Large-Scale Integration of Microarray Data:

Investigating the Pathologies of Cancer and Infectious Diseases

A Thesis

Submitted to the Faculty

of

Drexel University

by

Noor Dawany

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

June 2010

Dedications

To my late grandmother, Dr. Siranoosh Raihani, for being a great inspiration…

Acknowledgements

First and foremost I would like to acknowledge my parents, Bassam and Nada, for their unconditional support throughout my educational life. I would not be here if it were not for their sacrifices and encouragement. I also want to thank my brother, Sahil, for telling me almost nine years ago to toughen up and continue; I have never looked back since then.

I owe my sincerest gratitude to my advisor, Dr. Aydin Tozeren, for his patience, his support and for all the effort he has put into this accomplishment. He is responsible for where I am today and where I will be, for that I am forever grateful.

I would also like to thank my colleagues at the Center for Integrated Bioinformatics: Mahdi Saramdy, Will Dampier, Yichuan Liu and Perry Evans, and my friends from the Biomedical Engineering Department for their help and for entertaining conversations that commemorate our lives as PhD students at Drexel University.

Finally, a big thank you for the support of old friends, for always being there for me…

Table of Contents

List of Tables

List of Figures

Abstract

Large-Scale Integration of Microarray Data: Investigating the Pathologies of Cancer and Infectious Diseases
Noor Dawany
Aydin Tozeren, PhD

DNA microarray data provide a high-throughput technique for the genome-wide profiling of genes at the transcript level. With large amounts of microarray data deposited on various types and aspects of malignancies, microarray technology has revolutionized the study of cancer. Such experiments aid in the discovery of novel biomarkers and provide insight into disease diagnosis, prognosis and response to treatment. Nonetheless, microarray data contains non-biological obscuring variations and systemic biases, which can distort the extraction of true aberrations in gene expression. Moreover, the number of samples generated by a single experiment is typically less than is statistically required to support the large number of genes studied. As a result, biomarker gene lists produced from independent datasets show little overlap. Therefore, to understand the pathophysiology of cancers and the influence they exert on the cellular processes they override, methods for combining data from different sources are necessary.

Meta-analysis techniques have been utilized to address this issue by conducting an individual statistical analysis on each of the acquired datasets, then incorporating the results to generate a final gene list based on aggregated p-values or ranks. However, much of the publicly accessible cancer microarray datasets are unbalanced or asymmetric and therefore lack data from healthy samples. Consequently, critical and considerable amounts of data are overlooked. An integrative approach that combines data prior to analysis can incorporate asymmetric data. For this reason, a merge approach to the previously validated technique, the significance analysis of microarrays, is proposed. The merged SAM technique reproduced the known-cancer literature with higher coverage than meta-analysis in the five independent cancer tissues considered. The same

methodology was extended to a database of approximately 6000 healthy and cancer samples arising from thirteen tissues. The integrative approach has allowed for the identification of key genes common to the invasive paths of multiple cancers and can aid in drug discovery. Moreover, this integrative microarray approach was applied to viral data from HIV-1, hepatitis C and influenza to investigate the effect of these infections on iron-binding proteins. Iron is crucial for proteins involved in metabolism, DNA synthesis and immunity, accentuating such proteins as direct or indirect viral targets.

Chapter 1: Introduction

1.1 Motivation:

This research is dedicated to the investigation of complex diseases through gene expression analysis. A large effort has already been directed towards the study of cancer, however many results from these various studies are inconsistent. The goal of this research is to provide a robust statistical approach that can help integrate different datasets to provide a generalized overview of the genetic aberrances that can be introduced during the course of carcinogenesis. The proposed methodology is also designed to take advantage of the current distribution of gene expression data, which in many cases is more inclined towards providing data on diseases and anomalous states as opposed to healthy control samples. The initial application of this approach was directed towards the study of cancer, a multifaceted disease comprising over 200 different malignancies, in order to identify common genetic alterations that could serve as effective drug targets. However, the scope of its application is unlimited and has been extended to understand the repercussions induced by persistent viruses such as HIV and hepatitis C as opposed to cytopathic viruses whose etiological agents are cleared quickly by the immune system.

1.2 Transcription, Translation and Control of Gene Expression:

Biological systems, whether at the level of a single cell, a multicellular tissue or a multi-tissue organism are complex entities. From an engineering perspective, these entities consist of several physiochemical and mechanical processes, which are governed by genes and proteins [1]. The human genome is coded into double-stranded DNA. A substantial fraction of the DNA can be transcribed to allow for the expression of the coded information in directing the synthesis of RNA and eventually protein molecules. The transcription process alone is sufficient for producing the functional RNA molecules (Figure 1) including messenger RNA (mRNA), transfer RNA (tRNA)

and ribosomal RNA (rRNA) [2]. The genetic coding region, however, is not necessarily continuous; instead the exon coding regions are interspersed with introns or non-coding regions. Introns are spliced out following transcription and prior to leaving the nucleus to the cytoplasm as part of the post-transcriptional modifications that mRNA undergoes before being translated [3]. While nucleic acids store and transmit the genetic information of the cell, the information itself is expressed in the form of proteins, therefore the protein-coding information from mRNA is translated into amino acid units that are connected to each other using peptide bonds to form polypeptides (Figure 1). The formation of the final protein depends on the number of polypeptide chains it contains, as well as the final three-dimensional structure it assumes which is determined by internal and external interactions with the protein's environment. The multiplicity of protein functions within cells ranges from reaction catalysis, transport of molecules and ions to immune response [2].

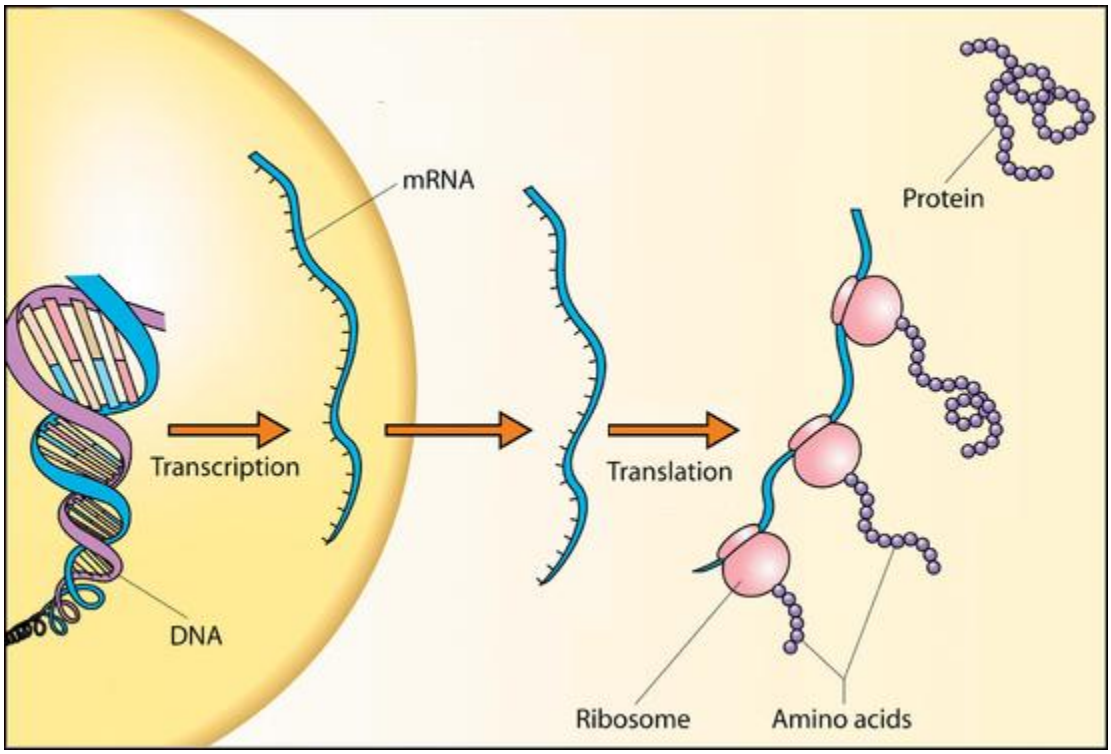

Figure 1 - Transcription and Translation (Adapted from [4])

In humans and multicellular organisms that possess more than one cell type, gene expression has to be controlled such that cells from distinct lineages develop differently and remain to be different. While the DNA code in the cells of an organism's pancreas and kidney are identical, they express different proteins at different levels. Therefore, during development and differentiation, cells have to control the expression of different sets of genes by switching them on and off as needed [3]. As transcription and translation are localized within different areas in the cell, gene expression can be accordingly regulated at more than one stage and location within cells [3, 5]. The primary mode however occurs in the nucleus as RNA polymerase interacts with the DNA promoter to initiate transcription. This binding can be sufficient for producing a few RNA molecules; however, the binding of transcription factors to enhancer sequences is essential for determining whether a gene will be transcribed. These transcription factors therefore help in regulating the time during which a gene is transcribed. This ensures that transcription occurs at the right developmental stage for example, and the right tissue location in which a gene is expressed. Nonetheless, additional regulatory instances can occur that could alter the way by which the primary mRNA transcript is processed or control the level of translation in the cytoplasm [3]. Hence, apart from their individual roles, the interactions between DNA, RNA and proteins are also important. DNA-protein interactions control gene expression, transcription, recombination, replication, packaging and repair. Similarly, since RNA is involved in various biological functions within the cell, RNA-protein complexes are also essential for these processes, including post-transcriptional regulation of gene expression and protein synthesis during translation [6].

