| hcv0145 | E2 | SETD2 | 29072 | 18985028 | Two Hybrid Test (MI:0018) |
|---|---|---|---|---|---|
| hcv0146 | E2 | SMEK2 | 57223 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0147 | F | C14orf135 | 64430 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0148 | F | ZNF83 | 55769 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0149 | F | PFDN2 | 5202 | 16876117 | Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Immunofluorescence (MI:0022); Coimmunoprecipitation (MI:0019) |
| hcv0150 | F | PFDN5 | 5204 | 18398700 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Confocal microscopy (MI:0663) |
| hcv0151 | F | AGT | 183 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0152 | F | AZGP1 | 563 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0153 | F | CTSB | 1508 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0154 | F | MPDU1 | 9526 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0155 | F | RAB14 | 51552 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0156 | F | SERPINC1 | 462 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0157 | F | ST3GAL1 | 6482 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0158 | F | vitronectin | 7448 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0159 | F | ZG16 | 653808 | 16237761 | Two Hybrid Test (MI:0018) |
| hcv0160 | NS2 | NS4A | 951475 | 10721731 | Two Hybrid Test (MI:0018); Affinity Chromatography (MI:0004) |
| hcv0161 | NS2 | NS3 | 951475 | 11559826 | Size-exclusion chromatography (MI:0071); Mutational analysis (MI:0074) |
| hcv0162 | NS2 | CIDEB | 27141 | 12595532 | Two Hybrid Test (MI:0018); Colocalization (MI:0403); Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0163 | NS2 | NS4B | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663) |
| hcv0164 | NS2 | NS4A | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663) |
| hcv0165 | NS2 | NS5B | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663) |
| hcv0166 | NS2 | NS2 | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Affinity Chromatography (MI:0004) |
| hcv0167 | NS2 | NS3 | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Affinity Chromatography (MI:0004); Immunostaining (MI:0022) |
| hcv0168 | NS2 | C7 | 730 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0169 | NS2 | FBLN5 | 10516 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0170 | NS2 | HOXD8 | 3234 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0171 | NS2 | NR4A1 | 3164 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0172 | NS2 | POU3F2 | 5454 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0173 | NS2 | SETD2 | 29072 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0174 | NS2 | TRIM27 | 5987 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0175 | NS2 | NS5A | 951475 | 12692242 | Coimmunoprecipitation (MI:0019) |
| hcv0176 | NS3 | NS4A | 951475 | 7494258 | Coimmunoprecipitation (MI:0019); Mutational analysis (MI:0074) |

| hcv0177 | NS3 | MBP | 4155 | 8647104 | Affinity Chromatography (MI:0004) |
|---------|-----|-----|------|---------|-----------------------------------|
| hcv0178 | NS3 | PRM1 | 5619 | 8647104 | Affinity Chromatography (MI:0004) |
| hcv0179 | NS3 | HIST4H4 | 121504 | 8647104 | Affinity Chromatography (MI:0004) |
| hcv0180 | NS3 | HIST3H2BB | 128312 | 8647104 | Affinity Chromatography (MI:0004) |
| hcv0181 | NS3 | PRKACA | 5566 | 8647104 | GST pull-down (MI:0059); autoradiography (MI:0833); Affinity Chromatography (MI:0004) |
| hcv0182 | NS3 | PRKACA | 5566 | 9060639 | Competition Binding Experiment (MI:0405); Far-Western blot (MI:0047) |
| hcv0183 | NS3 | NS3 | 951475 | 9187654 | Three dimensional structure (MI:0105) |
| hcv0184 | NS3 | TP53 | 7157 | 9827557 | Coimmunoprecipitation (MI:0019) |
| hcv0185 | NS3 | NS4A | 951475 | 9827557 | Coimmunoprecipitation (MI:0019) |
| hcv0186 | NS3 | NS4A | 951475 | 10220351 | Gel filtration (MI:0071); Affinity Chromatography (MI:0004) |
| hcv0187 | NS3 | HIST3H2BB | 128312 | 10405893 | Proteolytic fragmentation (MI:0079); microsequencing (MI:0093); specific histone binding assay (MI:0079); Gel filtration (MI:0071) |
| hcv0188 | NS3 | HIST4H4 | 121504 | 10405893 | Proteolytic fragmentation (MI:0079); microsequencing (MI:0093); specific histone binding assay (MI:0079); Gel filtration (MI:0071) |
| hcv0189 | NS3 | SERPINF2 | 5345 | 10570951 | Immunoblotting (MI:0113); Western blot (MI:0113) |
| hcv0190 | NS3 | SERPING1 | 710 | 10570951 | Immunoblotting (MI:0113); Western blot (MI:0113) |
| hcv0191 | NS3 | NS4A | 951475 | 10721731 | Two Hybrid Test (MI:0018) |
| hcv0192 | NS3 | PRMT5 | 10419 | 11152681 | Two Hybrid Test (MI:0018); Western blot (MI:0113) |
| hcv0193 | NS3 | PRMT1 | 3276 | 11483748 | Coimmunoprecipitation (MI:0019); Immunoblotting (MI:0113); Affinity Chromatography (MI:0004); Immunoblotting (MI:0113); GST pull-down (MI:0059) |
| hcv0194 | NS3 | NS3 | 951475 | 12465917 | Three dimensional structure (MI:0105) |
| hcv0195 | NS3 | NS4B | 951475 | 12692242 | Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); GST pull-down (MI:0059) |
| hcv0196 | NS3 | NS4A | 951475 | 12692242 | Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Immunostaining (MI:0022); GST pull-down (MI:0059) |
| hcv0197 | NS3 | NS3 | 951475 | 12692242 | Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); GST pull-down (MI:0059) |
| hcv0198 | NS3 | IRF3 | 3661 | 12702807 | Mutational analysis (MI:0074); Immunoblotting (MI:0113) |
| hcv0199 | NS3 | SNRPD1 | 6632 | 14524621 | Two Hybrid Test (MI:0018); Mutational analysis (MI:0074); Affinity Chromatography (MI:0004); Coimmunoprecipitation (MI:0019) |
| hcv0200 | NS3 | NS4A | 951475 | 14627400 | Three dimensional structure (MI:0105) |
| hcv0201 | NS3 | NS4A | 951475 | 14984200 | Three dimensional structure (MI:0105) |
| hcv0202 | NS3 | PSMB8 | 5696 | 15303969 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0203 | NS3 | SMAD3 | 4088 | 15334054 | Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0204 | NS3 | NS4A | 951475 | 15501035 | Three dimensional structure (MI:0105) |
| hcv0205 | NS3 | TLR2 | 7097 | 15521019 | Electrophoretic mobility shift assay (MI:0413); Western blot (MI:0113); Confocal microscopy (MI:0663); Flow cytometry (MI:0054) |
| hcv0206 | NS3 | TICAM1 | 148022 | 15767257 | Immunoblotting (MI:0113); Affinity Chromatography (MI:0004); Immunofluorescence (MI:0022); circular dichroism spectroscopy |

| | | | | | (MI:0016); Three dimensional structure (MI:0105) |
|---|---|---|---|---|---|
| hcv0207 | NS3 | PTBP2 | 58155 | 15823607 | Immunofluorescence (MI:0022); GST pull-down (MI:0059); Colocalization (MI:0403); Coimmunoprecipitation (MI:0019) |
| hcv0208 | NS3 | IKBKE | 9641 | 15841462 | Western blot (MI:0113); Coimmunoprecipitation (MI:0019) |
| hcv0209 | NS3 | TBK1 | 29110 | 15841462 | Western blot (MI:0113); Coimmunoprecipitation (MI:0019) |
| hcv0210 | NS3 | ERC1 | 23085 | 16033967 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059); Confocal microscopy (MI:0663); immunoelectron microscopy (MI:0040); Two Hybrid Test (MI:0018) |
| hcv0211 | NS3 | NS4A | 951475 | 16078825 | Three dimensional structure (MI:0105) |
| hcv0212 | NS3 | NS4A | 951475 | 16112862 | Three dimensional structure (MI:0105) |
| hcv0213 | NS3 | NS4A | 951475 | 16413182 | Three dimensional structure (MI:0105) |
| hcv0214 | NS3 | NS4A | 951475 | 16876765 | Confocal microscopy (MI:0663); Colocalization (MI:0403) |
| hcv0215 | NS3 | VISA | 57506 | 17289677 | Immunoblotting (MI:0113); Coimmunoprecipitation (MI:0019); Confocal microscopy (MI:0663) |
| hcv0216 | NS3 | NS5B | 951475 | 18179252 | Pull-down (MI:0096); surface plasmon resonance (MI:0107); fluorescence polarization binding assay (MI:0053) |
| hcv0217 | NS3 | ACTN1 | 87 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0218 | NS3 | ACTN2 | 88 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0219 | NS3 | AEBP1 | 165 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0220 | NS3 | ANKRD12 | 23253 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0221 | NS3 | ANKRD28 | 23243 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0222 | NS3 | ARFIP2 | 23647 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0223 | NS3 | ARHGEF6 | 9459 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0224 | NS3 | ARNT | 405 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0225 | NS3 | ARS2 | 51593 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0226 | NS3 | ASXL1 | 171023 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0227 | NS3 | B2M | 567 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0228 | NS3 | BCAN | 63827 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0229 | NS3 | BCKDK | 10295 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0230 | NS3 | BCL2A1 | 597 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0231 | NS3 | BCL6 | 604 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0232 | NS3 | BZRAP1 | 9256 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0233 | NS3 | C10orf18 | 54906 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0234 | NS3 | C10orf6 | 55719 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0235 | NS3 | C12orf41 | 54934 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0236 | NS3 | INF2 | 64423 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0237 | NS3 | C16orf7 | 9605 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0238 | NS3 | BEND5 | 79656 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0239 | NS3 | C1orf94 | 84970 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0240 | NS3 | C9orf30 | 91283 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0241 | NS3 | CALCOCO2 | 10241 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |

| hcv0242 | NS3 | CBY1 | 25776 | 18985028 | Two Hybrid Test (MI:0018) |
|---------|-----|------|-------|----------|---------------------------|
| hcv0243 | NS3 | CCDC21 | 64793 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0244 | NS3 | CCDC37 | 348807 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0245 | NS3 | CCDC52 | 152185 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0246 | NS3 | CCDC66 | 285331 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0247 | NS3 | INO80E | 283899 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0248 | NS3 | CCHCR1 | 54535 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0249 | NS3 | CD5L | 922 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0250 | NS3 | CDC23 | 8697 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0251 | NS3 | CELSR2 | 1952 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0252 | NS3 | CEP152 | 22995 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0253 | NS3 | CEP192 | 55125 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0254 | NS3 | CFP | 5199 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0255 | NS3 | CHPF | 79586 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0256 | NS3 | CORO1B | 57175 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0257 | NS3 | CSNK2B | 1460 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0258 | NS3 | CTGF | 1490 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0259 | NS3 | ALG13 | 79868 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0260 | NS3 | DEAF1 | 10522 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0261 | NS3 | DES | 1674 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0262 | NS3 | DLAT | 1737 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0263 | NS3 | DOCK7 | 85440 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0264 | NS3 | DPF1 | 8193 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0265 | NS3 | DPP7 | 29952 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0266 | NS3 | EEF1A1 | 1915 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0267 | NS3 | EFEMP1 | 2202 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0268 | NS3 | EFEMP2 | 30008 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0269 | NS3 | EIF1 | 10209 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0270 | NS3 | EIF4ENIF1 | 56478 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0271 | NS3 | FAM120B | 84498 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0272 | NS3 | FAM65A | 79567 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0273 | NS3 | FBF1 | 85302 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0274 | NS3 | FBLN1 | 2192 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0275 | NS3 | FBLN2 | 2199 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0276 | NS3 | FBLN5 | 10516 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0277 | NS3 | FBN1 | 2200 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0278 | NS3 | FBN3 | 84467 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0279 | NS3 | FES | 2242 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0280 | NS3 | FIGNL1 | 63979 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0281 | NS3 | FLAD1 | 80308 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0282 | NS3 | C19orf66 | 55337 | 18985028 | Two Hybrid Test (MI:0018) |

| hcv0283 | NS3 | FN1 | 2335 | 18985028 | Two Hybrid Test (MI:0018) |
|---------|-----|-----|------|----------|---------------------------|
| hcv0284 | NS3 | FRMPD4 | 9758 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0285 | NS3 | FRS3 | 10817 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0286 | NS3 | FTH1 | 2495 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0287 | NS3 | FUCA2 | 2519 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0288 | NS3 | GAA | 2548 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0289 | NS3 | GBP2 | 2634 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0290 | NS3 | GFAP | 2670 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0291 | NS3 | GNB2 | 2783 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0292 | NS3 | GON4L | 54856 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0293 | NS3 | HIVEP2 | 3097 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0294 | NS3 | HNRNPK | 3190 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0295 | NS3 | HOMER3 | 9454 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0296 | NS3 | IQWD1 | 55827 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0297 | NS3 | ITGB4 | 3691 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0298 | NS3 | JAG2 | 3714 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0299 | NS3 | JUN | 3725 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0300 | NS3 | KHDRBS1 | 10657 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0301 | NS3 | KIAA1549 | 57670 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0302 | NS3 | KIF17 | 57576 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0303 | NS3 | KIF7 | 374654 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0304 | NS3 | KPNA1 | 3836 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0305 | NS3 | L3MBTL3 | 84456 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0306 | NS3 | LAMA5 | 3911 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0307 | NS3 | LAMB2 | 3913 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0308 | NS3 | LAMC3 | 10319 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0309 | NS3 | LDB1 | 8861 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0310 | NS3 | LRRC7 | 57554 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0311 | NS3 | LRRCC1 | 85444 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0312 | NS3 | LTBP4 | 8425 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0313 | NS3 | LZTS2 | 84445 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0314 | NS3 | MAGED1 | 9500 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0315 | NS3 | MAPK7 | 5598 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0316 | NS3 | MEGF8 | 1954 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0317 | NS3 | MLLT4 | 4301 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0318 | NS3 | MLXIP | 22877 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0319 | NS3 | MORC4 | 79710 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0320 | NS3 | MORF4L1 | 10933 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0321 | NS3 | MVP | 9961 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0322 | NS3 | NAP1L1 | 4673 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0323 | NS3 | NAP1L2 | 4674 | 18985028 | Two Hybrid Test (MI:0018) |

| hcv0324 | NS3 | NCAN | 1463 | 18985028 | Two Hybrid Test (MI:0018) |
|---------|-----|------|------|----------|---------------------------|
| hcv0325 | NS3 | NDC80 | 10403 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0326 | NS3 | NEFL | 4747 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0327 | NS3 | NEFM | 4741 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0328 | NS3 | NELL1 | 4745 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0329 | NS3 | NELL2 | 4753 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0330 | NS3 | NID1 | 4811 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0331 | NS3 | NID2 | 22795 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0332 | NS3 | NOTCH1 | 4851 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0333 | NS3 | N-PAC | 84656 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0334 | NS3 | NUP62 | 23636 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0335 | NS3 | OBSCN | 84033 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0336 | NS3 | PARP4 | 143 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0337 | NS3 | PCYT2 | 5833 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0338 | NS3 | PDE4DIP | 9659 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0339 | NS3 | PDLIM5 | 10611 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0340 | NS3 | PICK1 | 9463 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0341 | NS3 | PKNOX1 | 5316 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0342 | NS3 | PLEKHG4 | 25894 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0343 | NS3 | PNPLA8 | 50640 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0344 | NS3 | PRRC1 | 133619 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0345 | NS3 | PSMB9 | 5698 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0346 | NS3 | PSME3 | 10197 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0347 | NS3 | PTPRN2 | 5799 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0348 | NS3 | RABEP1 | 9135 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0349 | NS3 | RAI14 | 26064 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0350 | NS3 | RASAL2 | 9462 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0351 | NS3 | RBM4 | 5936 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0352 | NS3 | RCN3 | 57333 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0353 | NS3 | RGNEF | 64283 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0354 | NS3 | RICS | 9743 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0355 | NS3 | RINT1 | 60561 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0356 | NS3 | RNF31 | 55072 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0357 | NS3 | ROGDI | 79641 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0358 | NS3 | KIAA2022 | 340533 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0359 | NS3 | RUSC2 | 9853 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0360 | NS3 | SBF1 | 6305 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0361 | NS3 | SDCCAG8 | 10806 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0362 | NS3 | SECISBP2 | 79048 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0363 | NS3 | 10-Sep | 151011 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0364 | NS3 | SERTAD1 | 29950 | 18985028 | Two Hybrid Test (MI:0018) |