Since the entire cell's genetic information is encoded within its DNA, the DNA must be faithfully replicated during every division cycle a cell undergoes. This ensures that the encoded genetic information is retained in the progeny cells. While DNA molecules possess a very stable

structure and have a longer half-life than RNA molecules, changes in the DNA sequence still occur. Alterations in DNA replication must be repaired and, in general, the location where the error exists is excised by DNA repair enzymes to readjust the base sequence. Since DNA is double stranded, the enzymes are required to identify which of the two strands is in fact damaged [3]. Despite employing repair mechanisms to control and fix DNA damage, cells can still undergo mutations that result in permanent changes to their genome. Genomic damage can arise from internal and external processes. Internally, damage can occur from errors during DNA replication, the chemical instability of some DNA bases and from free radicals produced during metabolism. On the other hand, external causes of DNA damage can be produced from interactions with ionizing radiation, ultraviolet radiation and certain chemicals, which can result in a cascade of mutations, especially if the damage is directed towards genes whose function is to ensure the accuracy of DNA replication [7]. Genetic variations can occur in different forms, ranging from deletions, insertions or single nucleotide polymorphisms, to large chromosomal anomalies like copy number variation, where whole sections of homologous sequences are gained or lost. Such changes can affect gene expression and alter gene dosage, leading to diseases, disorders or increased predisposition to genetic mutations [8].

1.3 Cancer Overview

Normal cells progressively transform into malignancies through the sequential acquisition of mutations that occur due to damage to the genome (Figure 2). The main targets for the progression of normal cells into malignant ones are genes involved in normal homeostatic mechanisms. These genes are therefore mainly involved in cell growth and death. Namely these targets are the oncogenes, which are activated by mutations and hence stimulate proliferation, or tumor suppressor genes, which typically code for proteins that behave as checkpoints during cell proliferation or death and are therefore inactivated by mutations [7]. Oncogene mutations can

occur through gene amplifications, chromosomal translocations, or mutations in the regulatory machinery of the gene. Examples of oncogenes include Jun oncogene, and myelocytomatosis viral oncogene homolog (*MYC*). Tumor suppressor genes, like tumor protein *p53* and breast cancer 1 and 2 (*BRCA1/2*) are targeted in the opposite manner with mutations resulting in shortened proteins due to deletions or insertions, missense mutations, or epigenetic silencing factors [9], like hypermethylation [10]. In some cases, inactivating mutations occur in genes that are involved in maintaining genomic integrity, thus facilitating the acquisition of additional mutations [7]. These stability genes could be involved in subtle DNA repair, including mismatch repair (*MMR*) and base-excision repair (*BER*) or can control processes involving larger chromosomal regions like segregation and recombination. Therefore, a single mutation is not sufficient to infer cancerous transformation. Instead, several mutations have to be acquired and while all genes can be affected by mutations in stability genes, only mutations in oncogene and tumor suppressor gene regulation affect the net proliferation of the cell [9].

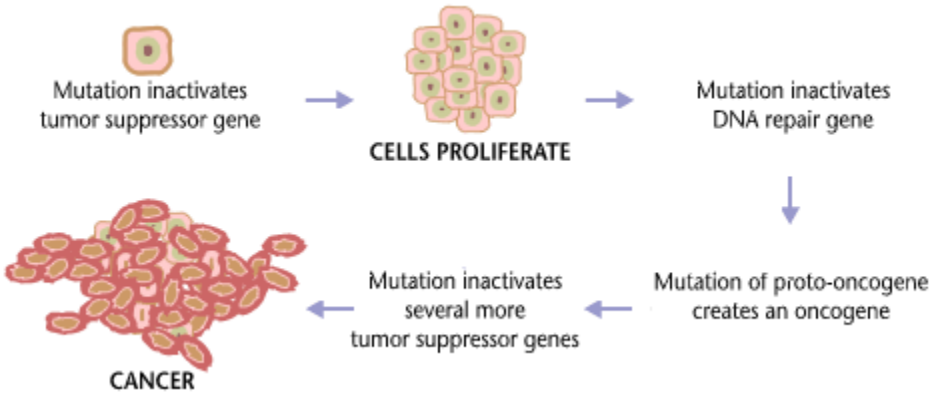

Figure 2 - Carcinogenesis: Transformation of normal cells into cancer (Adapted from [11])

Cancer is often preceded by chronic inflammation, although the role of inflammation in the malignant transformation is not fully understood. Examples include lung cancer following inflammation due to smoking and colon cancers following chronic inflammatory bowel disease.

As a result, several factors can induce cancer and these can be broadly categorized into *events* or *agents*. Events include environmental factors such as diet, occupation and chemical agents like carcinogens in tobacco. Meanwhile, several infectious viral agents can also promote cancer, including *Helicobacter pylori* causing gastric cancer, hepatitis B and C causing hepatocellular carcinoma and human papillomavirus leading to cervical cancer. While the specific genes involved might differ, the general mechanism of such infections involves triggering an inflammatory response through cytokines, chemokines and free radicals. The inflammatory response leads to the release of more free radicals which can contribute to the genetic mutations leading to the malignant transformation [12].

Regardless of cause or origin, once a cell has acquired and sustained a series of irreversible genetic alterations, a cancer develops through stochastic proliferation and differentiation [13]. Upon the completion of this transformation, a cancer cell must achieve two requirements to allow for its continual survival; it must overcome replicative senescence and obtain sufficient amounts of nutrients and oxygen supplies to support its high prolific activities by promoting angiogenesis [7]. Some cancers then metastasize beyond the site of their initial growth by entering the lymphatic system or the blood stream and localizing in a new tissue [14]. Because of the large heterogeneity of the different malignancies and the complex interplay between the various affected genes and pathways, the proliferation, progression and spread of cancer are all highly dependent on the alteration initially experienced by the primary cell and the resultant changes the cancer can impose on its surroundings.

1.4 Viral Infections and Hijacking Cellular Functions

There are two strategies that viruses generally employ for their survival. The first is a “hit and run” approach resulting in a quick infection, viral replication, cytolysis and transmission to a new

host. Such viruses are usually highly infective and easily transmissible like influenza and measles. However, other viruses are persistent, achieving long-term residence within the host. Either way, the virus must compete with the host cell for control of the cell's machinery, and it often dominates various components of the cellular mechanisms, imposing changes to the gene expression and pathway regulation through viral-host protein crosstalk [15].

1.4.1 Human Immunodeficiency Virus

The Human Immunodeficiency Virus (HIV) is the causative agent of the Acquired Immunodeficiency Syndrome (AIDS), perhaps the most serious viral infection to affect humans. HIV accounts for approximately 42 million cases worldwide, and the fatality rate is almost 100% [16]. There are two known types of HIV: HIV-1 and HIV-2, however the two types are not closely related to one another [17]. HIV-2 has a slower disease progression and limited impact on the survival of the majority of the infected adults compared to HIV-1 [18]. As a result, HIV-2 is confined to specific countries and has reduced pathogenecity. Moreover, there is better immune control of HIV-2 and a degree of CD4-independence. Most infections are caused by HIV-1 viruses [19], therefore from hereafter, the use of HIV will refer solely to HIV-1 infections.

The genome of HIV (Figure 3) is approximately 10 kilobases long and encodes 16 distinct proteins. The structural proteins are encoded into three HIV genes: group specific gene (*Gag*), polymerase (*Pol*), and envelope (*Env*). The remainder of the proteome consists of two regulatory proteins and four accessory proteins. The regulatory proteins: transcriptional transactivator (*Tat*) and regulator of virion gene expression (*Rev*) as well as the accessory protein negative effector (*Nef*) are expressed early in the viral cycle. *Tat* and *Nef* are necessary for inducing high levels of viral replication, while Rev regulates the gene expression transition from the early to late stages [20]. As HIV infects cells of the immune system, dendritic cells are believed to be among the

first cells that encounter the virus. Therefore, they mediate the transmission of HIV to CD+ T cells in the lymphoid tissue [21]. Viral entry into the immune system cells depends on the presence of specific chemokines receptors. HIV R5 strain binds to the CC-chemokine receptor 5 (*CCR5*) which is expressed on the surface of macrophages, dendritic cells and T cells. The X4 strain however can only infect T cells as it depends on CXC chemokines 4. Once inside, the virus is uncoated, its RNA is reverse transcribed and the resulting DNA is integrated into the cell's genome with a preference for active genes. Upon integration, the T cells become permissive allowing for the progression of the HIV infection. *Tat* controls the production of full-length viral transcripts. The mature HIV particles are then assembled and surrounded by the viral envelope in which the glycoproteins *Gp120* and *Gp41* are embedded, which are essential for viral binding to new cells [20].

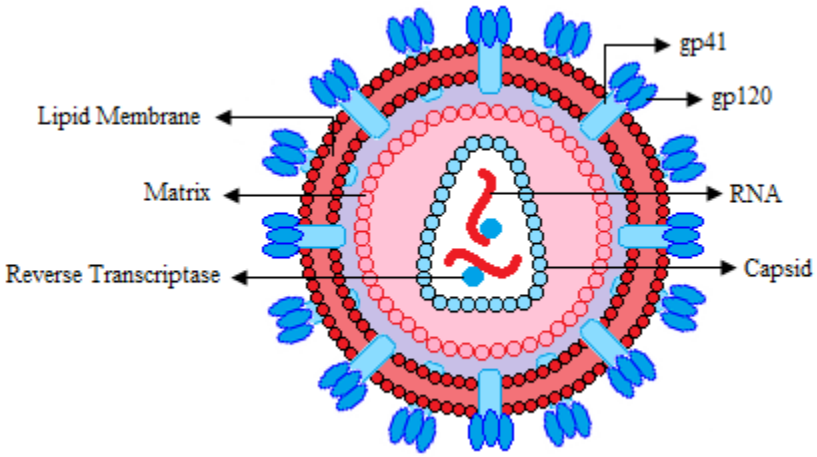

Figure 3 - Human Immunodeficiency Virus: Structure of HIV

What adds to HIV's chronicity is its ability to evade the host's immune system. Viruses can achieve that either by hiding in the microglial cells of the central nervous system since cell-mediated responses are naturally reduced there, or otherwise they enter a state of proviral latency in resting T cells. In addition, *Nef* decrease the expression of major histocompatibility complex

(MHC) class I molecules on the cell surface. Finally, once an infected cells has produced enough copies of the virus, HIV promotes apoptosis through the *FAS* and *TNF* death-inducing signaling pathways allowing for the virus to be released in order to infect new cells [20]. Consequently, following infection, there is a gradual loss of T cells resulting in a progressive immune deficiency that ultimately leads to opportunistic infections and death [22].

1.4.2 Hepatitis C

The hepatitis C virus is characterized by its high chronicity and it presents an international public health problem. It is transmitted primarily through the blood and is believed to infect approximately 3% of the world population. Hepatitis C infects the host's liver and can lead to acute hepatitis (20% of cases) or chronic hepatitis (up to 80% of cases). It can also lead to liver cirrhosis and has been associated with hepatocellular carcinoma [23-24]. The development of chronic hepatitis C depends on several factors including the viral genotype, the mode of viral acquisition and the immune response of the host [24].

Structurally, the virus has a positive sense, single-stranded RNA genome that is contained within a nucleocapsid [23-24]. The RNA and nucleocapsid are then packaged into an envelope that is derived from the host cell membrane upon viral release from the cell, in which viral-encoded glycoproteins are embedded (Figure 4). The RNA open reading frame encodes a polyprotein that is about 3300 amino acids long, which is cleaved inside the host cell to produce ten different polypeptides: the core peptide, two envelope peptides, six non-structural proteins in addition to a small hydrophobic protein [24-25]. The non-structural proteins comprise the RNA replicase complex of the virus needed for replication, while the core protein binds viral RNA to regulate its translation [25].

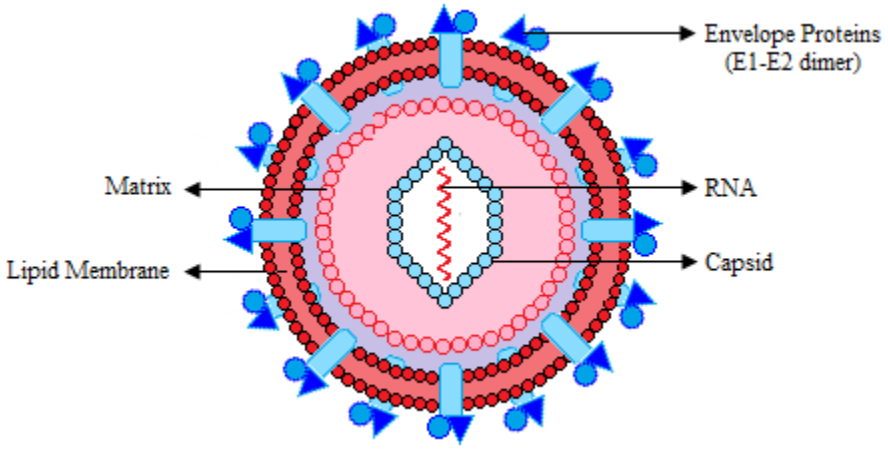

Figure 4 - Hepatitic C Virus: Structure of Hepatitis C

Aside from the typical role of the core protein, it possesses a wide array of functions as it interacts with several host proteins. First, *Core* has been shown to bind heterogeneous nuclear ribonuclearprotein K (*hnRNP K*), which in turn can recruit a range of molecules involved in signal transduction and transcriptional regulation. *hnRNP K* can also bind to DNA, RNA and transcriptional repressors and activators, in addition to acting as a shuttle between the nucleus and cytoplasm, implicating that it might be involved in RNA transport. *Core* also interacts with Lymphotoxin β receptor (*LT-βR*), a member of the tumor necrosis factor receptor (TNF-R) superfamily, expressed on the surface of most cell types. In addition, *Core* can bind to *TNF-R1* which mediates the tumor necrosis factor induction. More specifically, the viral protein associates with the death domain contained within *TNF-R1* that triggers the activation of consequent cell signaling pathways that control apoptosis. Studies have also reported *Core*'s interaction with an RNA helicase belonging to the DEAD box protein family, which are involved in several cellular activities such as mRNA splicing, RNA transport, ribosome assembly and translation, as well as controlling growth and differentiation [26]. As a result, the *Core* protein affects cell signaling, lipid metabolism, apoptosis, and carcinogenesis [25-26], although it is

unclear if these events are direct results of the viral infection or outcomes of protein over-expression within the cells [25].

Hepatitis C is also capable of avoiding the host's immune system by perturbing the host's ability to detect and destroy infected cells [26]. Over the course of an infection, Hepatitis C, like HIV, results in escape mutations. The host's immune response can be evaded by substituting the epitopes of nearby T cells by decreasing binding of the MHC and impairing antigen processing [27]. The virus can then persist and additional complications can arise that are associated with the autoimmune state of the host. These include the development of other syndromes like non-Hodgkin's lymphoma or coinfections with other viruses such as other hepatitis strains or HIV, thus further interfering with the cellular machinery and compromising the host's defense mechanisms [28].

1.4.3 Influenza A

The threat of human influenza epidemics is a recurrent issue which infrequently progresses into major worldwide pandemic as antigenically novel viruses are introduced to immunologically naïve human populations [29]. Influenza viruses have a segmented, negative sense RNA genome that is encapsidated by a viral nucleoprotein. Influenza is categorized into three types; A, B and C, based on the serological responses to their internal proteins [30-31]. However, influenza A has a greater impact on the human population as it is more common than type B [29], and generally causing the most serious respiratory illness compared to both types B [31] and C [30]. While the natural hosts for type A are aquatic birds, the viruses can infect a large variety of avian and mammalian species [29-31].

Influenza A viruses are composed of eight RNA segments that code for eleven viral proteins: two surface antigens, a nucleoprotein, two matrix proteins, three RNA polymerase proteins, and three non-structural proteins (Figure 5) including *PB1-F2* which plays a role in apoptosis [32]. These A-type viruses encompass a large variety of antigenically distinct subtypes based on the serological reactivities to their surface antigens, haemagglutinin (*HA*) and neuraminidase (*NA*) [30-31], for a total of 16 *HA* and 9 *NA* subtypes [33]. The *HA* glycoprotein is necessary for attaching to, and facilitating the entry and fusion into, the host cell. Meanwhile, *NA* is needed for breaking down cellular sialic acid to allow the virus to exit the host cell [34].

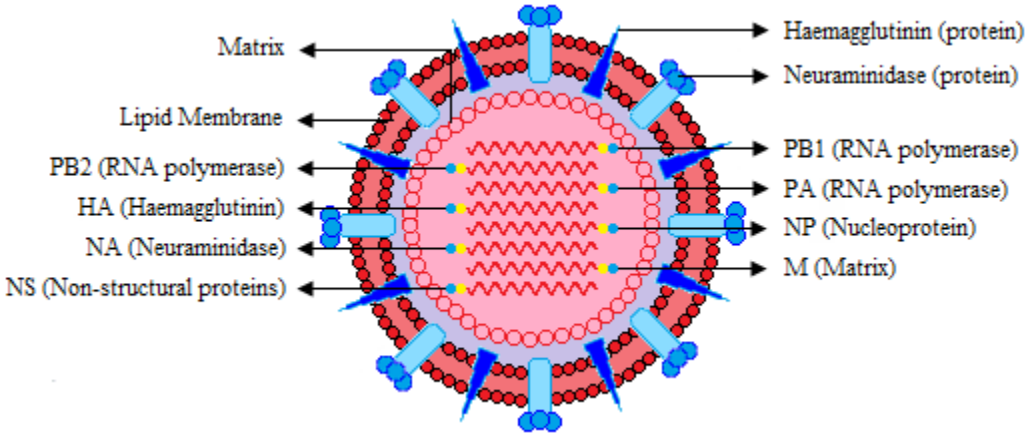

Figure 5 - Influenza A Virus: Structure of Influenza A

Upon binding to the host cell, the influenza virus activates proteins from the protein kinase C superfamily, which are linked downstream to several signaling pathways [35]. Once inside, like other viruses, influenza A exploits the host's cellular proteins and pathways to promote its own replication. Upon cellular entry, the viral core is disassembled and the genomic ribonucleoprotein (RNP) complexes are released into the cytoplasm and transported into the nucleus where the viral genome is transcribed. Most of the host factors therefore aid the virus by facilitating its replication [36]. Moreover, influenza A viruses induce the expression of several

cytokines and chemokines affecting their signaling cascades, including interferons α and β, as well as several interleukins. In addition three members of the mitogen-activated protein kinase (MAPK) family: extracellular-signal-regulated kinase, *JUN* N-terminal kinase, and *p38*, are believed to be activated due to influenza infections. MAPK signaling cascades are generally involved in several important cellular responses including cell activation, proliferation and differentiation, as well as immune response [35].