| hcv0365 | NS3 | SESTD1 | 91404 | 18985028 | Two Hybrid Test (MI:0018) |
|---------|-----|--------|-------|----------|---------------------------|
| hcv0366 | NS3 | SF3B2 | 10992 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0367 | NS3 | SIAH1 | 6477 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0368 | NS3 | SLIT1 | 6585 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0369 | NS3 | SLIT2 | 9353 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0370 | NS3 | SLIT3 | 6586 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0371 | NS3 | SMURF2 | 64750 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0372 | NS3 | SNX4 | 8723 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0373 | NS3 | SPOCK3 | 50859 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0374 | NS3 | SPON1 | 10418 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0375 | NS3 | SRPX2 | 27286 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0376 | NS3 | SSX2IP | 117178 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0377 | NS3 | STAB1 | 23166 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0378 | NS3 | STAT3 | 6774 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0379 | NS3 | SVEP1 | 79987 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0380 | NS3 | SYNE1 | 23345 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0381 | NS3 | SYNPO2 | 171024 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0382 | NS3 | TAF1 | 6872 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0383 | NS3 | TBC1D2B | 23102 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0384 | NS3 | TBXAS1 | 6916 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0385 | NS3 | TGFB1I1 | 7041 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0386 | NS3 | THAP1 | 55145 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0387 | NS3 | TMEM63B | 55362 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0388 | NS3 | TRIM23 | 373 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0389 | NS3 | TRIM27 | 5987 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0390 | NS3 | TRIO | 7204 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0391 | NS3 | TRIP11 | 9321 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0392 | NS3 | TXNDC11 | 51061 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0393 | NS3 | UBA3 | 9039 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0394 | NS3 | USHBP1 | 83878 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0395 | NS3 | UXT | 8409 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0396 | NS3 | VCAN | 1462 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0397 | NS3 | VIM | 7431 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0398 | NS3 | VWF | 7450 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0399 | NS3 | XAB2 | 56949 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0400 | NS3 | XRN2 | 22803 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0401 | NS3 | YY1AP1 | 55249 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0402 | NS3 | ZBTB1 | 22890 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0403 | NS3 | ZCCHC7 | 84186 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0404 | NS3 | ZHX3 | 23051 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0405 | NS3 | ZMYM2 | 7750 | 18985028 | Two Hybrid Test (MI:0018) |

| hcv0406 | NS3 | ZNF281 | 23528 | 18985028 | Two Hybrid Test (MI:0018) |
|---------|-----|--------|-------|----------|---------------------------|
| hcv0407 | NS3 | ZNF410 | 57862 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0408 | NS3 | ZZZ3 | 26009 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0409 | NS3 | NS5B | 951475 | 10721731 | Two Hybrid Test (MI:0018) |
| hcv0410 | NS3 | NS3 | 951475 | 11119590 | Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018) |
| hcv0411 | NS3 | NS5B | 951475 | 12235135 | Far-Western blot (MI:0047) |
| hcv0412 | NS3 | NS4B | 951475 | 12235135 | Far-Western blot (MI:0047) |
| hcv0413 | NS3 | NS5B | 951475 | 12692242 | Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018) |
| hcv0414 | NS3 | CASP8 | 841 | 15476874 | Coimmunoprecipitation (MI:0019) |
| hcv0415 | NS3 | NS3 | 951475 | 16166788 | Two Hybrid Test (MI:0018) |
| hcv0416 | NS3 | NS4A | 951475 | 16324764 | Immunofluorescence (MI:0022); Coimmunoprecipitation (MI:0019) |
| hcv0417 | NS3 | RSPH3 | 83861 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0418 | NS4A | NS3 | 951475 | 8861917 | Three dimensional structure (MI:0105) |
| hcv0419 | NS4A | NS3 | 951475 | 9568891 | Three dimensional structure (MI:0105) |
| hcv0420 | NS4A | NS3 | 951475 | 10702283 | Three dimensional structure (MI:0105) |
| hcv0421 | NS4A | NS3 | 951475 | 12465917 | Three dimensional structure (MI:0105) |
| hcv0422 | NS4A | NS4A | 951475 | 12692242 | Affinity Chromatography (MI:0004); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); GST pull-down (MI:0059) |
| hcv0423 | NS4A | NS4B | 951475 | 12692242 | Affinity Chromatography (MI:0004); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Immunostaining (MI:0022); Affinity Chromatography (MI:0004) |
| hcv0424 | NS4A | NS5B | 951475 | 12692242 | Affinity Chromatography (MI:0004); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Immunostaining (MI:0022) |
| hcv0425 | NS4A | CREB3 | 10488 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0426 | NS4A | ELAC2 | 60528 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0427 | NS4A | HOXD8 | 3234 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0428 | NS4A | NR4A1 | 3164 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0429 | NS4A | TRAF3IP3 | 80342 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0430 | NS4A | UBQLN1 | 29979 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0431 | NS4A | NS5A | 951475 | 12692242 | Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Affinity Chromatography (MI:0004); Immunostaining (MI:0022) |
| hcv0432 | NS4A | MT2A | 4502 | 16248944 | Two Hybrid Test (MI:0018) |
| hcv0433 | NS4A | TUT1 | 64852 | 16248944 | Two Hybrid Test (MI:0018) |
| hcv0434 | NS4A | MT-CO2 | 4513 | 16248944 | Two Hybrid Test (MI:0018) |
| hcv0435 | NS4A | EEF1A1 | 1915 | 16927014 | GST pull-down (MI:0059) |
| hcv0436 | NS4A | CAMLG | 819 | 17429534 | Two Hybrid Test (MI:0018) |
| hcv0437 | NS4A | NS5A | 951475 | 9261364 | Coimmunoprecipitation (MI:0019) |
| hcv0438 | NS4B | NS4A | 951475 | 9261364 | Coimmunoprecipitation (MI:0019) |
| hcv0439 | NS4B | ATF6B | 1388 | 12445808 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Confocal microscopy (MI:0663) |

| | | | | | |
|---|---|---|---|---|---|
| hcv0440 | NS4B | ATF6A | 22926 | 12445808 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Confocal microscopy (MI:0663) |
| hcv0441 | NS4B | NS5B | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Immunostaining (MI:0022); Affinity Chromatography (MI:0004) |
| hcv0442 | NS4B | NS4B | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Affinity Chromatography (MI:0004) |
| hcv0443 | NS4B | NS5B | 951475 | 12235135 | Far-Western blot (MI:0047) |
| hcv0444 | NS4B | TNXB | 7148 | 12445808 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Confocal microscopy (MI:0663) |
| hcv0445 | NS4B | NS5A | 951475 | 12692242 | Coimmunoprecipitation (MI:0019) |
| hcv0446 | NS4B | NS5A | 951475 | 15016871 | Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0447 | NS4B | RTN3 | 10313 | 16261208 | Two Hybrid Test (MI:0018) |
| hcv0448 | NS4B | RBP4 | 5950 | 16261208 | Two Hybrid Test (MI:0018) |
| hcv0449 | NS4B | NDUFV3 | 4731 | 16261208 | Two Hybrid Test (MI:0018) |
| hcv0450 | NS4B | FGG | 2266 | 16261208 | Two Hybrid Test (MI:0018) |
| hcv0451 | NS4B | MT-CO3 | 4514 | 16261208 | Two Hybrid Test (MI:0018) |
| hcv0452 | NS4B | RAB5A | 5868 | 17301141 | Coimmunoprecipitation (MI:0019) |
| hcv0453 | NS5A | EIF2AK2 | 5610 | 9143277 | Two Hybrid Test (MI:0018) |
| hcv0454 | NS5A | CSNK2A1 | 1457 | 10208859 | GST pull-down (MI:0059); Affinity Chromatography (MI:0004); Kinase assay (MI:0424) (MI:0424); Mutational analysis (MI:0074) |
| hcv0455 | NS5A | GRB2 | 2885 | 10318918 | Immunoblotting (MI:0113); Affinity Chromatography (MI:0004); Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019) |
| hcv0456 | NS5A | VAPA | 9218 | 10544080 | Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022); Immunostaining (MI:0022); Affinity Chromatography (MI:0004) |
| hcv0457 | NS5A | SRCAP | 10847 | 10702287 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004); Immunostaining (MI:0022); Pull-down (MI:0096) |
| hcv0458 | NS5A | IPO5 | 3843 | 10799599 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019);  Affinity Chromatography (MI:0004) |
| hcv0459 | NS5A | TP53 | 7157 | 11152513 | Pull-down (MI:0096); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Colocalization (MI:0403); Affinity Chromatography (MI:0004); Immunostaining (MI:0022) |
| hcv0460 | NS5A | CDC2 | 983 | 11278402 | Western blot (MI:0113); Coimmunoprecipitation (MI:0019) |
| hcv0461 | NS5A | CDK2 | 1017 | 11278402 | Western blot (MI:0113); Coimmunoprecipitation (MI:0019) |
| hcv0462 | NS5A | APOA1 | 335 | 11878923 | Immunofluorescence (MI:0022); Confocal microscopy (MI:0663); Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004); Immunostaining (MI:0022) |
| hcv0463 | NS5A | TP53 | 7157 | 12101418 | Coimmunoprecipitation (MI:0019); Confocal microscopy (MI:0663); Affinity Chromatography (MI:0004) |
| hcv0464 | NS5A | TAF9 | 6880 | 12101418 | Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022) |
| hcv0465 | NS5A | PIK3R1 | 5295 | 12186904 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |

| hcv0466 | NS5A | TP53 | 7157 | 12379483 | Coimmunoprecipitation (MI:0019); Pull-down (MI:0096); Affinity Chromatography (MI:0004) |
|---------|------|------|------|----------|------------------------------------------------------------------------------------------|
| hcv0467 | NS5A | TBP | 6908 | 12379483 | Coimmunoprecipitation (MI:0019); Pull-down (MI:0096); Affinity Chromatography (MI:0004) |
| hcv0468 | NS5A | BIN1 | 274 | 12604805 | Mass spectrometry (MI:0943); Mutational analysis (MI:0074); Coimmunoprecipitation (MI:0019); Colocalization (MI:0403); Affinity Chromatography (MI:0004); Immunostaining (MI:0022) |
| hcv0469 | NS5A | PITX1 | 5307 | 12620797 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019) |
| hcv0470 | NS5A | NS5B | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Immunostaining (MI:0022); Affinity Chromatography (MI:0004) |
| hcv0471 | NS5A | NS5A | 951475 | 12692242 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018); Confocal microscopy (MI:0663); Affinity Chromatography (MI:0004) |
| hcv0472 | NS5A | BAX | 581 | 12925958 | Immunofluorescence (MI:0022); Coimmunoprecipitation (MI:0019) |
| hcv0473 | NS5A | SSB | 6741 | 12963047 | Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0474 | NS5A | HCK | 3055 | 14993658 | Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0475 | NS5A | LCK | 3932 | 14993658 | Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0476 | NS5A | LYN | 4067 | 14993658 | Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0477 | NS5A | FYN | 2534 | 14993658 | Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0478 | NS5A | VAPA | 9218 | 15016871 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059); autoradiography (MI:0833); Affinity Chromatography (MI:0004) |
| hcv0479 | NS5A | OAS1 | 4938 | 15039538 | GST pull-down (MI:0059); Immunoblotting (MI:0113); Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022) |
| hcv0480 | NS5A | JAK1 | 3716 | 15063116 | Coimmunoprecipitation (MI:0019) |
| hcv0481 | NS5A | APOE | 348 | 15326295 | Two Hybrid Test (MI:0018) |
| hcv0482 | NS5A | VAPA | 9218 | 15326295 | Western blot (MI:0113); Two Hybrid Test (MI:0018) |
| hcv0483 | NS5A | NS5A | 951475 | 15326295 | Two Hybrid Test (MI:0018) |
| hcv0484 | NS5A | AHNAK | 79026 | 15607035 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059) |
| hcv0485 | NS5A | SFRP4 | 6424 | 15607035 | Two Hybrid Test (MI:0018) |
| hcv0486 | NS5A | CRABP1 | 1381 | 15607035 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059) |
| hcv0487 | NS5A | FTH1 | 2495 | 15607035 | Two Hybrid Test (MI:0018) |
| hcv0488 | NS5A | CCDC86 | 79080 | 15607035 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059) |
| hcv0489 | NS5A | TACSTD2 | 4070 | 15607035 | Two Hybrid Test (MI:0018) |
| hcv0490 | NS5A | PI4KA | 5297 | 15607035 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059) |
| hcv0491 | NS5A | PTMA | 5757 | 15607035 | Two Hybrid Test (MI:0018) |
| hcv0492 | NS5A | ARAP1 | 116985 | 15607035 | Two Hybrid Test (MI:0018); GST pull-down (MI:0059) |
| hcv0493 | NS5A | NDRG1 | 10397 | 15607035 | Two Hybrid Test (MI:0018) |
| hcv0494 | NS5A | MGP | 4256 | 15607035 | Two Hybrid Test (MI:0018) |
| hcv0495 | NS5A | CEP57 | 9702 | 15607035 | Two Hybrid Test (MI:0018) |
| hcv0496 | NS5A | C9orf6 | 54942 | 15607035 | Two Hybrid Test (MI:0018) |

| hcv0497 | NS5A | FBXL2 | 25827 | 15893726 | [(3)H]mevalonate labeling (MI:2131); Coimmunoprecipitation (MI:0019) |
|---|---|---|---|---|---|
| hcv0498 | NS5A | NS5A | 951475 | 15902263 | Three dimensional structure (MI:0105) |
| hcv0499 | NS5A | CSK | 1445 | 16139795 | Affinity Chromatography (MI:0004); Mass spectrometry (MI:0943); GST pull-down (MI:0059) |
| hcv0500 | NS5A | SRC | 6714 | 16139795 | Affinity Chromatography (MI:0004); Mass spectrometry (MI:0943); GST pull-down (MI:0059) |
| hcv0501 | NS5A | APOB | 338 | 16203724 | Coimmunoprecipitation (MI:0019) |
| hcv0502 | NS5A | VAPB | 9217 | 16227268 | Coimmunoprecipitation (MI:0019); Mutational analysis (MI:0074) |
| hcv0503 | NS5A | RAF1 | 5894 | 16405965 | Affinity Chromatography (MI:0004); Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022); Western blot (MI:0113) |
| hcv0504 | NS5A | TGFBR1 | 7046 | 16407286 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019) |
| hcv0505 | NS5A | PPP2R4 | 5524 | 16460864 | GST pull-down (MI:0059); His pull-down (MI:0061); Immunofluorescence (MI:0022); Coimmunoprecipitation (MI:0019); Immunoblotting (MI:0113) |
| hcv0506 | NS5A | BIN1 | 274 | 16530520 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Confocal microscopy (MI:0663); Mutational analysis (MI:0074) |
| hcv0507 | NS5A | TRAF2 | 7186 | 16581780 | Coimmunoprecipitation (MI:0019) |
| hcv0508 | NS5A | FKBP8 | 23770 | 16844119 | Two Hybrid Test (MI:0018) |
| hcv0509 | NS5A | FKBP8 | 23770 | 17024179 | Coimmunoprecipitation (MI:0019) |
| hcv0510 | NS5A | HSP90AA1 | 3320 | 17024179 | Coimmunoprecipitation (MI:0019) |
| hcv0511 | NS5A | PTPLAD1 | 51495 | 18160438 | Two Hybrid Test (MI:0018) |
| hcv0512 | NS5A | ACLY | 47 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0513 | NS5A | ARFIP1 | 27236 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0514 | NS5A | AXIN1 | 8312 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0515 | NS5A | BEND7 | 222389 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0516 | NS5A | CADPS | 8618 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0517 | NS5A | CADPS2 | 93664 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0518 | NS5A | CEP120 | 153241 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0519 | NS5A | CENPC1 | 1060 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0520 | NS5A | CEP250 | 11190 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0521 | NS5A | CEP63 | 80254 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0522 | NS5A | DNAJA3 | 9093 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0523 | NS5A | EFEMP1 | 2202 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0524 | NS5A | FHL2 | 2274 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0525 | NS5A | GOLGA2 | 2801 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0526 | NS5A | GPS2 | 2874 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0527 | NS5A | IGLL1 | 3543 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0528 | NS5A | ITGAL | 3683 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0529 | NS5A | LIMS2 | 55679 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0530 | NS5A | MOBK1B | 55233 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |

| hcv0531 | NS5A | NAP1L1 | 4673 | 18985028 | Two Hybrid Test (MI:0018) |
|---------|------|--------|------|----------|---------------------------|
| hcv0532 | NS5A | NAP1L2 | 4674 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0533 | NS5A | NFE2 | 4778 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0534 | NS5A | NUCB1 | 4924 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0535 | NS5A | PARVG | 64098 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0536 | NS5A | PMVK | 10654 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0537 | NS5A | PPP1R13L | 10848 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0538 | NS5A | PSMB9 | 5698 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0539 | NS5A | RPL18A | 6142 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0540 | NS5A | RRBP1 | 6238 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0541 | NS5A | SHARPIN | 81858 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0542 | NS5A | SMYD3 | 64754 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0543 | NS5A | SORBS2 | 8470 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0544 | NS5A | SORBS3 | 10174 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0545 | NS5A | THBS1 | 7057 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0546 | NS5A | TMF1 | 7110 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0547 | NS5A | TP53BP2 | 7159 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0548 | NS5A | TRIOBP | 11078 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0549 | NS5A | TXNDC11 | 51061 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0550 | NS5A | UBASH3A | 53347 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0551 | NS5A | USP19 | 10869 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0552 | NS5A | VPS52 | 6293 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0553 | NS5A | GIN1 | 54826 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0554 | NS5A | ZNF646 | 9726 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0555 | NS5A | NS5B | 951475 | 11801599 | GST pull-down (MI:0059); Affinity Chromatography (MI:0004) |
| hcv0556 | NS5A | TRAF2 | 7186 | 11821416 | Coimmunoprecipitation (MI:0019); Colocalization (MI:0403); Immunostaining (MI:0022); Affinity Chromatography (MI:0004) |
| hcv0557 | NS5A | TRADD | 8717 | 11886269 | Pull-down (MI:0096); Coimmunoprecipitation (MI:0019); Colocalization (MI:0403) |
| hcv0558 | NS5A | GAB1 | 2549 | 12186904 | Coimmunoprecipitation (MI:0019) |
| hcv0559 | NS5A | NS3 | 951475 | 12692242 | GST pull-down (MI:0059) |
| hcv0560 | NS5A | PIK3CB | 5291 | 14709551 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |
| hcv0561 | NS5A | FBXL20 | 84961 | 15893726 | Coimmunoprecipitation (MI:0019) |
| hcv0562 | NS5A | STAT1 | 6772 | 17275127 | Coimmunoprecipitation (MI:0019); Confocal microscopy (MI:0663) |
| hcv0563 | NS5A | MYD88 | 4615 | 17567694 | Coimmunoprecipitation (MI:0019) |
| hcv0564 | NS5A | PDPK1 | 5170 | 17616579 | GST pull-down (MI:0059); Kinase assay (MI:0424) (MI:0424) |
| hcv0565 | NS5A | AKT1 | 207 | 17616579 | GST pull-down (MI:0059); Kinase assay (MI:0424) (MI:0424) |
| hcv0566 | NS5A | VPS35 | 55737 | 17616579 | Two Hybrid Test (MI:0018) |
| hcv0567 | NS5A | MAPK12 | 6300 | 17616579 | Kinase assay (MI:0424) (MI:0424); GST pull-down (MI:0059) |
| hcv0568 | NS5A | IPO4 | 79711 | 17616579 | Two Hybrid Test (MI:0018) |

| hcv0569 | NS5A | GSK3B | 2932 | 17616579 | GST pull-down (MI:0059); Kinase assay (MI:0424) (MI:0424) |
| hcv0570 | NS5A | AHSA1 | 10598 | 17616579 | Two Hybrid Test (MI:0018) |
| hcv0571 | NS5A | GSK3A | 2931 | 17616579 | GST pull-down (MI:0059); Kinase assay (MI:0424) (MI:0424) |
| hcv0572 | NS5A | CDK6 | 1021 | 17616579 | Kinase assay (MI:0424) (MI:0424); GST pull-down (MI:0059) |
| hcv0573 | NS5A | TBC1D20 | 128637 | 17686842 | Two Hybrid Test (MI:0018) |
| hcv0574 | NS5B | VAPA | 9218 | 10544080 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022); Affinity Chromatography (MI:0004); Immunostaining (MI:0022) |
| hcv0575 | NS5B | EIF4A2 | 1974 | 11922617 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Affinity Chromatography (MI:0004) |
| hcv0576 | NS5B | NCL | 4691 | 12427757 | GST pull-down (MI:0059); Affinity Chromatography (MI:0004); Immunostaining (MI:0022) |
| hcv0577 | NS5B | NS5B | 951475 | 12692242 | Two Hybrid Test (MI:0018) |
| hcv0578 | NS5B | HAO1 | 54363 | 14623081 | Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022); Confocal microscopy (MI:0663); Two Hybrid Test (MI:0018) |
| hcv0579 | NS5B | TTC4 | 7268 | 14623081 | Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022); Confocal microscopy (MI:0663); Two Hybrid Test (MI:0018) |
| hcv0580 | NS5B | ACTN1 | 87 | 14623081 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019); Immunofluorescence (MI:0022); Confocal microscopy (MI:0663); Immunostaining (MI:0022) |
| hcv0581 | NS5B | VAPA | 9218 | 15016871 | Coimmunoprecipitation (MI:0019); GST pull-down (MI:0059) |
| hcv0582 | NS5B | DDX5 | 1655 | 15113910 | Coimmunoprecipitation (MI:0019); Immunostaining (MI:0022); Two Hybrid Test (MI:0018) |
| hcv0583 | NS5B | PKN2 | 5586 | 15364941 | Coimmunoprecipitation (MI:0019); Colocalization (MI:0403); Kinase assay (MI:0424) (MI:0424); Metabolic labeling (MI:2131) |
| hcv0584 | NS5B | PTBP2 | 58155 | 15823607 | Immunofluorescence (MI:0022); GST pull-down (MI:0059); Colocalization (MI:0403); Coimmunoprecipitation (MI:0019) |
| hcv0585 | NS5B | FBXL2 | 25827 | 15893726 | [(3)H]mevalonate labeling (MI:2131); Coimmunoprecipitation (MI:0019) |
| hcv0586 | NS5B | PPIB | 5479 | 15989969 | GST pull-down (MI:0059); Coimmunoprecipitation (MI:0019); Colocalization (MI:0403); Cross Linking (MI:0030) |
| hcv0587 | NS5B | VAPB | 9217 | 16227268 | Coimmunoprecipitation (MI:0019); Mutational analysis (MI:0074); Affinity Chromatography (MI:0004) |
| hcv0588 | NS5B | CHUK | 1147 | 16581780 | Coimmunoprecipitation (MI:0019) |
| hcv0589 | NS5B | CEP250 | 11190 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0590 | NS5B | CEP68 | 23177 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0591 | NS5B | HOXD8 | 3234 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0592 | NS5B | MGC2752 | 65996 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0593 | NS5B | MOBK1B | 55233 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0594 | NS5B | NR4A1 | 3164 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0595 | NS5B | OS9 | 10956 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
| hcv0596 | NS5B | PKM2 | 5315 | 18985028 | Two Hybrid Test (MI:0018) |

| hcv0597 | NS5B | PSMB9 | 5698 | 18985028 | Two Hybrid Test (MI:0018); Co-affinity purification (MI:0025) |
|---------|------|-------|------|----------|------------------------------------------------|
| hcv0598 | NS5B | SETD2 | 29072 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0599 | NS5B | SHARPIN | 81858 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0600 | NS5B | TUBB2C | 10383 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0601 | NS5B | NS5B | 951475 | 11907226 | Gel filtration (MI:0071); Cross Linking (MI:0030); Two Hybrid Test (MI:0018); Three dimensional structure (MI:0105); Temperature sensitivity (MI:0271) |
| hcv0602 | NS5B | UBQLN1 | 29979 | 12634373 | Two Hybrid Test (MI:0018); Coimmunoprecipitation (MI:0019) |
| hcv0603 | NS5B | CINP | 51550 | 14623081 | Two Hybrid Test (MI:0018) |
| hcv0604 | NS5B | PSMB4 | 5692 | 14623081 | Two Hybrid Test (MI:0018) |
| hcv0605 | NS5B | 6-Sep | 23157 | 17229681 | Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018) |
| hcv0606 | NS5B | HNRNPA1 | 3178 | 17229681 | Coimmunoprecipitation (MI:0019); Two Hybrid Test (MI:0018) |
| hcv0607 | p7 | P7 | 951475 | 12560074 | Cross Linking (MI:0030); Immunoblotting (MI:0113) |
| hcv0608 | p7 | FMNL1 | 752 | 16094715 | Two Hybrid Test (MI:0018) |
| hcv0609 | p7 | H19 | 283120 | 16094715 | Two Hybrid Test (MI:0018) |
| hcv0610 | p7 | ISLR | 3671 | 16094715 | Two Hybrid Test (MI:0018) |
| hcv0611 | p7 | MS4A6A | 64231 | 16094715 | Two Hybrid Test (MI:0018) |
| hcv0612 | p7 | NUP214 | 8021 | 16094715 | Two Hybrid Test (MI:0018) |
| hcv0613 | p7 | SSR4 | 6748 | 16094715 | Two Hybrid Test (MI:0018) |
| hcv0614 | p7 | STRBP | 55342 | 16094715 | Two Hybrid Test (MI:0018) |
| hcv0615 | p7 | CREB3 | 10488 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0616 | p7 | FBLN2 | 2199 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0617 | p7 | FXYD6 | 53826 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0618 | p7 | LMNB1 | 4001 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0619 | p7 | SLIT2 | 9353 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0620 | p7 | UBQLN1 | 29979 | 18985028 | Two Hybrid Test (MI:0018) |
| hcv0621 | p7 | UBQLN4 | 56893 | 18985028 | Two Hybrid Test (MI:0018) |

# Appendix IV (Chapter 3)

HCV proteins assigned with information relating to their genotype/strain sources.

| HCVpro ID | Molecule A | HCV Genotype | HCV isolate/strain | HCV Molecule A Protein Accession | Molecule B Gene Symbols | HCV Molecule B Protein Accession | Genotype Source |
|---|---|---|---|---|---|---|---|
| hcv0001 | CORE | HCV genotype 1a | H77 | NP_751919 | LTBR | | VirHostNet |
| hcv0002 | CORE | HCV genotype 1b | | | APOA2 | | HCVpro |
| hcv0003 | CORE | HCV genotype 1b | | | TP53 | | HCVpro |
| hcv0004 | CORE | HCV genotype 1a | H77 | NP_751919 | LTBR | | VirHostNet |
| hcv0005 | CORE | HCV genotype 1a | H77 | NP_751919 | HNRNPK | | BIND |
| hcv0006 | CORE | HCV genotype 1a | H77 | NP_751919 | CORE | NP_751919 | BIND |
| hcv0007 | CORE | HCV genotype 1a | H77 | NP_751919 | DDX3X | | VirHostNet |
| hcv0008 | CORE | HCV genotype 1a | H77 | NP_751919 | DDX3X | | BIND |
| hcv0009 | CORE | HCV genotype 1a | | AAA45676 | DDX3X | | BIND |
| hcv0010 | CORE | HCV genotype 1a | H77 | NP_751919 | APOA2 | | VirHostNet |
| hcv0011 | CORE | HCV genotype 1a | H77 | NP_751919 | TP53 | | BIND |
| hcv0012 | CORE | HCV genotype 1b | | | YWHAB | | HCVpro |
| hcv0013 | CORE | HCV genotype 1b | | | YWHAZ | | HCVpro |
| hcv0014 | CORE | HCV genotype 1a | H77 | NP_751919 | CREB3 | | VirHostNet |
| hcv0015 | CORE | HCV genotype 1a | strain H | AAA45534 | CORE | AAA45534 | HCVpro |
| hcv0016 | CORE | HCV genotype 1b | Isolate S98 | BAC54273 | CDKN1A | | BIND |
| hcv0017 | CORE | HCV genotype 1b | | | TAF11 | | HCVpro |
| hcv0018 | CORE | HCV genotype 1b | | | TP53 | | HCVpro |
| hcv0019 | CORE | HCV genotype 1a | H77 | NP_751919 | C1QBP | | VirHostNet |
| hcv0020 | CORE | HCV genotype 1b | Isolate S98 | BAC54273 | TNFRSF1A | | BIND |
| hcv0021 | CORE | HCV genotype 1a | H77 | NP_751919 | FADD | | VirHostNet |
| hcv0022 | CORE | Genotype 1b | Korean isolate | BAA01943.1 | TNF | | HCVpro |
| hcv0023 | CORE | Genotype 1b | Korean isolate | BAA01943.1 | TRADD | | HCVpro |
| hcv0024 | CORE | Genotype 1b | Korean isolate | BAA01943.1 | TRAF2 | | HCVpro |
| hcv0025 | CORE | HCV genotype 1a | H77 | NP_751919 | RXRA | | VirHostNet |
| hcv0026 | CORE | HCV genotype 1a | H77 | NP_751919 | STAT3 | | VirHostNet |
| hcv0027 | CORE | HCV genotype 1a | H77 | NP_671491 | TSN | | BIND |
| hcv0028 | CORE | HCV genotype 1a | H77 | NP_751919 | FUNDC2 | | VirHostNet |

| hcv0029 | CORE | HCV genotype 1a | | | TP73 | | HCVpro |
|---------|------|------------------|---|---|------|---|--------|
| hcv0030 | CORE | HCV genotype 1a | | | TP53 | | HCVpro |
| hcv0031 | CORE | HCV genotype 1b | | | JAK1 | | HCVpro |
| hcv0032 | CORE | HCV genotype 1b | | | JAK2 | | HCVpro |
| hcv0033 | CORE | HCV genotype 1a | | | FAS | | HCVpro |
| hcv0034 | CORE | HCV genotype 1a | H77 | NP_751919 | CCNH | | VirHostNet |
| hcv0035 | CORE | HCV genotype 1a | H77 | NP_751919 | TBP | | VirHostNet |
| hcv0036 | CORE | HCV genotype 1b | | AAR06173.1 | TP53BP2 | | BIND |
| hcv0037 | CORE | HCV genotype 1a | H77 | NP_751919 | SMAD3 | | VirHostNet |
| hcv0038 | CORE | HCV genotype 1b | L02 | AB077729.1 | EP300 | | HCVpro |
| hcv0039 | CORE | HCV genotype 1b | L02 | AB077729.1 | CREBBP | | HCVpro |
| hcv0040 | CORE | | | | TLR2 | | Not assigned |
| hcv0041 | CORE | HCV genotype 1a | H77 | NP_751919 | HLA-A | | VirHostNet |
| hcv0042 | CORE | HCV genotype 1a | H77 | NP_751919 | HLA-E | | VirHostNet |
| hcv0043 | CORE | HCV genotype 1a | H77 | NP_751919 | STAT1 | | VirHostNet |
| hcv0044 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | ACP1 | | HCVpro |
| hcv0045 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | CFL1 | | HCVpro |
| hcv0046 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | FKBP7 | | HCVpro |
| hcv0047 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | HSPD1 | | HCVpro |
| hcv0048 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | KRT18 | | HCVpro |
| hcv0049 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | KRT19 | | HCVpro |
| hcv0050 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | KRT8 | | HCVpro |
| hcv0051 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | SLC22A7 | | HCVpro |
| hcv0052 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | TATDN1 | | HCVpro |
| hcv0053 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | GLRX3 | | HCVpro |
| hcv0054 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | VIM | | HCVpro |
| hcv0055 | CORE | HCV genotype 1a | H77 | NP_751919 | NPM1 | | VirHostNet |
| hcv0056 | CORE | HCV genotype 1a | H77 | NP_751919 | YY1 | | VirHostNet |
| hcv0057 | CORE | HCV genotype 1a | H77 | NP_751919 | PML | | VirHostNet |
| hcv0058 | CORE | HCV genotype 1a | strain H77 | NP_751919 | DICER1 | | VirHostNet |
| hcv0059 | CORE | HCV genotype 1a | | | PSME3 | | HCVpro |
| hcv0060 | CORE | HCV genotype 1a | | | DICER1 | | HCVpro |
| hcv0061 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | AGRN | | HCVpro |
| hcv0062 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | BCAR1 | | HCVpro |

| hcv0063 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | CD68 | | HCVpro |
|---------|------|-----------------|--------------|----------|------|--|--------|
| hcv0064 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | COL4A2 | | HCVpro |
| hcv0065 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | DDX3Y | | HCVpro |
| hcv0066 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | EGFL7 | | HCVpro |
| hcv0067 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | FBLN2 | | HCVpro |
| hcv0068 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | FBLN5 | | HCVpro |
| hcv0069 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | GAPDH | | HCVpro |
| hcv0070 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | GRN | | HCVpro |
| hcv0071 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | HIVEP2 | | HCVpro |
| hcv0072 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | HOXD8 | | HCVpro |
| hcv0073 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | LPXN | | HCVpro |
| hcv0074 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | LRRTM1 | | HCVpro |
| hcv0075 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | LTBP4 | | HCVpro |
| hcv0076 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | MAGED1 | | HCVpro |
| hcv0077 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | MEGF6 | | HCVpro |
| hcv0078 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | MMRN2 | | HCVpro |
| hcv0079 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | NR4A1 | | HCVpro |
| hcv0080 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | PABPN1 | | HCVpro |
| hcv0081 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | PAK4 | | HCVpro |
| hcv0082 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | PLSCR1 | | HCVpro |
| hcv0083 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | RNF31 | | HCVpro |
| hcv0084 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | SETD2 | | HCVpro |
| hcv0085 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | SLC31A2 | | HCVpro |
| hcv0086 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | VWF | | HCVpro |
| hcv0087 | CORE | HCV genotype 1b | isolate con1 | AJ238799 | ZNF271 | | HCVpro |
| hcv0088 | CORE | HCV genotype 1b | | | YWHAE | | HCVpro |
| hcv0089 | CORE | HCV genotype 1a | H77 | NP_751919 | NS5A | NP_751927 | BIND |
| hcv0090 | CORE | HCV genotype 1a | H77 | NP_751919 | NS5B | NP_751928 | VirHostNet |
| hcv0091 | CORE | HCV genotype 1a | H77 | NP_751919 | RSF1 | | VirHostNet |
| hcv0092 | CORE | HCV genotype 1a | H77 | NP_751919 | E1 | NP_751920 | BIND |
| hcv0093 | CORE | HCV genotype 1a | H77 | NP_751919 | SP110 | | VirHostNet |
| hcv0094 | CORE | HCV genotype 1a | H77 | NP_751919 | KPNA1 | | VirHostNet |
| hcv0095 | CORE | HCV genotype 1b | pCV-J4L8 | AAC15730 | DDX5 | | HCVpro |
| hcv0096 | CORE | HCV genotype 1b | Singapore strain | | NS5A | | HCVpro |