In humans, the influenza virus causes an infection that can vary in severity from asymptotic to a serious systemic illness. While it is unlikely that the virus replicates to a great extent outside of the respiratory tract, viral genes have been detected in peripheral blood mononuclear cells but without evidence of viral replication. In addition the viraemia is believed to occur as a result of respiratory complications enabling viral entry into the blood [31]. Therefore, the complex interactions that occur between viral and host proteins are essential for the virulence and the spread of the infection [30, 37].

In summary, while the cell machinery has been designed to help maintain a homeostatic environment ensuring healthy growth and proliferation, external events and infectious agents can interfere with these control mechanisms. Perturbations are introduced directly as mutations occur, or indirectly as invading pathogens usurp the cellular machinery. These changes often influence several pathways that occur downstream of the affected proteins, resulting in disturbances in the cell's regular function. As a result, the severity of these induced changes depends on the pathways concerned. Hence, the most deleterious infections and tumors are those that are capable of efficiently hijacking cellular machinery and directing nutrient supplies to meet their own demands, creating an environment fit for their continued proliferation and growth, and effectively avoiding or compromising the host's immune system.

Our understanding of the changes and influences associated with the invasion and/or progression of cancers and viral diseases is incomplete. One of the approaches that can help in identifying these perturbations at the transcript level is gene expression microarrays. A large effort has already been applied to the study of cancer [38-48], and similar analyses have been constructed to identify viral-induced changes [49-52], although these experiments have not been as extensive as those available for cancer. Nonetheless, such data allows for the identification of important genes and biomarkers that could aid in not only understanding the pathogenesis of different diseases, but also help in identifying candidate drug targets. Gene expression microarray analysis is further discussed in the next chapter.

Chapter 2: Gene Expression Microarrays

2.1 Introduction

Over the course of the past few years, the genomes of several organisms have been sequenced, however deciphering the DNA sequence does not reveal the function of genes, or the changes that occur in organisms due to disease, infections or even aging, for example. Microarrays have emerged as a useful tool for the simultaneous analysis of the expression of thousands of genes. They can help identify differentially expressed genes between two states which can facilitate the discovery of functionally important genes. In general terms, microarrays are affinity matrices in which labeled RNA or DNA is hybridized in solution to DNA molecules attached to the surface of the chip [53].

There are two basic types of DNA microarrays; complimentary DNA (cDNA) and oligonucleotide arrays. In the former, mRNA is obtained from two samples and labeled with different fluorescent dyes (Cy3 for reference sample and Cy5 for test sample). The experiment is then conducted as a competitive assay in which the two samples are hybridized to the same microarray chip and relative mRNA levels for each gene can be determined from the Cy3/Cy5 signal; a method that is termed two-color or two-channel microarrays (Figure 6a) [54]. In oligonucleotide arrays, such as Affymetrix GeneChip, each gene is represented by at least one set of 11-20 probe pairs, where each probe pair consists of 25 base pairs-long perfect match (PM) oligonucleotide probe and a mismatch (MM) probe of equal length. The MM probe matches the PM sequence with the exception of 1 base pair in the middle of the probe (13^{th} position) [54-55]. The purpose of MM probes is to measure non-specific binding [55]. The information across all the probe sets is then integrated and a probe set signal is produced. Oligonucleotide arrays are therefore one channel arrays in which only one sample is hybridized per chip (Figure 6b), and the

signal intensities from different chips can be analyzed and compared [54]. Single-color microarrays therefore require double the amount of chips needed for a two-color microarray experiment.

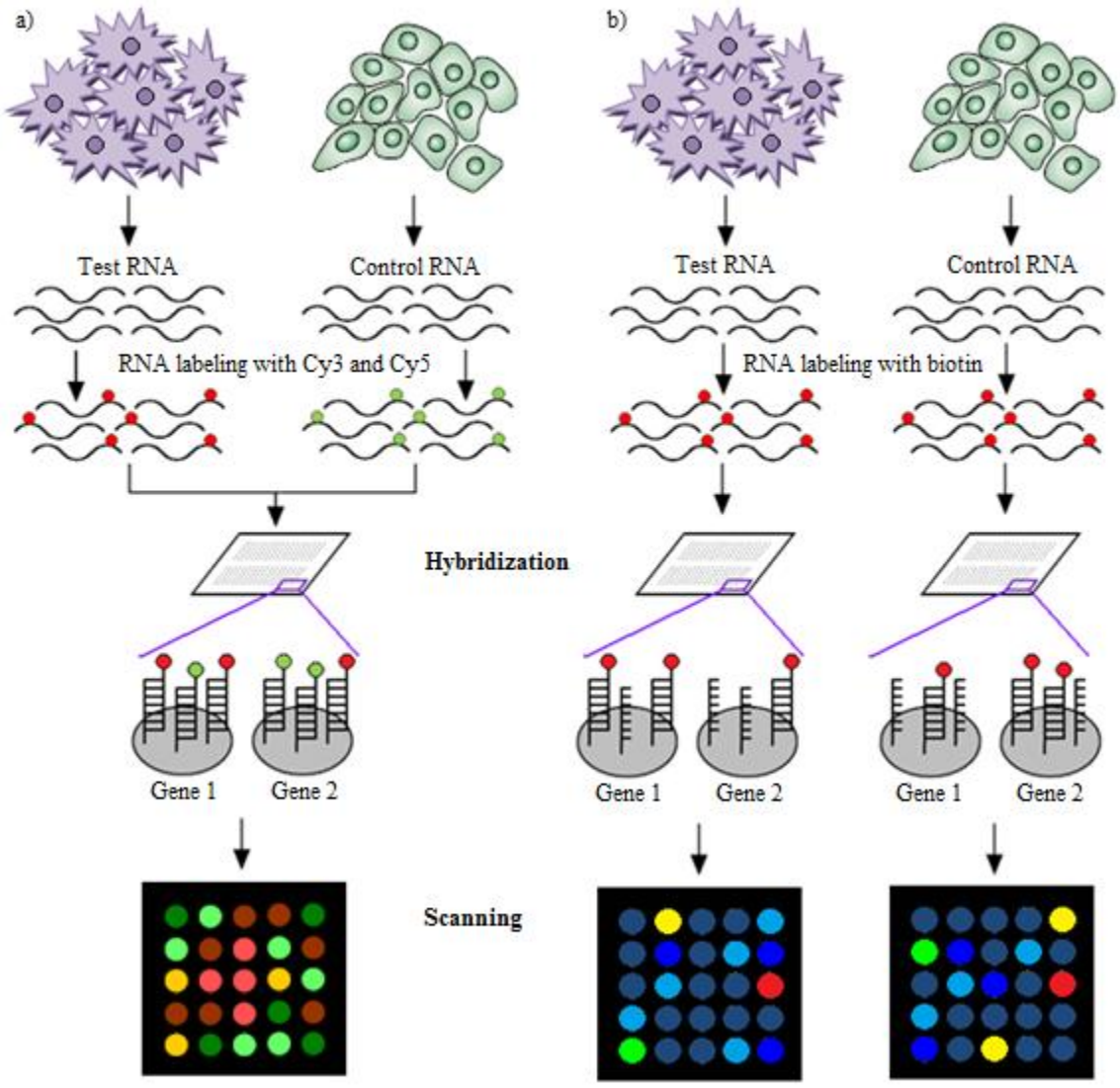

Figure 6 - DNA Microarray: Hybridization using a) two-channel and b) single-channel microarray platforms

While single-color arrays produce gene expression intensities for every sample used [54], two-color cDNA chips only provide a relative difference between two biological samples and

therefore are not a reliable source for the evaluation of absolute gene expression levels [56]. In addition, most cDNA arrays are in-house made by spotting or gridding the sequences onto a glass chip [57-59]. Commercially produced Affymetrix chips, on the other hand, use photolithography, utilizing ultraviolet light to direct where the oligonucleotide synthesis occurs on a siliconized glass surface [58-59]. Comparing data obtained from different samples as well as from different labs therefore requires normalization. As a result of the differences between the two types of microarray technologies, different standardization methods exist in each case. The normalization of custom-made cDNA chips is a complex problem in which the standard used is often based on the purpose for which the array was produced [58], hence the focus from hereon will be shifted to Affymetrix microarray chips.

2.2 Microarray Normalization:

When examining microarray data, two types of variation exist: informative variation and obscuring variation. Informative or interesting variations result from the conditions behind the study such as alterations accompanying a disease state, the effect of a protein or gene knockout, changes in environmental conditions (such as nutrients or temperature), introduction of infectious agents, mutations, or cellular stresses. Obscuring variations, on the other hand, can occur during the process of carrying out the experiment and can interfere with the interesting biological variations that occur between the two conditions. Obscuring variations can arise during the preparation of samples including variations during mRNA extraction, temperature fluctuations or reagent quality. In addition, some variations can occur while manufacturing the arrays such as the hybridization efficiency of the probes and probe concentrations. Finally, other obscuring variations arise from the processing of arrays, either during the hybridization of samples (differences in the amount of sample applied, buffer concentration and cross-hybridization interferences) or after array hybridization (variation in fluorescent intensity, optical

measurements and imaging algorithms) [60]. Thus the purpose of normalization prior to analysis is to deal with these obscuring variations [61] and unless the arrays are appropriately normalized, comparing values from different samples can lead to misleading results [62].

2.2.1 Robust Multichip Average Algorithm:

In order to produce gene expression levels that are representative of the hybridized DNA or mRNA samples, the probe intensities for each probe set have to be summarized. The robust multichip average (RMA) algorithm utilizes PM intensities only, as opposed to some of the other algorithms that use both PM and MM probes. The rationale behind this exclusion is due to reports that have revealed that the typical subtraction of MM values to correct for non-specific binding is not necessarily appropriate since the mathematical subtraction does not directly correspond to biological subtraction [55]. The pre-processing of Affymetrix microarray data includes three main steps: background-adjustment, normalization and final summarization of probe expression levels. Since all arrays are assumed to have a common mean background level, the PM intensities are adjusted to remove the background effect thus providing a more accurate absolute level of probe expression. Following background-correction, probe values are normalized using quantile normalization [62]. Values are transformed using the empirical distribution of each array and the empirical distribution of the averaged sample quantiles [61]. The purpose of quantile normalization is to make the distribution of probe intensities the same across all the arrays [61-62] and has been shown to produce favorable outcomes in terms of speed, variance and bias criteria when compared to other normalization algorithms[61]. Finally, for each probe on the array the background-corrected, normalized and $\log_2$-transformed PM intensities are fit into a linear additive model to remove probe-specific affinities. Median polishing is used to protect against outlier probes and to estimate the model parameters, resulting in the robustness of RMA [62].

2.2.2 Reference Robust Multichip Average

Classic microarray normalization and summarization methods including RMA and other multiple array-dependent algorithms present a major limitation in that the final model is applied to all the test samples used. In other words, the training samples and the test samples are the same. This dependency restricts the expansion of the model to additional data due to the lack of archived parameters that can be applied to an updated database of microarray samples. As a result, data from two studies cannot be directly compared if each has been normalized separately since each analysis used different data to define the normalization parameters and estimate the probe effects. This requires that the normalization technique be reapplied to the data as a whole to avoid pre-processing bias, a process that can create several constraints when dealing with large amounts of data, including time and memory restrictions. The reference robust multichip average (refRMA) algorithm, however, allows for the construction of a static normalization scheme that can be applied to added data on a continual basis [63]. The normalization process is termed static since the previously normalized data are not re-normalized with the addition of new data.

In short, a large number of biologically distinct Affymetrix microarrays are used to train the RMA model. Similar steps are applied to the training data as with the classical RMA, namely background-adjustment, quantile normalization and median polishing. The training process then produces two archived vectors; a probe effect vector compiled from the individual log-scale probe affinity effects and a normalization vector compiled based on the transformed PM intensities. The resulting vectors can then be extended to new test data by using the predetermined group of arrays to estimate the effects and the average empirical distribution that should be used for the added data. The final step differs, however, in that a full median polish summarization cannot be performed so the median is taken across probes from each probe set resulting in probe set level summaries [63].

2.2.3 Custom Chip Definition Files

Microarray data requires the use of chip definition files (CDF) in order to process the raw information obtained from the data files. Affymetrix CDF files encode the physical design of the chip. They also contain the sequence details that can be used to link the oligonucleotide probes that are present on the chip to the investigated transcripts [64]. Much attention has been directed towards statistical algorithms for normalizing data and detecting differentially expressed genes, yet problems related to probe and probe set identity can result in significant errors, especially when expression changes are not dramatic. Affymetrix had initially utilized the complete information available during the design stage, but with the immense progress achieved in genome sequencing and annotation in the past years, the Affymetrix probe set designs have become suboptimal [65]. Therefore, a gap exists in the correspondence between the probes and probe sets from the Affymetrix chips with the genes and transcripts [64-65]. Probe set annotation is constantly updated by Affymetrix and it has deviated from the original one-to-one correspondence between probe set and transcription locus. Nonetheless, the updates affect the qualitative attributes of the probe sets that control the effective matching between probes and genome sequences [64]. Analysis of chip definition files has revealed that several of the old probe sets do not truly reflect the expression levels of several significant genes in a given tissue [65].

Entrez custom CDF files are part of a collection of custom CDF files created by Dai et al. [65]. The process includes mapping probe sequences to individual sequences found in UniGene, dbSNP, and the genome sequence of the species and then aligning these sequences. Probes matching non-transcribed regions are excluded and only probes that have one perfect match with the corresponding genome sequence are retained. The probes in all probe sets are also required to

be aligned in the same direction on the genome. Finally, each probe set has to contain at least three probe pairs [65]. The resulting expression data is representative of an individual gene as opposed to the original Affymetrix CDF that produce gene intensities per probes where one gene can be represented by more than one probe and one probe can be mapped to multiple genes. The development of custom CDF files has been shown to significantly improve the outcomes of differential expression microarray analysis [64-65].

2.3 Microarray Analysis and Differential Gene Expression

Microarray technology has proved to be useful as it allows for the simultaneous quantification of thousands of genes in a high-throughput and cost-effect manner [66]. In many cases, the objective of the microarray data is to identify genes that are differentially expressed between the different conditions considered [67]. As a result, a large variety of methods have emerged for the analysis of microarray gene expression data. One of the simplest calculations is computing the fold change of a gene [68]. However, this is a statistically inefficient approach due to the systemic and biological variations that occur in such experiments. While some biases can be effectively removed through normalization, sample-to-sample variation cannot be accounted for in this manner. Hence, the use of fold change as the sole statistic of significant genes can increase the number of false positives (type I error) or false negative (type II error) identified. It is more appropriate to detect significantly altered genes by calculating a statistic based on replicate array data, then ranking genes and determining a cutoff value [69]. Statistical methods for ranking genes include the student's t-test [70], analysis of variance (ANOVA; [71]), Mann-Whitney test [72] or the Bayesian method [73-74].

The statistical cutoff can then be set, however it has to balance the false positives and the false negatives. In addition, since a microarray chip contains thousands of genes, setting a cutoff of

0.05 for an experiment studying 10,000 genes results in 500 genes falsely inferred as significant, exaggerating type I error. Statistical tests therefore have to deal with problems arising from multiple hypothesis testing. One example is the Bonferroni correction were the significance cutoff is divided by the number of genes, but such a correction can be too stringent. Hence, it is more practical to control the expected proportion of false positives by controlling the false discovery rate (FDR), as is the case with the significance analysis of microarrays (SAM; [75]) test [69].

2.3.1 Significance Analysis of Microarrays

Similar to the aforementioned tests, SAM determines significantly altered genes by assigning a score to each gene that is based on the change of the gene's expression relative to the standard deviation of the repeated measurements for that gene. This calculation is similar to that of a t-test However, a value is added to the standard deviation in the denominator, minimizing the coefficient of variation. Significance is associated with larger scores passing the set cutoff. The FDR is then calculated to determine what percentage of the genes were identified as significant by chance. FDRs are estimated by using random permutations of the gene expression measurements and calculating the expected relative difference for a gene from these permutations. SAM's performance has been shown to be superior to that of other conventional microarray analysis methods [75]

Nonetheless, despite the many statistical methods designed to deal with the different issues that arise from determining significant genes, discrepancies in the results from similar studies still occur [76]. Microarray studies using different datasets can report non-reproducible findings or produce results that are not robust even to the slightest data perturbations. While such problems can occur due to multiple reasons including improper analysis or insufficient control of false

positives, as previously discussed, the lab-dependency of results is exacerbated by the small number of samples utilized in individual studies. These numbers are generally much smaller than what is statistically needed to support the thousands of genes analyzed. Therefore, one solution to increase the statistical power, reliability and generalizability of microarray-based results is to combine information from several existing experiments [46].

2.3.2 Meta-Analysis

The term meta-analysis is used to describe statistical approaches that combine the results of independent but relevant studies. Meta-analysis techniques have been widely used for clinical trials and epidemiological studies [77-82], as well as for microarray analysis [46-48, 83-86]. In general, gene lists can be obtained from different studies and compared and a final gene list is produced which reflects the results of the multiple studies. However, it is more preferable to obtain and re-analyze gene expression data from each experiment. The results are then aggregated into a final gene list by considering the statistical enrichment of a gene across all studies. There are four main methods for combining information across studies: vote counting, combining ranks, combining p-values and combining effect sizes [46]. In vote counting, a gene receives a vote each time it appears to be significant in a list. However, the main difficulty with the approach is determining the minimum number of votes required to deem a gene significant. This can be further complicated as some genes may not even be analyzed on specific platforms. As a result, resampling methods [87-88] are required to estimate the significance of the different findings [89]. When combining ranks, the top-ranked genes are obtained from each study and the location of the genes within these lists is used to assess their overall significance [46]. Approaches for aggregating ranks utilize different algorithms including Markov chains [86] and Monte Carlo permutations [76]. P-values can also be combined across studies; in the Fisher's sum [47, 90] approach, for example, the logarithms of the different p-values for a gene are

accrued across all experiments. Finally, effect size can be aggregated, where the effect size represents a measure of the strength of the relationship between the two gene states considered, to obtain significant and meaningful results. Combining effect sizes is described in the following section as part of the inverse-variance methodology [46].