| hcv0097 | CORE | HCV genotype 1a | strain H77 | NP_751919 | EIF2AK2 | | VirHostNet |
|---------|------|-----------------|------------|-----------|---------|---|-----------|
| hcv0098 | CORE | HCV genotype 1a | strain H77 | NP_751919 | PPARA | | VirHostNet |
| hcv0099 | CORE | HCV genotype 1a | strain H77 | NP_751919 | ACY3 | | VirHostNet |
| hcv0100 | CORE | HCV genotype 1a | H77 | NP_751919 | MCL1 | | VirHostNet |
| hcv0101 | CORE | HCV genotype 1a | H77 | NP_751919 | CORE | NP_751919 | VirHostNet |
| hcv0102 | CORE | HCV genotype 1a | H77 | NP_751919 | CORE | | VirHostNet |
| hcv0103 | E1 | HCV genotype 1a | H77 | NP_751920 | E2 | NP_751921 | VirHostNet |
| hcv0104 | E1 | HCV genotype 1a | H77 | NP_751920 | CORE | NP_751919 | VirHostNet |
| hcv0105 | E1 | HCV genotype 1a | H77 | NP_751920 | LTF | | VirHostNet |
| hcv0106 | E1 | HCV genotype 1a | H77 | NP_751920 | CALR | | VirHostNet |
| hcv0107 | E1 | HCV genotype 1a | H77 | NP_751920 | CANX | | VirHostNet |
| hcv0108 | E1 | HCV genotype 1a | H77 | NP_751920 | HSPA5 | | VirHostNet |
| hcv0109 | E1 | HCV genotype 1a | Isolate H | P27958 | E2 | P27958 | BIND |
| hcv0110 | E1 | HCV genotype 1a | H77 | NP_751920 | E2 | NP_751921 | BIND |
| hcv0111 | E1 | HCV genotype 1a | H77 | AF009606 | CD209 | | HCVpro |
| hcv0112 | E1 | HCV genotype 1a | H77 | AF009606 | CLEC4M | | HCVpro |
| hcv0113 | E1 | HCV genotype 1a | H77 | NP_671491 | E2 | | BIND |
| hcv0114 | E1 | HCV genotype 1b | | | CD209 | | HCVpro |
| hcv0115 | E1 | HCV genotype 1b | isolate con1 | AJ238799 | JUN | | HCVpro |
| hcv0116 | E1 | HCV genotype 1b | isolate con1 | AJ238799 | NR4A1 | | HCVpro |
| hcv0117 | E1 | HCV genotype 1b | isolate con1 | AJ238799 | PFN1 | | HCVpro |
| hcv0118 | E1 | HCV genotype 1b | isolate con1 | AJ238799 | SETD2 | | HCVpro |
| hcv0119 | E1 | HCV genotype 1b | isolate con1 | AJ238799 | TMSB4X | | HCVpro |
| hcv0120 | E1 | HCV genotype 1a | strain H | AAA45534 | NS5A | AAA45534 | HCVpro |
| hcv0121 | E1 | HCV genotype 1a | H77 | NP_751920 | E2 | NP_751921 | VirHostNet |
| hcv0122 | E2 | HCV genotype 1a | H77 | NP_751921 | LTF | | VirHostNet |
| hcv0123 | E2 | HCV genotype 1a | H77 | NP_751921 | CALR | | VirHostNet |
| hcv0124 | E2 | HCV genotype 1a | H77 | NP_751921 | CANX | | VirHostNet |
| hcv0125 | E2 | HCV genotype 1a | H77 | NP_751921 | HSPA5 | | VirHostNet |
| hcv0126 | E2 | HCV genotype 1a | H77 | NP_751921 | EIF2AK2 | | VirHostNet |
| hcv0127 | E2 | HCV genotype 1a | strain H | AAA45534 | NS3 | AAA45534 | HCVpro |
| hcv0128 | E2 | HCV genotype 1a | H77 | NP_671491 | CD81 | | BIND |
| hcv0129 | E2 | HCV genotype 1a | isolate H | P27958 | CD81 | | BIND |
| hcv0130 | E2 | HCV genotype 1a | H77 | NP_751921 | SCARB1 | | VirHostNet |

| hcv0131 | E2 | HCV genotype 1b | | | CD81 | | HCVpro |
|---------|-----|-----------------|-------------|-----------|-----------|----------|------------|
| hcv0132 | E2 | HCV genotype 1b | | | LTF | | HCVpro |
| hcv0133 | E2 | HCV genotype 1b | | | TF | | HCVpro |
| hcv0134 | E2 | HCV genotype 1a | H77 | NP_751921 | CD81 | | BIND |
| hcv0135 | E2 | HCV genotype 1a | H77 | NP_751921 | EIF2AK3 | | VirHostNet |
| hcv0136 | E2 | HCV genotype 1a | H77 | AF009606 | CD209 | | HCVpro |
| hcv0137 | E2 | HCV genotype 1a | H77 | AF009606 | CLEC4M | | HCVpro |
| hcv0138 | E2 | HCV genotype 1a | H77 | NP_751921 | SDC2 | | VirHostNet |
| hcv0139 | E2 | HCV genotype 1b | | | CD209 | | HCVpro |
| hcv0140 | E2 | HCV genotype 1b | isolate con1 | AJ238799 | HOXD8 | | HCVpro |
| hcv0141 | E2 | HCV genotype 1b | isolate con1 | AJ238799 | ITGB1 | | HCVpro |
| hcv0142 | E2 | HCV genotype 1b | isolate con1 | AJ238799 | FAM135A | | HCVpro |
| hcv0143 | E2 | HCV genotype 1b | isolate con1 | AJ238799 | NR4A1 | | HCVpro |
| hcv0144 | E2 | HCV genotype 1b | isolate con1 | AJ238799 | PSMA6 | | HCVpro |
| hcv0145 | E2 | HCV genotype 1b | isolate con1 | AJ238799 | SETD2 | | HCVpro |
| hcv0146 | E2 | HCV genotype 1b | isolate con1 | AJ238799 | SMEK2 | | HCVpro |
| hcv0147 | F | HCV genotype 1b | | | C14orf135 | | HCVpro |
| hcv0148 | F | HCV genotype 1b | | | ZNF83 | | HCVpro |
| hcv0149 | F | HCV genotype 1a | strain H77 | NP_803170 | PFDN2 | | VirHostNet |
| hcv0150 | F | HCV genotype 1a | strain H77 | NP_803170 | PFDN5 | | VirHostNet |
| hcv0151 | F | HCV genotype 1b | | | AGT | | HCVpro |
| hcv0152 | F | HCV genotype 1b | | | AZGP1 | | HCVpro |
| hcv0153 | F | HCV genotype 1b | | | CTSB | | HCVpro |
| hcv0154 | F | HCV genotype 1b | | | MPDU1 | | HCVpro |
| hcv0155 | F | HCV genotype 1b | | | RAB14 | | HCVpro |
| hcv0156 | F | HCV genotype 1b | | | SERPINC1 | | HCVpro |
| hcv0157 | F | HCV genotype 1b | | | ST3GAL1 | | HCVpro |
| hcv0158 | F | HCV genotype 1b | | | vitronectin | | HCVpro |
| hcv0159 | F | HCV genotype 1b | | | ZG16 | | HCVpro |
| hcv0160 | NS2 | HCV genotype 1a | strain H | AAA45534 | NS4A | AAA45534 | HCVpro |
| hcv0161 | NS2 | HCV genotype 1b | Strain J | | NS3 | | HCVpro |
| hcv0162 | NS2 | HCV genotype 1a | Strain H | | CIDEB | | BIND |
| hcv0163 | NS2 | HCV genotype 1a | H77 | AF009606 | NS4B | AF009606 | HCVpro |
| hcv0164 | NS2 | HCV genotype 1a | H77 | AF009606 | NS4A | AF009606 | HCVpro |

| hcv0165 | NS2 | HCV genotype 1a | H77 | AF009606 | NS5B | AF009606 | HCVpro |
|---------|-----|------------------|-----|----------|------|----------|--------|
| hcv0166 | NS2 | HCV genotype 1a | H77 | AF009606 | NS2 | AF009606 | HCVpro |
| hcv0167 | NS2 | HCV genotype 1a | H77 | AF009606 | NS3 | AF009606 | HCVpro |
| hcv0168 | NS2 | HCV genotype 1b | isolate con1 | AJ238799 | C7 | | HCVpro |
| hcv0169 | NS2 | HCV genotype 1b | isolate con1 | AJ238799 | FBLN5 | | HCVpro |
| hcv0170 | NS2 | HCV genotype 1b | isolate con1 | AJ238799 | HOXD8 | | HCVpro |
| hcv0171 | NS2 | HCV genotype 1b | isolate con1 | AJ238799 | NR4A1 | | HCVpro |
| hcv0172 | NS2 | HCV genotype 1b | isolate con1 | AJ238799 | POU3F2 | | HCVpro |
| hcv0173 | NS2 | HCV genotype 1b | isolate con1 | AJ238799 | SETD2 | | HCVpro |
| hcv0174 | NS2 | HCV genotype 1b | isolate con1 | AJ238799 | TRIM27 | | HCVpro |
| hcv0175 | NS2 | HCV genotype 1a | H77 | AF009606 | NS5A | AF009606 | HCVpro |
| hcv0176 | NS3 | | | | NS4A | | Not assigned |
| hcv0177 | NS3 | HCV genotype 1b | | AAA72945 | MBP | | VirHostNet |
| hcv0178 | NS3 | HCV genotype 1b | | AAA72945 | PRM1 | | VirHostNet |
| hcv0179 | NS3 | HCV genotype 1b | | AAA72945 | HIST4H4 | | VirHostNet |
| hcv0180 | NS3 | HCV genotype 1b | | AAA72945 | HIST3H2BB | | VirHostNet |
| hcv0181 | NS3 | HCV genotype 1b | | AAA72945 | PRKACA | | VirHostNet |
| hcv0182 | NS3 | HCV genotype 1b | HCV isolate BK | P26663 | PRKACA | | BIND |
| hcv0183 | NS3 | HCV genotype 1a | H77 | NP_671491 | NS3 | NP_671491 | BIND |
| hcv0184 | NS3 | HCV genotype 1a | H77 | AAB66324 | TP53 | | VirHostNet |
| hcv0185 | NS3 | HCV genotype 1a | H77 | AAB66324 | NS4A | AAB66324 | VirHostNet |
| hcv0186 | NS3 | HCV genotype 1b | HCV isolate BK | P26663 | NS4A | P26663 | BIND |
| hcv0187 | NS3 | HCV genotype 1a | H77 | AAB66324 | HIST3H2BB | | VirHostNet |
| hcv0188 | NS3 | HCV genotype 1a | H77 | AAB66324 | HIST4H4 | | VirHostNet |
| hcv0189 | NS3 | HCV genotype 1b | Isolate G01 | | SERPINF2 | | HCVpro |
| hcv0190 | NS3 | HCV genotype 1b | Isolate G01 | | SERPING1 | | HCVpro |
| hcv0191 | NS3 | HCV genotype 1a | strain H | AAA45534 | NS4A | AAA45534 | HCVpro |
| hcv0192 | NS3 | HCV genotype 1a | H77 | NP_803144 | PRMT5 | | BIND |
| hcv0193 | NS3 | HCV genotype 1b | Con1 | CAB46677 | PRMT1 | | BIND |
| hcv0194 | NS3 | HCV genotype 1a | H77 | AAB66324 | NS3 | | VirHostNet |
| hcv0195 | NS3 | HCV genotype 1a | H77 | AF009606 | NS4B | AF009606 | HCVpro |
| hcv0196 | NS3 | HCV genotype 1a | H77 | AF009606 | NS4A | AF009606 | HCVpro |
| hcv0197 | NS3 | HCV genotype 1a | H77 | AF009606 | NS3 | AF009606 | HCVpro |
| hcv0198 | NS3 | HCV genotype 1a | | | IRF3 | | HCVpro |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| hcv0199 | NS3 | HCV genotype 1a | | AAA45676.1 | SNRPD1 | | BIND |
| hcv0200 | NS3 | | | | NS4A | | Not assigned |
| hcv0201 | NS3 | | | | NS4A | | Not assigned |
| hcv0202 | NS3 | HCV genotype 1b | Isolate HCV-S1 | AAL00900 | PSMB8 | | BIND |
| hcv0203 | NS3 | HCV genotype 1a | H77 | NP_803144 | SMAD3 | | VirHostNet |
| hcv0204 | NS3 | | | | NS4A | | Not assigned |
| hcv0205 | NS3 | | | | TLR2 | | Not assigned |
| hcv0206 | NS3 | HCV genotype 1a | H77 | NP_803144 | TICAM1 | | VirHostNet |
| hcv0207 | NS3 | HCV genotype 1a | H77 | NP_803144 | PTBP2 | | VirHostNet |
| hcv0208 | NS3 | HCV genotype 1a | H77 | NP_803144 | IKBKE | | VirHostNet |
| hcv0209 | NS3 | HCV genotype 1a | H77 | NP_803144 | TBK1 | | VirHostNet |
| hcv0210 | NS3 | HCV genotype 1a | H77 | NP_803144 | ERC1 | | VirHostNet |
| hcv0211 | NS3 | | | | NS4A | | Not assigned |
| hcv0212 | NS3 | | | | NS4A | | Not assigned |
| hcv0213 | NS3 | | | | NS4A | | Not assigned |
| hcv0214 | NS3 | | | | NS4A | | Not assigned |
| hcv0215 | NS3 | HCV genotype 1a | | | VISA | | HCVpro |
| hcv0216 | NS3 | HCV genotype 1b | | | NS5B | | HCVpro |
| hcv0217 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ACTN1 | | HCVpro |
| hcv0218 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ACTN2 | | HCVpro |
| hcv0219 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | AEBP1 | | HCVpro |
| hcv0220 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ANKRD12 | | HCVpro |
| hcv0221 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ANKRD28 | | HCVpro |
| hcv0222 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ARFIP2 | | HCVpro |
| hcv0223 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ARHGEF6 | | HCVpro |
| hcv0224 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ARNT | | HCVpro |
| hcv0225 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ARS2 | | HCVpro |
| hcv0226 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ASXL1 | | HCVpro |
| hcv0227 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | B2M | | HCVpro |
| hcv0228 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | BCAN | | HCVpro |
| hcv0229 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | BCKDK | | HCVpro |
| hcv0230 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | BCL2A1 | | HCVpro |
| hcv0231 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | BCL6 | | HCVpro |
| hcv0232 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | BZRAP1 | | HCVpro |

| hcv0233 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | C10orf18 | | HCVpro |
|---|---|---|---|---|---|---|---|
| hcv0234 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | C10orf6 | | HCVpro |
| hcv0235 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | C12orf41 | | HCVpro |
| hcv0236 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | INF2 | | HCVpro |
| hcv0237 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | C16orf7 | | HCVpro |
| hcv0238 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | BEND5 | | HCVpro |
| hcv0239 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | C1orf94 | | HCVpro |
| hcv0240 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | C9orf30 | | HCVpro |
| hcv0241 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CALCOCO2 | | HCVpro |
| hcv0242 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CBY1 | | HCVpro |
| hcv0243 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CCDC21 | | HCVpro |
| hcv0244 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CCDC37 | | HCVpro |
| hcv0245 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CCDC52 | | HCVpro |
| hcv0246 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CCDC66 | | HCVpro |
| hcv0247 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | INO80E | | HCVpro |
| hcv0248 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CCHCR1 | | HCVpro |
| hcv0249 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CD5L | | HCVpro |
| hcv0250 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CDC23 | | HCVpro |
| hcv0251 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CELSR2 | | HCVpro |
| hcv0252 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CEP152 | | HCVpro |
| hcv0253 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CEP192 | | HCVpro |
| hcv0254 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CFP | | HCVpro |
| hcv0255 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CHPF | | HCVpro |
| hcv0256 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CORO1B | | HCVpro |
| hcv0257 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CSNK2B | | HCVpro |
| hcv0258 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | CTGF | | HCVpro |
| hcv0259 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ALG13 | | HCVpro |
| hcv0260 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | DEAF1 | | HCVpro |
| hcv0261 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | DES | | HCVpro |
| hcv0262 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | DLAT | | HCVpro |
| hcv0263 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | DOCK7 | | HCVpro |
| hcv0264 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | DPF1 | | HCVpro |
| hcv0265 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | DPP7 | | HCVpro |
| hcv0266 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | EEF1A1 | | HCVpro |