2.3.2.1 Inverse-Variance

The inverse-variance model [91] is believed to be the most comprehensive meta-analysis approach that can be applied for the comparison of two-class gene expression microarrays [46]. It has been used in several microarray-based meta-analysis studies [46, 83-84, 92]. For each gene in a study, the effect size and the variance associated with it are calculated. The effect size can be computed using an adjusted value of the initial distance parameter used by Cohen [93] based on the typical t-test value where the difference between two independent means is standardized by dividing the difference by the common within-population standard deviation. The adjusted value was introduced by Hedges and Olkin [94] as a correction factor for the calculation of the unbiased estimate of the effect size. This adjusts the overestimation of the effect size in studies with small sample sizes that results from standardizing the mean difference [46, 94]. A weighted-average, inversely-proportional to the variance of the study-specific estimates, is then used to combine the effect sizes of a gene across the different studies [46].

The inverse-variance technique possesses several advantages. First, the method takes into account information from all available genes. In addition, it can combine data from different platform technologies including one-color Affymetrix chips and two-color cDNA microarrays. Given the differences in the genes included from one platform to another, whether within the same technology (i.e. different Affymetrix platforms), or across technologies (i.e. Affymetrix vs. cDNA), some genes will be studied more frequently than others. It is therefore essential that a

statistical approach treats both frequently and rarely studied genes equally, and the inverse-variance addresses this issue by calculating the weighted average of the effect sizes. Thus, it weighs the contribution of a study by its precision, with more weight emphasis given to larger sample studies. And finally the parameters calculated by the inverse-variance; the pooled effect size and the standard error, can be biologically interpreted [46].

2.4 Databases

Several online databases are now available, providing access for biological information and data. Two of the biggest efforts are the National Center for Biological Information (NCBI; http://www.ncbi.nlm.nih.gov) and the European Bioinforamtics Institute (EBI; http://www.ebi.ac.uk). These databases include nucleotide sequence and microarray data, protein information, gene descriptions, and disease annotations, among others. Functional annotation databases for gene subsets of interest are also important for providing information on the connectivity of these genes and the roles they play within the cell. These databases help categorize genes and proteins based on similarities in their characteristics to provide higher order functions.

2.4.1 Microarray Databases:

With the increasing utilization of microarray data and with raw data availability being required by the Microarray and Gene Expression Data Society (MGED Society; [95]), microarray databases have aided in simplifying the acquisition of publically accessible data. NCBI's Gene Expression Ominbus (GEO; [96-97]) and EBI's ArrayExpress Archive [98] are two such repositories. Both databases support the retrieval of microarray data for analysis from a variety of organisms. They contain extensive amounts of data from different experimental settings including disease states and stages, genetic interventions and gene-knockouts, time series, and manipulative treatments

and drug therapy [96-98]. The availability of multiple microarray datasets is useful for noise reduction and adding sensitivity to the results extracted from the data.

2.4.2 Functional Annotation Databases:

As sequencing projects continue to provide gene catalogs, the functional annotation of these genes is essential yet incomplete. Biological functions within a cell usually cannot be attributed to a single gene or molecule but instead to a group of genes or molecules. The Kyoto Encyclopedia of Genes and Genomes (KEGG; [99-101]) is one of the databases that aim to combine genomic information with higher order functional information by integrating the available literature on cellular processes. KEGG links gene sets within a network of interacting molecules to provide complexes and pathways in which these genes function. These pathways cover a variety of processes such as metabolic pathways, diseases, genetic and environmental information processing, and signaling pathways. Apart from the visual networks of the pathways, KEGG also provides information within its sub-databases including the compound database for chemical structures, enzyme database for enzymatic nomenclature, and the reaction database containing reaction formulas [99-101].

The continuous accumulation of biological data and information has also produced a need for unifying annotation standards. The Gene Ontology (GO; [102]) Consortium has therefore established the structured vocabulary needed to facilitate communication between researchers as well as to provide consistent descriptions of gene products. The main categories described within the consortium are molecular function, biological processes and cellular components. Molecular function focuses on the activities at the molecular level rather than discussing the entities (molecules or complexes) that are responsible for the actions, or where and when the activity occurs. Biological processes, on the other hand, describe the biological goals that are

accomplished by one or more molecular functions. Finally, the cellular component term refers to the subcellular locations of structures and macromolecular complexes. The main ontologies are divided into several subcategories at different hierarchical levels with increasing degrees of specificity, resulting in a dynamic, controlled vocabulary for annotating genes and proteins [102-103].

Given a set of genes of interest that could be identified from a microarray experiment, the genes can be annotated using KEGG pathways and GO categories. Over-representation of such terms using a hypergeometric test can indicate functional enrichment of gene sets. These significant pathways and GO terms can then indicate which activities and processes are affected by the perturbation introduced to the system, providing an intelligible account of the cell's state under the specified conditions.

Chapter 3: Asymmetric integration of microarray data outperforms meta-analysis approach

3.1 Summary

This chapter focuses on the integrative and meta-analysis approaches used in the analysis of microarray data. Much of the public access cancer microarray data is asymmetric, belonging to datasets containing no samples from normal tissue. Asymmetric data cannot be used in standard meta-analysis approaches such as the inverse variance method, but are necessary for obtaining large sample sizes for statistical power enrichment. Noting that plenty of normal tissue microarray samples exist in studies not involving cancer, the viability and accuracy of an integrated microarray analysis approach based on significance analysis of microarrays (merged SAM) using a collection of data from separate diseased and normal samples was investigated. The research focused on five solid cancer types (colon, kidney, liver, lung, and pancreas), where available microarray data allowed for the comparison between meta-analysis and integrated approaches. Results from the merged SAM significantly overlapped gene lists from the validated inverse-variance method. In addition, both meta-analysis and merged SAM approaches successfully captured the aberrances in the cell cycle that commonly occur in the different cancer types. However, the integrated SAM analysis replicated the known cancer literature (excluding microarray studies) with much more accuracy than the meta-analysis. The merged SAM test is therefore a powerful, robust approach for combining data from similar platforms and for analyzing asymmetric datasets, including those with only normal or only cancer samples that cannot be utilized by meta-analysis methods. The integrated SAM approach can also be used in comparing global gene expression between various subtypes of cancer arising from the same tissue.

3.2 Background

Microarray studies typically provide intensity levels for thousands of genes. However, not only are the individual datasets usually small in size, but the inferences made from individual studies are often inconsistent with similar studies [76]. As thousands of microarray samples have accumulated in publicly accessible databases in the last decade [96-98], several statistical methods have been developed to allow for the combination and comparison of data from multiple sources. Among the many methodologies that exist that deal with merging different microarray datasets, are the permutation tests [47-48], parametric tests and clustering [104], rank-aggregation procedures [86, 105], rank products [106], METRADISC [76], and inverse-variance [46, 83-84]. The utilization of vast amounts of microarray data provided by different groups is considered to increase the reliability of results and weaken the effects of lab-specific noise [107].

The meta-analysis procedures cited above combine results from different studies. Each dataset is analyzed separately. Genes are associated with an effect size or a p-value. These are then combined across all analyses and a top-ranked gene list is generated based on the aggregated effect size or p-value [108]. While some meta-analysis methods require the use of raw data [46-48], others can depend solely on the ranking of genes from various studies [86, 105]. The meta-analysis is robust in the sense that it allows for comparisons across different platforms and analytical techniques (cDNA and oligonucleotide microarrays). However, the most important limitation the meta-analysis poses is that it requires datasets to include both control and test samples. Previous studies showed that aggregating data prior to obtaining results is usually more powerful than obtaining separate statistics from each dataset and then integrating the results [109]. Therefore, based on the grounds of previous studies that revealed the predictive potential of integrated microarray [110-112], this study considers a large-scale merge approach to the significance analysis of microarrays (SAM; [75]) test that can utilize asymmetric datasets.

To test the performances of the meta-analysis and the merged SAM approach, microarray data was compiled from 31 laboratories, resulting in a database containing 339 healthy tissue samples and 1,429 cancer samples from five different tissue types using comparable Affymetrix platforms. The tumor tissue types considered in this study –colon, kidney, liver, lung, and pancreas – had multiple microarray datasets containing both normal and disease samples. The meta-analysis approach has already been employed by a few cancer microarray studies either focusing on a single tissue type [47, 84, 113-115] or across different tissues in order to identify gene sets associated with common cancer mechanisms [46, 48, 116]. For the purpose of this study, the inverse-variance (IV) test was adopted from the work of Ramasamy et al. [46] to compare the quality of our results, since it is believed to be the most comprehensive meta-analysis method for two-class microarray gene expression analyses. With this large-scale database, significantly altered gene lists were generated for each individual tissue as well as across all five tissue types, using both the IV and the merged SAM tests. The results revealed that the merged SAM analysis, when based on large-scale data, not only significantly overlaps the results produced by the IV meta-analysis, but also provides gene lists that replicate the known cancer literature at least as well as the IV test.

3.3 Materials and Methods

3.3.1 Microarray dataset selection

A total of 31 Affymetrix microarray datasets containing 1,768 unique samples from human cancer (1,429) and corresponding healthy control tissues (339) were collected from the Gene Expression Omnibus (GEO; [96-97] and Array Express [98] online repositories. Samples were selected for five different tissue types: colon, kidney, liver, lung and pancreas, then categorized into cancer and control subsets to allow for intra- and inter-tissue comparisons. The cancer samples were not restricted to a single type of malignancy in order to provide a generalized

pathogenic approach shared by cancers. The microarray data were limited to those hybridized on the Affymetrix human microarray platforms HG-U133A, HG-U133A 2.0, and the HG-U133 Plus 2.0 due to the large overlap between the three platforms. In addition, the inclusion criteria restricted that each dataset was obtained from a peer-reviewed study and contained a minimum of 20 usable microarray samples (Figure 7).

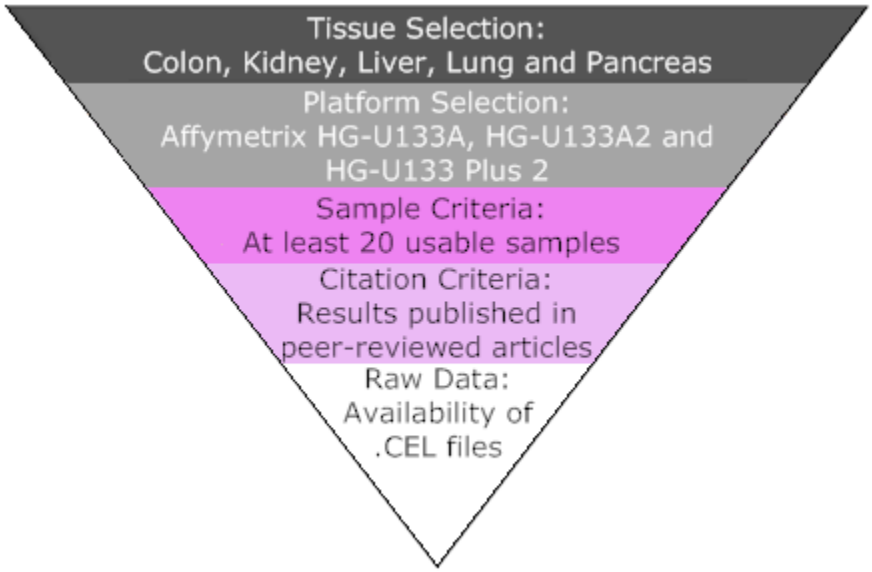

Figure 7 - Dataset Inclusion Criteria: Selection method used for the inclusion of Affymetrix datasets utilized in the analyses in this study.

3.3.2 Normalization and differential expression

For Affymetrix chips, raw microarray CEL files were read using the platform-compatible custom ENTREZG CDF file (version 12) [117] in order to obtain Entrez gene intensities. Where multiple replicates from the same source were available, the gene intensities were averaged across replicates. Nineteen out of thirty-one datasets contained samples for both the normal and cancer tissues and therefore could be used in meta-analysis. Individual datasets were background adjusted normalized with median polish using the robust multi-array analysis (RMA) in MATLAB [62]. For each tissue, the corresponding log-transformed data were transferred into R

[118] and the metaGEM package [46] was utilized to conduct the meta-analysis using inverse variance (IV1; Figure 8). The false discovery rate (FDR) was set at 0.001%. Moreover, the samr package [119] in R was used to conduct the significance analysis of microarrays (SAM) test [75] on each individual dataset. A hundred permutations were performed and results were restricted to significant genes with an FDR of 0.

While IV analyzes each dataset separately before combining the results, SAM can be applied to previously merged data. This merger was achieved by using the refRMA algorithm [63], designed for large microarray datasets to compute the robust multichip averages. Background adjustment was applied to each array. Quantile normalization was performed on a 909-array training set composed of all HG-U133 Plus 2.0 arrays used in this study. Median polished outputs of the training set was finally used to adjust the normalized gene intensities thus allowing for the integration of data from all three platforms together, limiting results to the 9,409 genes common to these platforms. A merged SAM test was then applied to the combined data of each tissue using the same datasets included in the IV1 test based on the aforementioned parameters (100 permutations and 0 FDR).

As noted above, the IV test is limited to datasets that contain both cancer and normal tissues. The merged SAM method, however, allows for the inclusion of datasets containing solely normal or solely cancer samples. Thus, to test the effect of adding such datasets, microarray samples from all datasets of the same tissue were combined together and another series of SAM analyses were applied using the same test parameters as above. For the purpose of this research, the first set of SAM tests, based on the data from the 19 datasets containing both normal and cancer tissues, is referred to as SAM1 (Figure 8). The second method in which all samples from the 31 datasets could be utilized is denoted as SAM2 (Figure 8). For each tissue, the lists of top 400 differentially

expressed genes from the IV and both SAM tests were selected. These gene lists were used to identify significantly enriched KEGG pathways at a p-value ≤ 0.05 using DAVID Bioinformatics resources [120-121].

3.3.3 Common transcriptional profiles across all five tissue types

To identify consistent changes that are associated with multiple cancer tissue types, an IV1 test was conducted on all 19 Affymetrix datasets containing both cancer and normal samples together, regardless of tissue type. Similarly, a SAM test was performed on the same samples (SAM1) and another SAM test was applied to all 1,768 available Affymetrix samples from the five tissues considered (SAM2). The same test parameters were used as previously mentioned. After determining the genes that behave consistently across all the different cancer types, the top 400 genes were selected from the gene lists produced by each of the methods. Enriched KEGG pathways were identified for all lists at a p-value cutoff of 0.05.

3.3.4 Expanding IV analysis to cDNA data

An additional five datasets using cDNA microarray platforms were obtained from GEO. The datasets contained cancer versus normal samples from colon, kidney and lung tissues for a total of 292 cancer and 169 normal samples. No publicly-accessible data could be found for the other two tissues. The IV analyses for these three tissues as well as the combined tissue test were re-run (IV2; Figure 8) to investigate the cost of excluding these datasets from the merged SAM approach that relies solely on Affymetrix data. Similar test parameters were applied; restricting results to genes with an FDR less than 0.001% and top 400 gene lists were utilized for identifying enriched KEGG pathways, as described above.

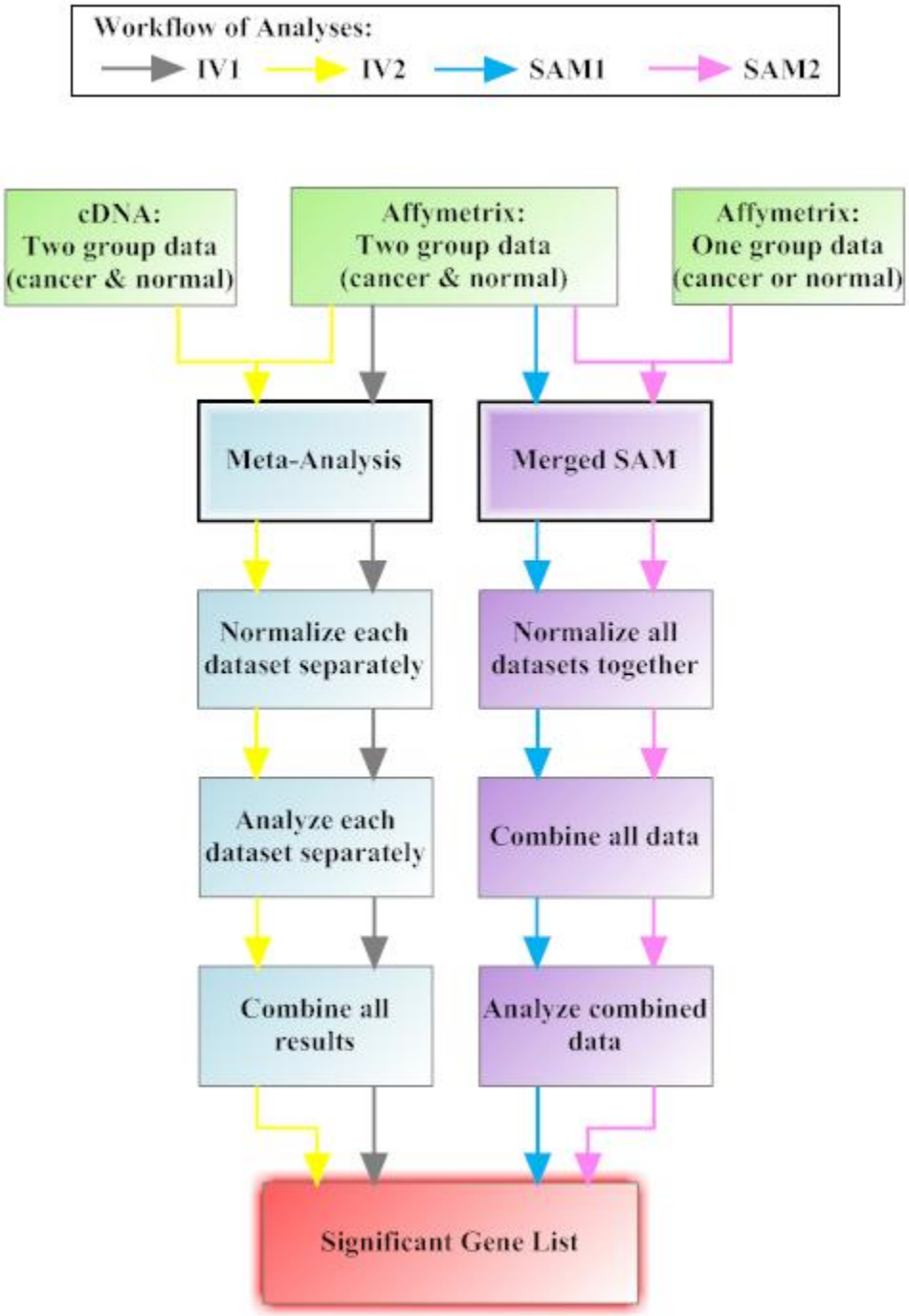

Figure 8 - Analysis Workflow: Flowchart depicting the steps involved in each of the four analyses considered: IV1, IV2, SAM1 and SAM2