| hcv0267 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | EFEMP1 | | HCVpro |
|---------|-----|-----------------|--------------|----------|---------|--|--------|
| hcv0268 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | EFEMP2 | | HCVpro |
| hcv0269 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | EIF1 | | HCVpro |
| hcv0270 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | EIF4ENIF1 | | HCVpro |
| hcv0271 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FAM120B | | HCVpro |
| hcv0272 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FAM65A | | HCVpro |
| hcv0273 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FBF1 | | HCVpro |
| hcv0274 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FBLN1 | | HCVpro |
| hcv0275 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FBLN2 | | HCVpro |
| hcv0276 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FBLN5 | | HCVpro |
| hcv0277 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FBN1 | | HCVpro |
| hcv0278 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FBN3 | | HCVpro |
| hcv0279 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FES | | HCVpro |
| hcv0280 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FIGNL1 | | HCVpro |
| hcv0281 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FLAD1 | | HCVpro |
| hcv0282 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | C19orf66 | | HCVpro |
| hcv0283 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FN1 | | HCVpro |
| hcv0284 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FRMPD4 | | HCVpro |
| hcv0285 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FRS3 | | HCVpro |
| hcv0286 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FTH1 | | HCVpro |
| hcv0287 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | FUCA2 | | HCVpro |
| hcv0288 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | GAA | | HCVpro |
| hcv0289 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | GBP2 | | HCVpro |
| hcv0290 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | GFAP | | HCVpro |
| hcv0291 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | GNB2 | | HCVpro |
| hcv0292 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | GON4L | | HCVpro |
| hcv0293 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | HIVEP2 | | HCVpro |
| hcv0294 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | HNRNPK | | HCVpro |
| hcv0295 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | HOMER3 | | HCVpro |
| hcv0296 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | IQWD1 | | HCVpro |
| hcv0297 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ITGB4 | | HCVpro |
| hcv0298 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | JAG2 | | HCVpro |
| hcv0299 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | JUN | | HCVpro |
| hcv0300 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | KHDRBS1 | | HCVpro |

| hcv0301 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | KIAA1549 | | HCVpro |
|---------|-----|-----------------|--------------|----------|----------|---|--------|
| hcv0302 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | KIF17 | | HCVpro |
| hcv0303 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | KIF7 | | HCVpro |
| hcv0304 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | KPNA1 | | HCVpro |
| hcv0305 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | L3MBTL3 | | HCVpro |
| hcv0306 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | LAMA5 | | HCVpro |
| hcv0307 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | LAMB2 | | HCVpro |
| hcv0308 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | LAMC3 | | HCVpro |
| hcv0309 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | LDB1 | | HCVpro |
| hcv0310 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | LRRC7 | | HCVpro |
| hcv0311 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | LRRCC1 | | HCVpro |
| hcv0312 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | LTBP4 | | HCVpro |
| hcv0313 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | LZTS2 | | HCVpro |
| hcv0314 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | MAGED1 | | HCVpro |
| hcv0315 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | MAPK7 | | HCVpro |
| hcv0316 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | MEGF8 | | HCVpro |
| hcv0317 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | MLLT4 | | HCVpro |
| hcv0318 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | MLXIP | | HCVpro |
| hcv0319 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | MORC4 | | HCVpro |
| hcv0320 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | MORF4L1 | | HCVpro |
| hcv0321 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | MVP | | HCVpro |
| hcv0322 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NAP1L1 | | HCVpro |
| hcv0323 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NAP1L2 | | HCVpro |
| hcv0324 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NCAN | | HCVpro |
| hcv0325 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NDC80 | | HCVpro |
| hcv0326 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NEFL | | HCVpro |
| hcv0327 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NEFM | | HCVpro |
| hcv0328 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NELL1 | | HCVpro |
| hcv0329 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NELL2 | | HCVpro |
| hcv0330 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NID1 | | HCVpro |
| hcv0331 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NID2 | | HCVpro |
| hcv0332 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NOTCH1 | | HCVpro |
| hcv0333 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | N-PAC | | HCVpro |
| hcv0334 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | NUP62 | | HCVpro |

| hcv0335 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | OBSCN | | HCVpro |
|---------|-----|-----------------|--------------|----------|-------|---|--------|
| hcv0336 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PARP4 | | HCVpro |
| hcv0337 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PCYT2 | | HCVpro |
| hcv0338 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PDE4DIP | | HCVpro |
| hcv0339 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PDLIM5 | | HCVpro |
| hcv0340 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PICK1 | | HCVpro |
| hcv0341 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PKNOX1 | | HCVpro |
| hcv0342 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PLEKHG4 | | HCVpro |
| hcv0343 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PNPLA8 | | HCVpro |
| hcv0344 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PRRC1 | | HCVpro |
| hcv0345 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PSMB9 | | HCVpro |
| hcv0346 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PSME3 | | HCVpro |
| hcv0347 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | PTPRN2 | | HCVpro |
| hcv0348 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RABEP1 | | HCVpro |
| hcv0349 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RAI14 | | HCVpro |
| hcv0350 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RASAL2 | | HCVpro |
| hcv0351 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RBM4 | | HCVpro |
| hcv0352 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RCN3 | | HCVpro |
| hcv0353 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RGNEF | | HCVpro |
| hcv0354 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RICS | | HCVpro |
| hcv0355 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RINT1 | | HCVpro |
| hcv0356 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RNF31 | | HCVpro |
| hcv0357 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ROGDI | | HCVpro |
| hcv0358 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | KIAA2022 | | HCVpro |
| hcv0359 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RUSC2 | | HCVpro |
| hcv0360 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SBF1 | | HCVpro |
| hcv0361 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SDCCAG8 | | HCVpro |
| hcv0362 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SECISBP2 | | HCVpro |
| hcv0363 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | 10-Sep | | HCVpro |
| hcv0364 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SERTAD1 | | HCVpro |
| hcv0365 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SESTD1 | | HCVpro |
| hcv0366 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SF3B2 | | HCVpro |
| hcv0367 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SIAH1 | | HCVpro |
| hcv0368 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SLIT1 | | HCVpro |

| hcv0369 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SLIT2 | | HCVpro |
|---|---|---|---|---|---|---|---|
| hcv0370 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SLIT3 | | HCVpro |
| hcv0371 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SMURF2 | | HCVpro |
| hcv0372 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SNX4 | | HCVpro |
| hcv0373 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SPOCK3 | | HCVpro |
| hcv0374 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SPON1 | | HCVpro |
| hcv0375 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SRPX2 | | HCVpro |
| hcv0376 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SSX2IP | | HCVpro |
| hcv0377 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | STAB1 | | HCVpro |
| hcv0378 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | STAT3 | | HCVpro |
| hcv0379 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SVEP1 | | HCVpro |
| hcv0380 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SYNE1 | | HCVpro |
| hcv0381 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | SYNPO2 | | HCVpro |
| hcv0382 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TAF1 | | HCVpro |
| hcv0383 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TBC1D2B | | HCVpro |
| hcv0384 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TBXAS1 | | HCVpro |
| hcv0385 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TGFB1I1 | | HCVpro |
| hcv0386 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | THAP1 | | HCVpro |
| hcv0387 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TMEM63B | | HCVpro |
| hcv0388 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TRIM23 | | HCVpro |
| hcv0389 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TRIM27 | | HCVpro |
| hcv0390 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TRIO | | HCVpro |
| hcv0391 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TRIP11 | | HCVpro |
| hcv0392 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | TXNDC11 | | HCVpro |
| hcv0393 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | UBA3 | | HCVpro |
| hcv0394 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | USHBP1 | | HCVpro |
| hcv0395 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | UXT | | HCVpro |
| hcv0396 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | VCAN | | HCVpro |
| hcv0397 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | VIM | | HCVpro |
| hcv0398 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | VWF | | HCVpro |
| hcv0399 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | XAB2 | | HCVpro |
| hcv0400 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | XRN2 | | HCVpro |
| hcv0401 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | YY1AP1 | | HCVpro |
| hcv0402 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ZBTB1 | | HCVpro |

| hcv0403 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ZCCHC7 | | HCVpro |
|---|---|---|---|---|---|---|---|
| hcv0404 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ZHX3 | | HCVpro |
| hcv0405 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ZMYM2 | | HCVpro |
| hcv0406 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ZNF281 | | HCVpro |
| hcv0407 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ZNF410 | | HCVpro |
| hcv0408 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | ZZZ3 | | HCVpro |
| hcv0409 | NS3 | HCV genotype 1a | strain H | AAA45534 | NS5B | AAA45534 | HCVpro |
| hcv0410 | NS3 | HCV genotype 1a | H77 | NP_803144 | NS3 | NP_803144 | VirHostNet |
| hcv0411 | NS3 | HCV genotype 1b | | | NS5B | | HCVpro |
| hcv0412 | NS3 | HCV genotype 1b | | | NS4B | | HCVpro |
| hcv0413 | NS3 | HCV genotype 1a | H77 | AF009606 | NS5B | AF009606 | HCVpro |
| hcv0414 | NS3 | HCV genotype 1a | H77 | NP_803144 | CASP8 | | VirHostNet |
| hcv0415 | NS3 | HCV genotype 1b | Singapore strain | | NS3 | | HCVpro |
| hcv0416 | NS3 | HCV genotype 1a | H77 | NP_803144 | NS4A | NP_751925 | VirHostNet |
| hcv0417 | NS3 | HCV genotype 1b | isolate con1 | AJ238799 | RSPH3 | | HCVpro |
| hcv0418 | NS4A | HCV genotype 1a | H77 | NP_751925 | NS3 | NP_803144 | VirHostNet |
| hcv0419 | NS4A | HCV genotype 1b | Isolate Bk | 1JXP_D | NS3 | BAA09073 | BIND |
| hcv0420 | NS4A | HCV isolate 1 | Isolate Taiwanese | 1DY9_D | NS3 | BAA28515 | BIND |
| hcv0421 | NS4A | HCV genotype 1a | H77 | AAB66324 | NS3 | | VirHostNet |
| hcv0422 | NS4A | HCV genotype 1a | H77 | AF009606 | NS4A | AF009606 | HCVpro |
| hcv0423 | NS4A | HCV genotype 1a | H77 | AF009606 | NS4B | AF009606 | HCVpro |
| hcv0424 | NS4A | HCV genotype 1a | H77 | AF009606 | NS5B | AF009606 | HCVpro |
| hcv0425 | NS4A | HCV genotype 1b | isolate con1 | AJ238799 | CREB3 | | HCVpro |
| hcv0426 | NS4A | HCV genotype 1b | isolate con1 | AJ238799 | ELAC2 | | HCVpro |
| hcv0427 | NS4A | HCV genotype 1b | isolate con1 | AJ238799 | HOXD8 | | HCVpro |
| hcv0428 | NS4A | HCV genotype 1b | isolate con1 | AJ238799 | NR4A1 | | HCVpro |
| hcv0429 | NS4A | HCV genotype 1b | isolate con1 | AJ238799 | TRAF3IP3 | | HCVpro |
| hcv0430 | NS4A | HCV genotype 1b | isolate con1 | AJ238799 | UBQLN1 | | HCVpro |
| hcv0431 | NS4A | HCV genotype 1a | H77 | AF009606 | NS5A | AF009606 | HCVpro |
| hcv0432 | NS4A | HCV genotype 1a | H77 | NP_751925 | MT2A | | VirHostNet |
| hcv0433 | NS4A | HCV genotype 1a | H77 | NP_751925 | TUT1 | | VirHostNet |
| hcv0434 | NS4A | HCV genotype 1a | H77 | NP_751925 | MT-CO2 | | VirHostNet |
| hcv0435 | NS4A | HCV genotype 1a | strain H77 | NP_751925 | EEF1A1 | | VirHostNet |
| hcv0436 | NS4A | HCV genotype 1a | strain H77 | NP_751925 | CAMLG | | VirHostNet |

| hcv0437 | NS4A | HCV genotype 1a | H77 | AAB66324 | NS5A | AAB66324 | VirHostNet |
|---------|------|-----------------|-----|----------|------|----------|------------|
| hcv0438 | NS4B | HCV genotype 1a | H77 | AAB66324 | NS4A | AAB66324 | VirHostNet |
| hcv0439 | NS4B | HCV genotype 1b | | | ATF6B | | HCVpro |
| hcv0440 | NS4B | HCV genotype 1b | | | ATF6A | | HCVpro |
| hcv0441 | NS4B | HCV genotype 1a | H77 | AF009606 | NS5B | AF009606 | HCVpro |
| hcv0442 | NS4B | HCV genotype 1a | H77 | AF009606 | NS4B | AF009606 | HCVpro |
| hcv0443 | NS4B | HCV genotype 1b | | | NS5B | | HCVpro |
| hcv0444 | NS4B | HCV genotype 1b | | | TNXB | | HCVpro |
| hcv0445 | NS4B | HCV genotype 1a | H77 | AF009606 | NS5A | AF009606 | HCVpro |
| hcv0446 | NS4B | HCV genotype 1b | | | NS5A | | HCVpro |
| hcv0447 | NS4B | HCV genotype 1a | H77 | NP_751926 | RTN3 | | VirHostNet |
| hcv0448 | NS4B | HCV genotype 1a | H77 | NP_751926 | RBP4 | | VirHostNet |
| hcv0449 | NS4B | HCV genotype 1a | H77 | NP_751926 | NDUFV3 | | VirHostNet |
| hcv0450 | NS4B | HCV genotype 1a | H77 | NP_751926 | FGG | | VirHostNet |
| hcv0451 | NS4B | HCV genotype 1a | H77 | NP_751926 | MT-CO3 | | VirHostNet |
| hcv0452 | NS4B | HCV genotype 1a | strain H77 | NP_751926 | RAB5A | | VirHostNet |
| hcv0453 | NS5A | HCV genotype 1a | H77 | NP_751927 | EIF2AK2 | | BIND |
| hcv0454 | NS5A | HCV genotype 1b | | | CSNK2A1 | | HCVpro |
| hcv0455 | NS5A | HCV genotype 1b | Con1 | CAB46677 | GRB2 | | BIND |
| hcv0456 | NS5A | HCV genotype 1a | H77 | NP_751927 | VAPA | | VirHostNet |
| hcv0457 | NS5A | HCV genotype 1a | H77 | NP_751927 | SRCAP | | BIND |
| hcv0458 | NS5A | HCV genotype 1a | H77 | NP_751927 | IPO5 | | BIND |
| hcv0459 | NS5A | HCV genotype 1a | H77 | NP_751927 | TP53 | | BIND |
| hcv0460 | NS5A | HCV genotype 1a | strain H | AAA45534 | CDC2 | | HCVpro |
| hcv0461 | NS5A | HCV genotype 1a | strain H | AAA45534 | CDK2 | | HCVpro |
| hcv0462 | NS5A | HCV genotype 1a | H77 | NP_751927 | APOA1 | | BIND |
| hcv0463 | NS5A | HCV genotype 1b | | | TP53 | | HCVpro |
| hcv0464 | NS5A | HCV genotype 1b | | | TAF9 | | HCVpro |
| hcv0465 | NS5A | HCV genotype 1b | | | PIK3R1 | | HCVpro |
| hcv0466 | NS5A | HCV genotype 1b | | | TP53 | | HCVpro |
| hcv0467 | NS5A | HCV genotype 1b | | | TBP | | HCVpro |
| hcv0468 | NS5A | HCV genotype 1b | isolate con1 | CAB46677 | BIN1 | | BIND |
| hcv0469 | NS5A | HCV genotype 1a | H77 | NP_751927 | PITX1 | | VirHostNet |
| hcv0470 | NS5A | HCV genotype 1a | H77 | AF009606 | NS5B | AF009606 | HCVpro |

| hcv0471 | NS5A | HCV genotype 1a | H77 | AF009606 | NS5A | AF009606 | HCVpro |
|---------|------|-----------------|-----|----------|------|----------|--------|
| hcv0472 | NS5A | HCV genotype 1a | H77 | NP_751927 | BAX | | VirHostNet |
| hcv0473 | NS5A | HCV genotype 1a | H77 | NP_671491.1 | SSB | | BIND |
| hcv0474 | NS5A | HCV genotype 1a | H77 | AF009606 | HCK | | HCVpro |
| hcv0475 | NS5A | HCV genotype 1a | H77 | AF009606 | LCK | | HCVpro |
| hcv0476 | NS5A | HCV genotype 1a | H77 | AF009606 | LYN | | HCVpro |
| hcv0477 | NS5A | HCV genotype 1a | H77 | AF009606 | FYN | | HCVpro |
| hcv0478 | NS5A | HCV genotype 1b | | | VAPA | | HCVpro |
| hcv0479 | NS5A | HCV genotype 1a | H77 | NP_751927 | OAS1 | | VirHostNet |
| hcv0480 | NS5A | HCV genotype 1a | H77 | NP_751927 | JAK1 | | BIND |
| hcv0481 | NS5A | HCV genotype 1b | Con1 | CAB46677 | APOE | | HCVpro |
| hcv0482 | NS5A | HCV genotype 1b | Con1 | CAB46677 | VAPA | | HCVpro |
| hcv0483 | NS5A | HCV genotype 1b | Con1 | CAB46677 | NS5A | | HCVpro |
| hcv0484 | NS5A | HCV genotype 1b | Con1 | CAB46677 | AHNAK | | VirHostNet |
| hcv0485 | NS5A | HCV genotype 1b | Con1 | CAB46677 | SFRP4 | | VirHostNet |
| hcv0486 | NS5A | HCV genotype 1b | Con1 | CAB46677 | CRABP1 | | VirHostNet |
| hcv0487 | NS5A | HCV genotype 1b | Con1 | CAB46677 | FTH1 | | VirHostNet |
| hcv0488 | NS5A | HCV genotype 1b | Con1 | CAB46677 | CCDC86 | | VirHostNet |
| hcv0489 | NS5A | HCV genotype 1b | Con1 | CAB46677 | TACSTD2 | | VirHostNet |
| hcv0490 | NS5A | HCV genotype 1b | Con1 | CAB46677 | PI4KA | | VirHostNet |
| hcv0491 | NS5A | HCV genotype 1b | Con1 | CAB46677 | PTMA | | VirHostNet |
| hcv0492 | NS5A | HCV genotype 1b | Con1 | CAB46677 | ARAP1 | | VirHostNet |
| hcv0493 | NS5A | HCV genotype 1b | Con1 | CAB46677 | NDRG1 | | VirHostNet |
| hcv0494 | NS5A | HCV genotype 1b | Con1 | CAB46677 | MGP | | VirHostNet |
| hcv0495 | NS5A | HCV genotype 1b | Con1 | CAB46677 | CEP57 | | VirHostNet |
| hcv0496 | NS5A | HCV genotype 1b | Con1 | CAB46677 | C9orf6 | | VirHostNet |
| hcv0497 | NS5A | HCV genotype 1b | | | FBXL2 | | HCVpro |
| hcv0498 | NS5A | HCV genotype 1b | strain MD13 | AAF65944.1 | NS5A | | BIND |
| hcv0499 | NS5A | HCV genotype 1a | H77 | NP_751927 | CSK | | VirHostNet |
| hcv0500 | NS5A | HCV genotype 1a | H77 | NP_751927 | SRC | | VirHostNet |
| hcv0501 | NS5A | HCV genotype 1a | H77 | NP_751927 | APOB | | VirHostNet |
| hcv0502 | NS5A | HCV genotype 1b | Strain J1 | D89815 | VAPB | | HCVpro |
| hcv0503 | NS5A | HCV genotype 1a | H77 | NP_751927 | RAF1 | | VirHostNet |
| hcv0504 | NS5A | HCV genotype 1b | Korean isolate | | TGFBR1 | | HCVpro |

| hcv0505 | NS5A | HCV genotype 1a | strain H77 | NP_751927 | PPP2R4 | | VirHostNet |
|---|---|---|---|---|---|---|---|
| hcv0506 | NS5A | HCV genotype 1a | strain H77 | NP_751927 | BIN1 | | VirHostNet |
| hcv0507 | NS5A | HCV genotype 1b | | | TRAF2 | | HCVpro |
| hcv0508 | NS5A | HCV genotype 1a | strain H77 | NP_751927 | FKBP8 | | VirHostNet |
| hcv0509 | NS5A | HCV genotype 1a | strain H77C | | FKBP8 | | HCVpro |
| hcv0510 | NS5A | HCV genotype 1a | strain H77C | | HSP90AA1 | | HCVpro |
| hcv0511 | NS5A | HCV genotype 1a | strain H77 | NP_751927 | PTPLAD1 | | VirHostNet |
| hcv0512 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | ACLY | | HCVpro |
| hcv0513 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | ARFIP1 | | HCVpro |
| hcv0514 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | AXIN1 | | HCVpro |
| hcv0515 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | BEND7 | | HCVpro |
| hcv0516 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | CADPS | | HCVpro |
| hcv0517 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | CADPS2 | | HCVpro |
| hcv0518 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | CEP120 | | HCVpro |
| hcv0519 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | CENPC1 | | HCVpro |
| hcv0520 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | CEP250 | | HCVpro |
| hcv0521 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | CEP63 | | HCVpro |
| hcv0522 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | DNAJA3 | | HCVpro |
| hcv0523 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | EFEMP1 | | HCVpro |
| hcv0524 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | FHL2 | | HCVpro |
| hcv0525 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | GOLGA2 | | HCVpro |
| hcv0526 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | GPS2 | | HCVpro |
| hcv0527 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | IGLL1 | | HCVpro |
| hcv0528 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | ITGAL | | HCVpro |
| hcv0529 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | LIMS2 | | HCVpro |
| hcv0530 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | MOBK1B | | HCVpro |
| hcv0531 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | NAP1L1 | | HCVpro |
| hcv0532 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | NAP1L2 | | HCVpro |
| hcv0533 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | NFE2 | | HCVpro |
| hcv0534 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | NUCB1 | | HCVpro |
| hcv0535 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | PARVG | | HCVpro |
| hcv0536 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | PMVK | | HCVpro |
| hcv0537 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | PPP1R13L | | HCVpro |
| hcv0538 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | PSMB9 | | HCVpro |

| hcv0539 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | RPL18A | | HCVpro |
|---------|------|-----------------|--------------|----------|--------|--|--------|
| hcv0540 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | RRBP1 | | HCVpro |
| hcv0541 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | SHARPIN | | HCVpro |
| hcv0542 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | SMYD3 | | HCVpro |
| hcv0543 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | SORBS2 | | HCVpro |
| hcv0544 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | SORBS3 | | HCVpro |
| hcv0545 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | THBS1 | | HCVpro |
| hcv0546 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | TMF1 | | HCVpro |
| hcv0547 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | TP53BP2 | | HCVpro |
| hcv0548 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | TRIOBP | | HCVpro |
| hcv0549 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | TXNDC11 | | HCVpro |
| hcv0550 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | UBASH3A | | HCVpro |
| hcv0551 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | USP19 | | HCVpro |
| hcv0552 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | VPS52 | | HCVpro |
| hcv0553 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | GIN1 | | HCVpro |
| hcv0554 | NS5A | HCV genotype 1b | isolate con1 | AJ238799 | ZNF646 | | HCVpro |
| hcv0555 | NS5A | HCV genotype 1a | H77 | NP_751927 | NS5B | NP_751928 | BIND |
| hcv0556 | NS5A | HCV genotype 1a | H77 | NP_751927 | TRAF2 | | BIND |
| hcv0557 | NS5A | HCV genotype 1a | H77 | NP_751927 | TRADD | | VirHostNet |
| hcv0558 | NS5A | HCV genotype 1b | | | GAB1 | | HCVpro |
| hcv0559 | NS5A | HCV genotype 1a | H77 | AF009606 | NS3 | AF009606 | HCVpro |
| hcv0560 | NS5A | HCV genotype 1a | H77 | NP_751927 | PIK3CB | | VirHostNet |
| hcv0561 | NS5A | HCV genotype 1b | | | FBXL20 | | HCVpro |
| hcv0562 | NS5A | HCV genotype 1a | strain H77 | NP_751927 | STAT1 | | VirHostNet |
| hcv0563 | NS5A | HCV genotype 1a | strain H77 | NP_751927 | MYD88 | | VirHostNet |
| hcv0564 | NS5A | HCV genotype 2a | | | PDPK1 | | HCVpro |
| hcv0565 | NS5A | HCV genotype 2a | | | AKT1 | | HCVpro |
| hcv0566 | NS5A | HCV genotype 2a | | | VPS35 | | HCVpro |
| hcv0567 | NS5A | HCV genotype 2a | | | MAPK12 | | HCVpro |
| hcv0568 | NS5A | HCV genotype 2a | | | IPO4 | | HCVpro |
| hcv0569 | NS5A | HCV genotype 2a | | | GSK3B | | HCVpro |
| hcv0570 | NS5A | HCV genotype 2a | | | AHSA1 | | HCVpro |
| hcv0571 | NS5A | HCV genotype 2a | | | GSK3A | | HCVpro |
| hcv0572 | NS5A | HCV genotype 2a | | | CDK6 | | HCVpro |

| hcv0573 | NS5A | HCV genotype 1a | strain H77 | NP_751927 | TBC1D20 | | VirHostNet |
|---|---|---|---|---|---|---|---|
| hcv0574 | NS5B | HCV genotype 1a | H77 | NP_751928 | VAPA | | VirHostNet |
| hcv0575 | NS5B | HCV genotype 1a | H77 | NP_751928 | EIF4A2 | | BIND |
| hcv0576 | NS5B | HCV genotype 1a | H77 | NP_751928 | NCL | | BIND |
| hcv0577 | NS5B | HCV genotype 1a | H77 | AF009606 | NS5B | AF009606 | HCVpro |
| hcv0578 | NS5B | HCV genotype 1b | | | HAO1 | | HCVpro |
| hcv0579 | NS5B | HCV genotype 1b | | | TTC4 | | HCVpro |
| hcv0580 | NS5B | HCV genotype 1b | | | ACTN1 | | HCVpro |
| hcv0581 | NS5B | HCV genotype 1b | | | VAPA | | HCVpro |
| hcv0582 | NS5B | HCV genotype 1b | HCV-S1 | AAL00900 | DDX5 | | BIND |
| hcv0583 | NS5B | HCV genotype 1a | H77 | NP_751928 | PKN2 | | VirHostNet |
| hcv0584 | NS5B | HCV genotype 1a | H77 | NP_751928 | PTBP2 | | VirHostNet |
| hcv0585 | NS5B | HCV genotype 1b | | | FBXL2 | | HCVpro |
| hcv0586 | NS5B | HCV genotype 1a | H77 | NP_751928 | PPIB | | BIND |
| hcv0587 | NS5B | HCV genotype 1b | Strain J1 | D89815 | VAPB | | HCVpro |
| hcv0588 | NS5B | HCV genotype 1b | Korean isolate | | CHUK | | HCVpro |
| hcv0589 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | CEP250 | | HCVpro |
| hcv0590 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | CEP68 | | HCVpro |
| hcv0591 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | HOXD8 | | HCVpro |
| hcv0592 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | MGC2752 | | HCVpro |
| hcv0593 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | MOBK1B | | HCVpro |
| hcv0594 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | NR4A1 | | HCVpro |
| hcv0595 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | OS9 | | HCVpro |
| hcv0596 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | PKM2 | | HCVpro |
| hcv0597 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | PSMB9 | | HCVpro |
| hcv0598 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | SETD2 | | HCVpro |
| hcv0599 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | SHARPIN | | HCVpro |
| hcv0600 | NS5B | HCV genotype 1b | isolate con1 | AJ238799 | TUBB2C | | HCVpro |
| hcv0601 | NS5B | HCV genotype 1b | Isolate Con1 | CAB46677 | NS5B | CAB46677 | BIND |
| hcv0602 | NS5B | HCV genotype 1a | H77 | NP_751928 | UBQLN1 | | VirHostNet |
| hcv0603 | NS5B | HCV genotype 1b | | | CINP | | HCVpro |
| hcv0604 | NS5B | HCV genotype 1b | | | PSMB4 | | HCVpro |
| hcv0605 | NS5B | HCV genotype 1a | strain H77 | NP_751928 | 6-Sep | | VirHostNet |
| hcv0606 | NS5B | HCV genotype 1a | strain H77 | NP_751928 | HNRNPA1 | | VirHostNet |

| hcv0607 | p7 | HCV genotype 1b | strain (HC-J4) | | P7 | | HCVpro |
|---------|----|--------------------|-----------------|-----------|--------|--|--------|
| hcv0608 | p7 | HCV genotype 1b | | | FMNL1 | | HCVpro |
| hcv0609 | p7 | HCV genotype 1b | | | H19 | | HCVpro |
| hcv0610 | p7 | HCV genotype 1b | | | ISLR | | HCVpro |
| hcv0611 | p7 | HCV genotype 1b | | | MS4A6A | | HCVpro |
| hcv0612 | p7 | HCV genotype 1b | | | NUP214 | | HCVpro |
| hcv0613 | p7 | HCV genotype 1b | | | SSR4 | | HCVpro |
| hcv0614 | p7 | HCV genotype 1b | | | STRBP | | HCVpro |
| hcv0615 | p7 | HCV genotype 1b | isolate con1 | AJ238799 | CREB3 | | HCVpro |
| hcv0616 | p7 | HCV genotype 1b | isolate con1 | AJ238799 | FBLN2 | | HCVpro |
| hcv0617 | p7 | HCV genotype 1b | isolate con1 | AJ238799 | FXYD6 | | HCVpro |
| hcv0618 | p7 | HCV genotype 1b | isolate con1 | AJ238799 | LMNB1 | | HCVpro |
| hcv0619 | p7 | HCV genotype 1b | isolate con1 | AJ238799 | SLIT2 | | HCVpro |
| hcv0620 | p7 | HCV genotype 1b | isolate con1 | AJ238799 | UBQLN1 | | HCVpro |
| hcv0621 | p7 | HCV genotype 1b | isolate con1 | AJ238799 | UBQLN4 | | HCVpro |

# Appendix V (Chapter 3)

## Appendix Va

Hepatocellular carcinoma genes obtained from HCCNet and OncoDB.HCC databases

| Gene Name | Entrez Gene | Source | Source |
|-----------|-------------|--------|--------|
| ACLY | 47 | HCCNet | Oncodb.hcc |
| AEBP1 | 165 | HCCNet | |
| AGT | 183 | HCCNet | |
| AHNAK | 79026 | HCCNet | |
| AKT1 | 207 | HCCNet | |
| APOA1 | 335 | HCCNet | Oncodb.hcc |
| APOB | 338 | HCCNet | |
| APOE | 348 | HCCNet | Oncodb.hcc |
| ARNT | 405 | HCCNet | |
| AXIN1 | 8312 | HCCNet | Oncodb.hcc |
| AZGP1 | 563 | HCCNet | Oncodb.hcc |
| BCL2A1 | 597 | HCCNet | |
| C7 | 730 | HCCNet | |
| CANX | 821 | HCCNet | |
| CASP8 | 841 | HCCNet | |
| CD209 | 30835 | HCCNet | |
| CD5L | 922 | HCCNet | Oncodb.hcc |
| CD68 | 968 | HCCNet | |
| CD81 | 975 | HCCNet | Oncodb.hcc |
| CDC2 | 983 | HCCNet | |
| CDK2 | 1017 | HCCNet | |
| CDK6 | 1021 | HCCNet | |
| CDKN1A | 1026 | HCCNet | Oncodb.hcc |
| CEP152 | 22995 | HCCNet | |
| CEP68 | 23177 | HCCNet | |
| CFP | 5199 | HCCNet | |
| CHUK | 1147 | HCCNet | |
| CIDEB | 27141 | HCCNet | |
| CLEC4M | 10332 | HCCNet | |
| COL4A2 | 1284 | HCCNet | |
| CTGF | 1490 | HCCNet | Oncodb.hcc |
| CTSB | 1508 | HCCNet | Oncodb.hcc |
| DDX3X | 1654 | HCCNet | |

| | | | |
|---|---|---|---|
| DDX3Y | 8653 | HCCNet | |
| EEF1A1 | 1915 | HCCNet | |
| EFEMP1 | 2202 | HCCNet | |
| FAS | 355 | HCCNet | Oncodb.hcc |
| FBN1 | 2200 | HCCNet | |
| FGG | 2266 | HCCNet | Oncodb.hcc |
| FHL2 | 2274 | HCCNet | |
| FIGNL1 | 63979 | HCCNet | |
| FN1 | 2335 | HCCNet | |
| FYN | 2534 | HCCNet | |
| GAB1 | 2549 | HCCNet | |
| GAPDH | 2597 | HCCNet | |
| GBP2 | 2634 | HCCNet | |
| GRB2 | 2885 | HCCNet | |
| H19 | 283120 | HCCNet | Oncodb.hcc |
| HAO1 | 54363 | HCCNet | |
| HLA-A | 3105 | HCCNet | |
| HSPD1 | 3329 | HCCNet | |
| IKBKE | 9641 | HCCNet | |
| IRF3 | 3661 | HCCNet | |
| JAK2 | 3717 | HCCNet | |
| JUN | 3725 | HCCNet | |
| KRT19 | 3880 | HCCNet | Oncodb.hcc |
| KRT8 | 3856 | HCCNet | |
| LAMB2 | 3913 | HCCNet | |
| APOA2 | 336 | HCCNet | |
| MAGED1 | 9500 | HCCNet | |
| MBP | 4155 | HCCNet | |
| MCL1 | 4170 | HCCNet | Oncodb.hcc |
| MS4A6A | 64231 | HCCNet | |
| MT2A | 4502 | HCCNet | |
| MVP | 9961 | HCCNet | |
| NAP1L1 | 4673 | HCCNet | |
| NDC80 | 10403 | HCCNet | |
| NDRG1 | 10397 | HCCNet | |
| NOTCH1 | 4851 | HCCNet | |
| NUP62 | 23636 | HCCNet | |
| OAS1 | 4938 | HCCNet | |
| PDE4DIP | 9659 | HCCNet | |

| | | | |
|---|---|---|---|
| PDLIM5 | 10611 | HCCNet | |
| PKM2 | 5315 | HCCNet | |
| PPARA | 5465 | HCCNet | |
| PRKACA | 5566 | HCCNet | |
| PSMB4 | 5692 | HCCNet | |
| PTMA | 5757 | HCCNet | |
| RAB14 | 51552 | HCCNet | |
| RASAL2 | 9462 | HCCNet | |
| RGNEF | 64283 | HCCNet | |
| RXRA | 6256 | HCCNet | |
| SCARB1 | 949 | HCCNet | |
| SEPT6 | 23157 | HCCNet | |
| SERPINC1 | 462 | HCCNet | |
| SERPINF2 | 5345 | HCCNet | Oncodb.hcc |
| SERPING1 | 710 | HCCNet | Oncodb.hcc |
| SESTD1 | 91404 | HCCNet | |
| SIAH1 | 6477 | HCCNet | Oncodb.hcc |
| SLC22A7 | 10864 | HCCNet | |
| SLIT2 | 9353 | HCCNet | |
| SMAD3 | 4088 | HCCNet | |
| SMURF2 | 64750 | HCCNet | |
| SMYD3 | 64754 | HCCNet | |
| SRC | 6714 | HCCNet | Oncodb.hcc |
| STAB1 | 23166 | HCCNet | |
| STAT1 | 6772 | HCCNet | |
| STAT3 | 6774 | HCCNet | |
| SYNE1 | 23345 | HCCNet | |
| SYNPO2 | 171024 | HCCNet | |
| TACSTD2 | 4070 | HCCNet | |
| TBP | 6908 | HCCNet | |
| TF | 7018 | HCCNet | Oncodb.hcc |
| THBS1 | 7057 | HCCNet | Oncodb.hcc |
| TMF1 | 7110 | HCCNet | |
| TP53 | 7157 | HCCNet | Oncodb.hcc |
| TP53BP2 | 7159 | HCCNet | |
| UBQLN4 | 56893 | HCCNet | |
| VCAN | 1462 | HCCNet | |
| VIM | 7431 | HCCNet | Oncodb.hcc |
| VWF | 7450 | HCCNet | |

| | | | |
|---|---|---|---|
| YWHAE | 7531 | HCCNet | |
| YWHAZ | 7534 | HCCNet | |
| YY1 | 7528 | HCCNet | |
| ZCCHC7 | 84186 | HCCNet | |
| ZG16 | 653808 | HCCNet | |
| ACY3 | 91703 | HCCNet | |
| GRN | 2896 | | Oncodb.hcc |
| HSPA5 | 3309 | | Oncodb.hcc |
| TP73 | 7161 | | Oncodb.hcc |
| HSP90AA1 | 3320 | | Oncodb.hcc |
| BAX | 581 | | Oncodb.hcc |
| GSK3A | 2931 | | Oncodb.hcc |
| Vitronectin | 7448 | | Oncodb.hcc |
| ATF6A | 22926 | | Oncodb.hcc |
| KHDRBS1 | 10657 | | Oncodb.hcc |
| NR4A1 | 3164 | | Oncodb.hcc |
| IGLL1 | 3543 | | Oncodb.hcc |
| RAF1 | 5894 | | Oncodb.hcc |
| RBP4 | 5950 | | Oncodb.hcc |
| ITGB1 | 3688 | | Oncodb.hcc |
| YY1AP1 | 55249 | | Oncodb.hcc |
| HNRNPK | 3190 | | Oncodb.hcc |
| FADD | 8772 | | Oncodb.hcc |
| NPM1 | 4869 | | Oncodb.hcc |

## Appendix Vb

Hepatocellular carcinoma genes obtained from EHCO database with expression levels

| Gene_Symbol | Dataset | UP regulated | Down regulated |
|---|---|---|---|
| ACLY | GIS, SMD, UCSF, TOKYO | 4 | 0 |
| AEBP1 | SMD | 0 | 1 |
| AGRN | SMD, UCSF | 2 | 0 |
| AGT | SAGE, PubMed, mRNA | 1 | 1 |
| AKT1 | PubMed | 1 | 0 |
| ANKRD12 | KIM | 1 | 0 |
| APOA1 | SMD, SAGE, CGED | 0 | 3 |
| APOA2 | SAGE, PubMed, FUDAN | 1 | 1 |
| APOB | SAGE | 0 | 1 |
| APOE | SAGE | 0 | 1 |
| ARNT | PubMed, UCSF | 1 | 0 |
| ARS2 | FUDAN | 0 | 1 |
| AXIN1 | PubMed | 0 | 0 |
| AZGP1 | SMD, LEE, CGED | 0 | 3 |
| B2M | PubMed, mRNA, CGED | 1 | 1 |
| BAX | PubMed, mRNA | 1 | 0 |
| BCL2A1 | FUDAN | 0 | 1 |
| BIN1 | PubMed, mRNA, FUDAN, OncoFetal | 2 | 1 |
| C1QBP | TOKYO | 0 | 1 |
| C7 | GIS, SMD, mRNA | 0 | 3 |
| CALR | UCSF, CGED | 2 | 0 |
| CAMLG | SMD | 1 | 0 |
| CANX | SMD, SAGE, UCSF | 2 | 1 |
| CASP8 | PubMed | 0 | 0 |
| CCDC21 | SMD, UCSF | 2 | 0 |
| CCDC66 | SAGE | 1 | 0 |
| CD5L | GIS, KIM, TOKYO | 0 | 3 |
| CD68 | PubMed, LEE | 1 | 0 |
| CD81 | SMD, PubMed, mRNA | 1 | 1 |
| CDC2 | SMD, PubMed, UCSF, FUDAN | 3 | 0 |
| CDC23 | GIS, TOKYO | 2 | 0 |
| CDK2 | PubMed, FUDAN | 1 | 0 |
| CDKN1A | PubMed, mRNA, CGED, PASTEUR | 0 | 3 |
| CFP | PubMed, TOKYO | 0 | 1 |
| CHUK | PubMed, FUDAN | 0 | 1 |
| CLEC4M | mRNA | 0 | 1 |
| COL4A2 | SMD, UCSF | 2 | 0 |
| CRABP1 | PubMed | 0 | 0 |
| CSNK2B | SMD, UCSF, FUDAN | 3 | 0 |
| CTGF | PubMed, mRNA | 0 | 1 |
| CTSB | PubMed, Protein | 0 | 1 |
| DDX3X | PubMed, mRNA | 2 | 0 |
| DDX3Y | SMD, FUDAN | 0 | 2 |

| DES | PubMed, TOKYO | 0 | 1 |
|---|---|---|---|
| DNAJA3 | OncoFetal | 1 | 0 |
| DPP7 | SMD | 0 | 1 |
| EEF1A1 | Protein | 0 | 1 |
| EFEMP1 | SAGE | 1 | 0 |
| EIF1 | CGED | 0 | 1 |
| EIF2AK2 | PubMed | 0 | 0 |
| FADD | PubMed, CGED | 1 | 1 |
| FAS | SMD, PubMed | 0 | 1 |
| FBLN5 | mRNA | 0 | 1 |
| FBN1 | PubMed, LEE | 1 | 0 |
| FES | FUDAN | 1 | 0 |
| FGG | SMD, SAGE, mRNA, Protein, CGED | 1 | 5 |
| FLAD1 | SMD, UCSF | 2 | 0 |
| FN1 | mRNA, PASTEUR | 2 | 1 |
| FTH1 | mRNA | 1 | 0 |
| FUCA2 | CGED | 1 | 0 |
| FYN | SMD, PubMed | 0 | 1 |
| GAA | mRNA | 1 | 0 |
| GAB1 | SMD | 1 | 0 |
| GAPDH | PubMed, mRNA, Protein, OncoFetal | 4 | 0 |
| GFAP | PubMed | 0 | 0 |
| GLRX3 | Protein, FUDAN | 1 | 1 |
| GON4L | UCSF | 1 | 0 |
| GRB2 | Protein, PASTEUR | 0 | 2 |
| GRN | GIS, SMD, UCSF | 3 | 0 |
| H19 | SAGE, PubMed | 1 | 0 |
| HAO1 | Protein, CGED | 0 | 2 |
| HCK | LEE | 1 | 0 |
| HLA-A | SAGE, PubMed, mRNA | 1 | 1 |
| HNRNPA1 | LEE, mRNA | 2 | 0 |
| HSP90AA1 | mRNA, Protein, CGED | 3 | 0 |
| HSPA5 | GIS, SAGE, Protein, CGED | 3 | 1 |
| HSPD1 | SAGE, PubMed, mRNA, Protein | 3 | 1 |
| IGLL1 | SMD | 0 | 1 |
| IQWD1 | CGED | 1 | 0 |
| ITGB1 | GIS, PubMed, mRNA | 1 | 1 |
| ITGB4 | PASTEUR | 1 | 0 |
| JAK1 | PubMed | 0 | 0 |
| JAK2 | PubMed | 0 | 0 |
| JUN | SMD, FUDAN | 1 | 1 |
| KRT19 | SAGE, PubMed, LEE, mRNA, Protein | 5 | 0 |
| KRT8 | SMD, CGED, PASTEUR | 1 | 2 |
| LAMB2 | OncoFetal | 1 | 0 |
| LMNB1 | Protein | 1 | 0 |
| LTBP4 | TOKYO | 0 | 1 |
| LTBR | PubMed | 0 | 0 |
| LZTS2 | SAGE | 1 | 0 |

| | | | |
|---|---|---|---|
| MAPK12 | PubMed | 0 | 0 |
| MAPK7 | SMD | 0 | 1 |
| MBP | PubMed | 0 | 0 |
| MCL1 | SMD, CGED | 0 | 2 |
| MT2A | SAGE, mRNA, CGED, TOKYO | 0 | 4 |
| MYD88 | FUDAN | 0 | 1 |
| NAP1L1 | SMD, UCSF | 2 | 0 |
| NCL | GIS, PubMed, mRNA | 1 | 1 |
| NDRG1 | SMD, PubMed, LEE, UCSF, CGED | 4 | 0 |
| NEFL | PubMed | 0 | 0 |
| NOTCH1 | LEE | 1 | 0 |
| NPM1 | SMD, PubMed, LEE, mRNA, Protein, UCSF | 6 | 0 |
| NR4A1 | SMD, mRNA | 1 | 1 |
| NUCB1 | SAGE, CGED, OncoFetal | 1 | 2 |
| NUP62 | KIM | 1 | 0 |
| OS9 | mRNA | 0 | 1 |
| PABPN1 | PubMed | 0 | 0 |
| PDPK1 | KIM | 1 | 0 |
| PFDN2 | Protein | 1 | 0 |
| PIK3CB | PubMed | 1 | 0 |
| PIK3R1 | PubMed | 1 | 0 |
| PKM2 | LEE, mRNA | 2 | 0 |
| PLSCR1 | SMD | 0 | 1 |
| PML | PubMed | 0 | 0 |
| PPARA | PubMed | 0 | 0 |
| PPIB | GIS, SAGE, PubMed, mRNA | 3 | 1 |
| PSMB4 | GIS, CGED, FUDAN | 3 | 0 |
| PSMB8 | PubMed, FUDAN | 0 | 1 |
| PSMB9 | PubMed | 0 | 0 |
| PTPLAD1 | SMD, UCSF | 2 | 0 |
| PTPRN2 | SMD | 0 | 1 |
| RAB14 | mRNA | 0 | 1 |
| RAF1 | PubMed | 1 | 0 |
| RBP4 | SMD, SAGE, PubMed, Protein, CGED | 1 | 3 |
| RRBP1 | Protein, CGED | 0 | 2 |
| RSF1 | PubMed, CGED | 1 | 0 |
| RTN3 | SMD, UCSF | 2 | 0 |
| RXRA | LEE | 0 | 1 |
| SCARB1 | PubMed, CGED | 1 | 0 |
| SDC2 | SAGE, mRNA, CGED | 1 | 2 |
| SERPINC1 | SAGE, KIM, mRNA, FUDAN | 0 | 4 |
| SERPINF2 | SAGE, mRNA, FUDAN | 0 | 3 |
| SERPING1 | SMD, SAGE, mRNA, CGED, TOKYO | 1 | 4 |
| SHARPIN | UCSF | 1 | 0 |
| SIAH1 | SMD, PubMed | 0 | 1 |
| SLC22A7 | SAGE, CGED | 0 | 2 |
| SLC31A2 | SMD, KIM | 0 | 2 |
| SMYD3 | PubMed | 0 | 0 |

| | | | |
|---|---|---|---|
| SORBS2 | mRNA | 0 | 1 |
| SORBS3 | SMD | 0 | 1 |
| SRC | PubMed | 0 | 0 |
| SSB | PubMed | 0 | 0 |
| SSX2IP | FUDAN | 0 | 1 |
| STAT1 | PubMed | 0 | 0 |
| STAT3 | PubMed, Protein, CGED, PASTEUR | 0 | 3 |
| SVEP1 | SMD | 0 | 1 |
| TACSTD2 | SAGE | 1 | 0 |
| TAF1 | PubMed, mRNA | 1 | 0 |
| TAF9 | GIS, CGED | 2 | 0 |
| TF | SMD, SAGE, PubMed, KIM, mRNA, Protein, CGED | 1 | 6 |
| THBS1 | SMD, LEE, FUDAN | 1 | 2 |
| TLR2 | KIM | 0 | 1 |
| TMSB4X | SAGE, CGED | 1 | 1 |
| TNF | PubMed | 1 | 0 |
| TP53 | PubMed, mRNA, FUDAN, TOKYO | 0 | 3 |
| TP53BP2 | GIS, SMD, UCSF | 3 | 0 |
| TP73 | PubMed | 0 | 0 |
| TRADD | PubMed | 0 | 0 |
| TRAF2 | PubMed, LEE, mRNA | 1 | 1 |
| TRIM23 | FUDAN | 0 | 1 |
| TRIM27 | PubMed | 0 | 0 |
| TRIO | UCSF | 1 | 0 |
| TRIOBP | SMD | 0 | 1 |
| TRIP11 | UCSF | 1 | 0 |
| TSN | PubMed, mRNA, UCSF | 2 | 0 |
| UBA3 | FUDAN | 1 | 0 |
| UBQLN1 | FUDAN | 0 | 1 |
| UXT | PubMed | 0 | 0 |
| VCAN | PubMed | 0 | 0 |
| VIM | PubMed, mRNA, Protein | 3 | 0 |
| VPS35 | Protein | 1 | 0 |
| VPS52 | GIS | 1 | 0 |
| VWF | SMD, PubMed, UCSF | 2 | 0 |
| YWHAB | GIS | 1 | 0 |
| YWHAZ | SMD, UCSF | 2 | 0 |
| YY1AP1 | UCSF | 1 | 0 |

# Appendix VI (Chapter 3)

Appendix Va

Summary of data statistics of host proteins incorporated in HCVpro

| Features | Number of human proteins |
|---|---|
| Gene Name | 455 |
| Entrez Gene | 455 |
| HPRD ID | 453 |
| HCCNet | 116 |
| EHCO | 178 |
| Oncodb.hcc | 41 |
| Category | 455 |
| Description | 455 |
| Aliases | 446 |
| Previous Symbols | 117 |
| Previous Names | 126 |
| Chromosome | 455 |
| Entrez Gene Cytoband | 453 |
| Strand | 455 |
| Gene start | 453 |
| Gene end | 453 |
| Gene size | 454 |
| RefSeq RNA | 453 |
| RefSeq peptide | 452 |
| Ensembl Gene | 455 |
| UniGene | 453 |
| GenBank Accession | 439 |
| EMBL | 444 |
| OMIM | 383 |
| FUNCTION | 384 |
| SwissProt (name) | 447 |
| SwissProt (acc) | 447 |
| PDB id | 186 |
| IPI | 450 |
| Ensembl Protein ID | 454 |
| PharmGKB | 398 |
| InterPro | 408 |
| Pfam | 355 |
| KEGG | 163 |

| | |
|---|---|
| Reactome Pathway IdD | 86 |
| Nature Pathway | 42 |
| HPD | 85 |
| NetPath Pathway | 187 |
| GO | 435 |

## Appendix Vb

Summary of data statistics of viral proteins incorporated in HCVpro

| Features | Number of viral proteins |
|---|---|
| Short Name | 11 |
| Recommended Name | 11 |
| Other Names | 11 |
| Enzyme Commission | 3 |
| euHCVdb | 11 |
| OMIM | 11 |
| euhcvdb protein | 11 |
| GenBank | 11 |
| VBRC HCVdb Accession | 11 |
| Uniprot Post-translational modification | 10 |
| Uniprot Functions | 11 |
| Uniprot Subunit Structure | 10 |
| Uniprot Domain | 4 |
| Uniprot Subcellular Localization | 11 |
| Functions | 11 |
| Drug Development | 11 |
| Description | 11 |
| Molecular Mass by SDS-Page | 11 |
| Current State of Antiviral | 11 |
| Antiviral Progress | 7 |
| Uniprot IDs | 11 |
| Protein Type | 11 |
| Genome | 11 |
| Family | 11 |
| Genus | 11 |
| Strain Name | 11 |
| Molecule Type | 11 |
| Amino Acid Length | 11 |
| pI | 11 |

| | |
|---|---|
| Gene Size | 11 |
| Molecular Weight | 11 |
| Proteolytic Enzymes | 10 |
| NCBI CDD Pfam | 9 |
| Entrez Gene ID | 11 |
| GI | 11 |
| NCBI RefSeq | 11 |
| Pfam | 9 |
| Interpro | 9 |
| PDB | 7 |
| Recommended PubMed Readings | 11 |
| GO | 11 |

## Appendix Vc

Summary of data distribution in HCVpro

| Features | Number |
|---|---|
| Redundant virus-virus protein interactions | 72 |
| Non-redundant virus-virus protein interactions | 29 |
| Redundant virus-human protein interactions | 549 |
| Non-redundant virus-human protein interactions | 524 |
| Journal articles reporting interactions | 174 |
| BIND IDs mapped in HCVpro | 162 |
| VirusMint IDs mapped in HCVpro | 53 |
| HCVpro IDs | 621 |

# Appendix VII (Chapter 3)

## User manual for Hepatitis C Virus protein interaction database (HCVpro)

**Section A**

**1. Home Page**

The index page accessible via http://apps.sanbi.ac.za/hcvpro/ displays brief information about HCVpro and the menus (Search, manual, statistics, FAQ, links, download and contact)

**2. Querying database**



**3. Sub menu queries**

The enhanced user query interfaces enable users to perform searches via HCVpro by browsing the following sub query menus:

**Protein Select**

- HCV proteins: e.g. E2
- Human proteins: e.g. CD81
- Human Chromosome number: e.g. 10

**Evidence**

- Experiment type: e.g. coimmunoprecipitation
- PubMed ID: e.g. 10846074

**Identifier Search**

- HCVpro ID: e.g. hcv0128
- BIND ID: e.g. 183964
- MINT ID: e.g. MINT-14803

**String Search**

- Gene Name: e.g. Vimentin
- Gene ID: e.g. 7431
- Gene symbol: e.g. VIM

**Section B**

**Protein Search**

Here, we demonstrate querying of the HCVpro via the Hepatitis C Virus (HCV) proteins sub menu by browsing the list of HCV proteins. To retrieve interaction details on "NS5A" protein:
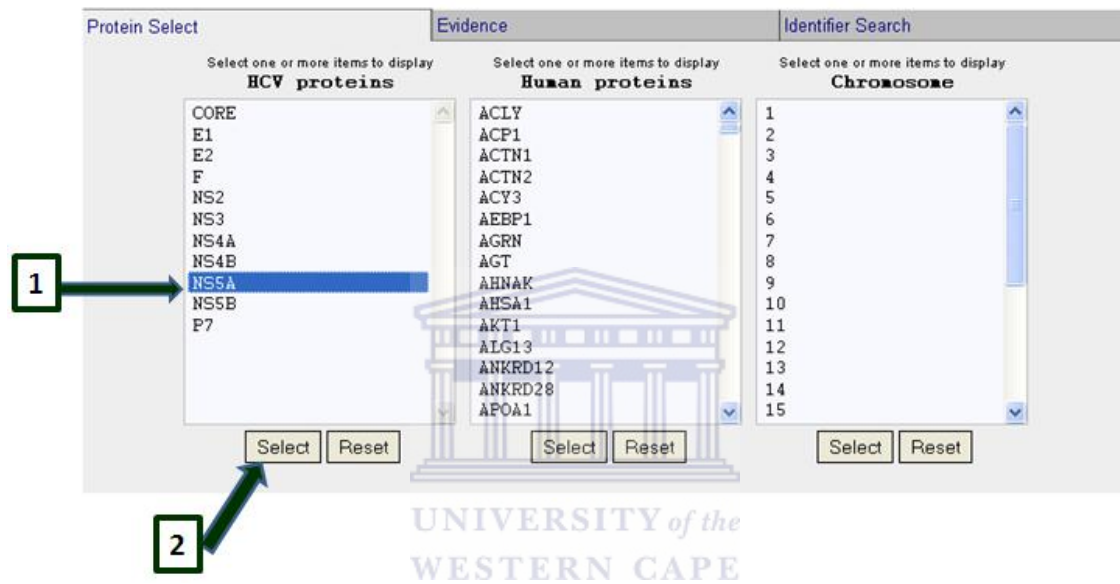


**1.** Click on "NS5A" protein

(NB: you may perform multiple queries by holding on the CTRL key and selecting the proteins simultaneously.)

**2.** Click on select to return a list of protein interactions as shown below:

**121 interactions found:**

| # | HCVpro ID | Molecule A | Molecule B Gene Symbols | PubMed ID | Details |
|---|---|---|---|---|---|
| 1. | hcv0453 | NS5A | EIF2AK2 | 9143277 | |
| 2. | hcv0454 | NS5A | CSNK2A1 | 10208859 | |
| 3. | hcv0455 | NS5A | GRB2 | 10318918 | |
| 4. | hcv0456 | NS5A | VAPA | 10544080 | |
| 5. | hcv0457 | NS5A | SRCAP | 10702287 | |
| 6. | hcv0458 | NS5A | IPO5 | 10799599 | |
| 7. | hcv0459 | NS5A | TP53 | 11152513 | |
| 8. | hcv0460 | NS5A | CDC2 | 11278402 | |
| 9. | hcv0461 | NS5A | CDK2 | 11278402 | |
| 10. | hcv0462 | NS5A | APOA1 | 11878923 | |
| 11. | hcv0463 | NS5A | TP53 | 12101418 | |
| 12. | hcv0464 | NS5A | TAF9 | 12101418 | |
| 13. | hcv0465 | NS5A | PIK3R1 | 12186904 | |
| 14. | hcv0466 | NS5A | TP53 | 12379483 | |
| 15. | hcv0467 | NS5A | TBP | 12379483 | |
| 16. | hcv0468 | NS5A | BIN1 | 12604805 | |
| 17. | hcv0469 | NS5A | PITX1 | 12620797 | |
| 18. | hcv0470 | NS5A | NS5B | 12692242 | |
| 19. | hcv0471 | NS5A | NS5A | 12692242 | |
| 20. | hcv0472 | NS5A | BAX | 12925958 | |
| 21. | hcv0473 | NS5A | SSB | 12963047 | |

**3.** Click on "SRCAP" protein to retrieve interaction details and enriched functional biological information on both "NS5A" and "SRCAP" proteins as below:

Click on all cross-referenced links to familiarize yourself with the information they provide.

**4.** Click on "Details Molecule A" link to retrieve comprehensive information on "NS5A"

For illustration view here:

http://apps.sanbi.ac.za/hcvpro/details_hcvprot.php?id=457&mol=a

**5.** Click on "Details Molecule B" to retrieve comprehensive information on "SRCAP"

For illustration view here: http://apps.sanbi.ac.za/hcvpro/details_humanprot.php?id=457

**Explained data fields**

Hepatocellular carcinoma associated genes are cross-referenced to the sources below:

- **HCCNet**: Hepatocellular carcinoma network database
- **EHCO**: Encyclopedia of hepatocellular carcinoma genes online
- **Oncodb.hcc**: Oncogenomic Database of hepatocellular carcinoma

NB: For details on EHCO gene expression levels consult: http://ehco.iis.sinica.edu.tw/

You may also consult the frequently asked questions and HCVpro data statistics at:

http://apps.sanbi.ac.za/hcvpro/FAQ.php and http://apps.sanbi.ac.za/hcvpro/Statistics.php/

**Section C**

Here, we demonstrate querying of the HCVpro via string search using the protein "vimentin"



**1.** Type "Vimentin" in the search area

**2.** Click on "search" to return a list of protein as shown below:

**3.** Click on "HCV interactions" as indicated above to access the protein interactions



- Click on "details" as indicated above to retrieve detailed interaction data on vimentin:



- Click on all cross-referenced links to familiarize yourself with the information they provide.

**4.** Click on "Details Molecule A" link to retrieve comprehensive information on "NS3"
For illustration view here:
http://apps.sanbi.ac.za/hcvpro/details_hcvprot.php?id=397&mol=a
**5.** Click on "Details Molecule B" to retrieve comprehensive information on "vimentin"
For illustration view here:
http://apps.sanbi.ac.za/hcvpro/details_humanprot.php?id=397

- The diagram below shows a snapshot of data on vimentin and links to GO ontologies, hepatocellular carcinoma databases, canonical pathways and other essential databases.
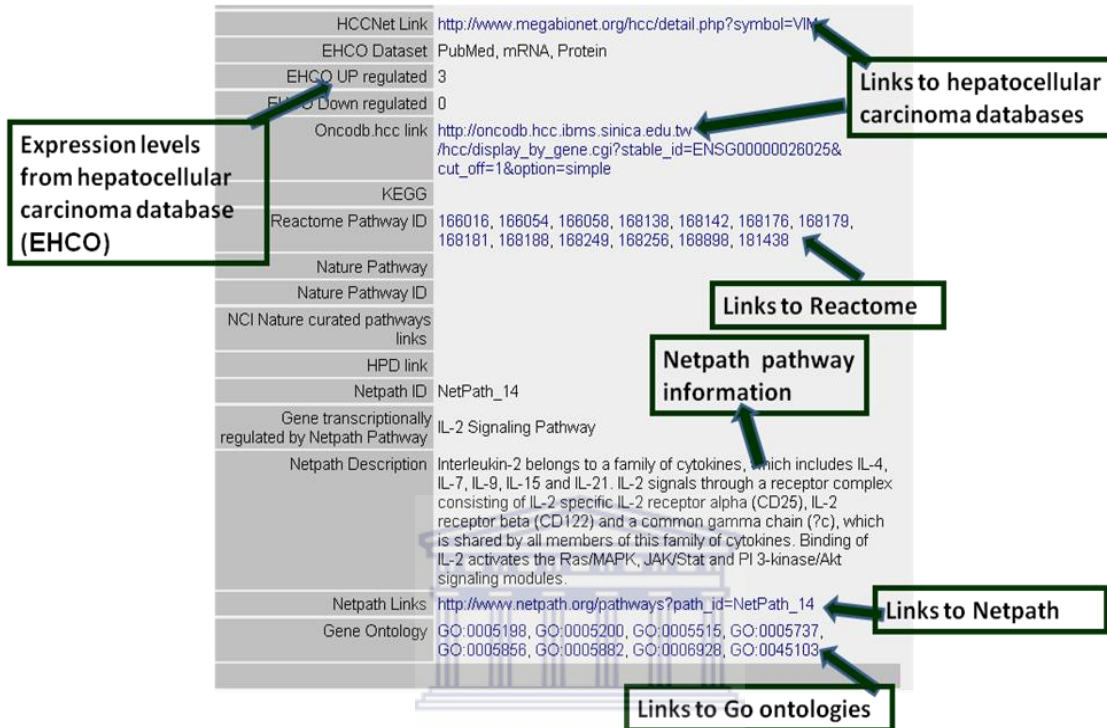
# Appendix VIII (Chapter 3)

## Frequently asked questions

### 1. What is HCVpro?

HCVpro is a freely accessible online knowledgebase comprising of Hepatitis C Virus (HCV) and human protein interactions.

### 2. Why was HCVpro created?

Currently HCV protein interactions data are found in literature and other publicly available databases but there is no single database solely focused on HCV protein interactions. HCVpro was thus created to offer "a one stop" knowledgebase on HCV protein interactions. Data integrated in HCVpro may enhance understanding of HCV life cycle, molecular function of its proteins, pathways and infection, and thereby augment efforts towards discovery of drugs, drug targets and diagnostic biomarkers.

### 3. What are the data sources in HCVpro?

HCVpro comprises of both manually verified literature- and database-curated interactions. Protein interactions databases employed for the curations were: BIND, VirusMint and VirHostNet. Additionally, functional annotation data were obtained from publicly available annotation resources and published literature.

### 4. What is unique about HCVpro, which cannot be found in other databases?

HCVpro provides enriched comprehensive and integrated data on HCV and human protein interactions, functional annotations, hepatocellular carcinoma (HCC) related gene expression, drug development, pathways, and cross-referenced links to other external

databases.

**5. What type of interaction data can be obtained from HCVpro?**

HCVpro provides two types of interactions:

- 1.HCV viral-viral protein interactions
- 2.HCV viral-human protein interactions

**6. How can I query HCVpro?**

The enhanced user query interface enables users to perform searches via the HCVpro by browsing the following sub query menus:

- Protein Select

  - HCV proteins: e.g. E2
  - Human proteins: e.g. CD81
  - Human Chromosome number: e.g. 10

- Evidence

  - Experiment type: e.g. coimmunoprecipitation
  - PubMed ID: e.g. 10846074

- Identifier Search

  - HCVpro ID: e.g. hcv0128
  - BIND ID: e.g. 183964
  - MINT ID: e.g. MINT-14803
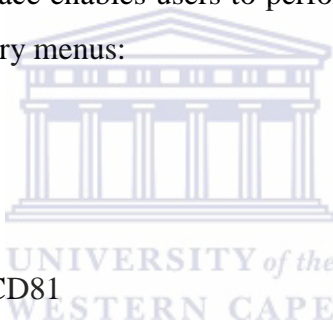
- String Search

  - Gene Name: e.g. Vimentin

  - Gene ID: e.g. 7431

  - Gene symbol: e.g. VIM

**7. Are there any future updates or modification in the pipeline?**

Yes, we would like to integrate Blast queries to enhance querying capabilities of HCVpro. Data on microRNA and siRNA would be integrated to augment efforts towards drug discovery. A complete HCV interactome would be provided in the background to enable users retrieve protein interaction sub-network and topological statistical features. Additionally, we will include features for computing druggability of protein interactions and also enable ligand and protein target screening online.

**8. Can researchers submit interactions to HCVpro?**

Yes, upon enquiries researchers may submit novel interactions either before or after publications in MIMIx compatible format. Data on interaction details such as domains, motif and residues involved in the interaction would be incorporated in the next update. Authors who have previously published HCV protein interactions would be called upon to submit interaction details or the developers involved in the HCVpro project will computationally infer as an alternative.

**The basic format for submission of interactions has been adapted from VirHostNet and is found below:**

| Protein A id | Protein B id | PSI-MI 2.5 method id | PubMed id |
| --- | --- | --- | --- |

Users may use the following links to convert interaction detection methods to PSI-MI 2.5:

http://obo.cvs.sourceforge.net/*checkout*/obo/obo/ontology/genomic-proteomic/protein/psi-mi.obo

or at

http://www.ebi.ac.uk/ontology-lookup/

**Click the links below for more details on PSI-MI data format:**

http://www.ncbi.nlm.nih.gov/pubmed/20078892

http://www.ncbi.nlm.nih.gov/pubmed/18936051

**Click the links below for more details on MIMIx data formats:**

http://www.ncbi.nlm.nih.gov/pubmed/17687370

**9. Can I download interaction data from HCVpro?**

Yes, users may download the full HCVpro interaction data in a tab-delimited format via the Download menu item.
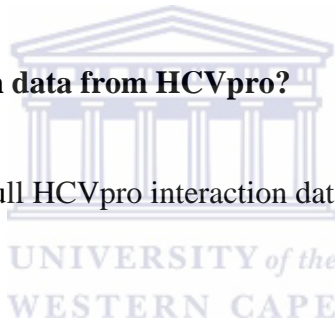
**10. Is there a user manual on HCVpro?**

Yes, users can access the manual via the Manual menu item.

**11. To whom can I report a bug or discrepancy?**

Please report all queries via the Contact menu item.