3.4 Results

3.4.1 Datasets and approaches

Three different clustering of microarray datasets were used to evaluate (a) the intersection of significant gene lists predicted by meta-analysis and merged SAM methods and (b) compare these predictions with research literature excluding microarray studies. Cluster 1 is composed of Affymetrix microarray datasets containing both cancer and normal samples for five different cancer tissues (Table 1). The gene set predictions resulting from analysis of this data with the use of meta-analysis and merged SAM are denoted as IV1 and SAM1, respectively. Each dataset was analyzed separately for the IV1 test and a final gene list was produced based on the weighted results from the individual datasets. The SAM1 test was applied to the same Affymetrix data from each tissue after their merger, with all samples being normalized together, regardless of dataset. Cluster 2 of microarray datasets used in intersection analysis and literature comparison contained cDNA microarray datasets in addition to the Affymetrix data in Cluster 1. The gene lists predicted by meta-analysis using these datasets were called IV2. Cluster 2 was used to take full advantage of the capability of meta-analysis in integrating microarray datasets from different technologies. Cluster 3 contained asymmetric Affymetrix data in addition to data in Cluster 1 (Table 1). The gene list corresponding to Cluster 3 data predicted by merged SAM is referred to as SAM2. Figure 9 shows the overall characteristics of the Affymetrix datasets used in the analysis. The intersections of the predicted gene lists obtained with the two methods on the three different dataset clusters are summarized in Table 2. The table also presents the p-values corresponding to the intersections based on a hypergeometric test.

Table 1- Overview of Data and Results: Datasets and distribution of microarray samples from the 5 cancer types used. Affyemtrix datasets containing both normal and cancer samples were utilized for the IV1 and SAM1 tests (sample cluster 1), IV2 contained all datasets used in IV1 in addition to all cDNA datasets (sample cluster 2), and all Affyemtrix datasets were merged for SAM2 analysis (sample cluster 3). (Platforms: A: HG-U133A, A2, HG-U133A2, P2: HG-U133 Plus 2)

Tissue	Accession #	Normal	Cancer	Platform
Colon	E-MTAB-57	22	25	A
	GSE4107	10	12	P2
	GSE4183	8	15	P2
	E-MEXP-1224	0	55	A
	E-MEXP-383	0	36	A
	E-TABM-176	55	0	P2
	GSE12945	0	36	A
	GSE17538	0	232	P2
	GSE6988	28	52	cDNA
	Total:	**123**	**463**	
Kidney	E-TABM-282	11	16	P2
	GSE11024†	12	60	P2
	GSE11151	3	57	P2
	GSE14762†	12	10	P2
	GSE15641	23	57	A
	GSE6344	10	10	A
	GSE7023	12	35	P2
	GSE10320	0	144	A
	GSE11904	0	21	A2
	GSE3	81	90	cDNA
	GSE7367	24	24	cDNA
	Total:	**164**	**524**	
Liver	GSE14323	19	47	A/A2
	GSE6764	10	35	P2
	E-TABM-292	0	32	A
	E-TABM-36	0	57	A
	GSE9843	0	69	P2
	Total:	**29**	**240**	
Lung	E-MEXP-231	9	49	A
	GSE10072	49	58	A
	GSE7670	27	27	A
	GSE10445	0	72	P2
	GSE12667	0	75	P2
	GSE2088	30	57	cDNA
	GSE8596	6	69	cDNA
	Total:	**121**	**407**	
Pancreas	E-MEXP-1121†	6	17	A
	E-MEXP-950	11	14	A
	GSE15471	39	39	P2
	GSE16515	15	36	P2
	Total:	**71**	**106**	

† Datasets included replicated samples

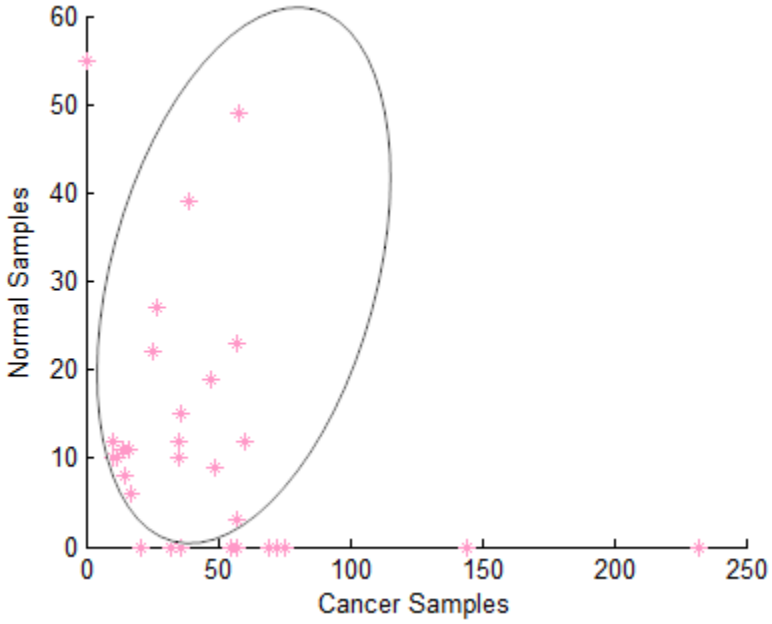

Figure 9 - Overview of Affymetrix Microarray Datasets Used: Distribution of all Affymetrix microarray data used based on the number of cancer versus normal samples in each dataset. Datasets used for IV1/SAM1 test are shown inside the ellipse. Additional datasets included in SAM2 only are located on axes

3.4.2 IV meta-analysis and merged SAM overlap significantly in results

As in previous microarray studies of cancer [40, 122-127], the gene lists produced by the two approaches used in this study indicate significant alterations of the transcriptional profile as the tissue is transformed from the normal to the cancer state, with up to thousands of genes possibly undergoing statistically significant expression changes. While the two methods applied to the three dataset clusters produced different lists of significant genes for each of the five tissues under consideration, there was a considerable overlap in the results (Table 2). The significance of the intersection between predicted gene lists increased consistently as the number of top-ranked genes used in comparison were increased from 10 to 400. In colon tissue, the overlap with IV1 was confined to 338 significant genes instead of 400, since that was the total number of genes passing the test criteria. At the 400 gene level p-values of the IV1/SAM1 intersection ranged from 2.66E-26 in pancreas to 8.42E-181 in lung, while the most significant overlap in IV1/SAM2

was in kidney (p-value = 1.02E-134). Comparison of the results of the two SAM methods produced even larger commonalities in the gene lists identified. Apart from the colon tissue, there was at least 60% overlap between the top 400 gene lists generated by the two SAM methods, for any given comparison. The match between the two SAM results became less pronounced with sharp increases in the number of samples added in SAM2. Nevertheless, even with 506 colon cancer samples included in SAM2 as opposed to the 92 used in SAM1, the overlap between the two methods (176 genes) remained significant. The overlap between IV1 and IV2 varied largely among the top ranked 400 genes with a minimum overlap of 144 genes in lung tissue and a maximum overlap of 355 genes in kidney, resulting in vanishing p-values in the latter case (Table 2).

To identify significantly altered genes across the five considered tissue types, the datasets from all tissues were pooled together. Again, SAM2 included additional datasets with cancer or normal samples only. Similarly, the significance of the overlap between the results increased as more top-ranked genes were considered, with p-values equal to 6.82E-97 and 2.80E-103 for the intersection at the top 400 genes level in IV1/SAM1 and IV1/SAM2, respectively (Table 2).

3.4.3 Cell cycle pathway is commonly enriched in cancers

The cellular pathways that were statistically enriched in the top 400 cancer-associated genes from the multiple tissues under consideration were identified using the DAVID Bioinformatics Resources' functional annotation tool as described in the Methods section. Enriched KEGG pathways common to at least two tissue types within a given test method or significantly associated with the combined 5-tissue comparisons are shown in Figure 10. The cell cycle pathway was statistically enriched in IV1, IV2, SAM1 and SAM2 gene lists across all tissue types (Figure 11). Among the key changes in the cell cycle in normal to cancer transition are the

differential expression of cyclins (A and B) and cyclin-dependent kinases (*CDK1* and *CDK4/6* complex). CDKs are the core of the regulatory apparatus of the cell cycle progression as changes in the kinases and cyclins drive the cell from one stage of the cell cycle to another [128].

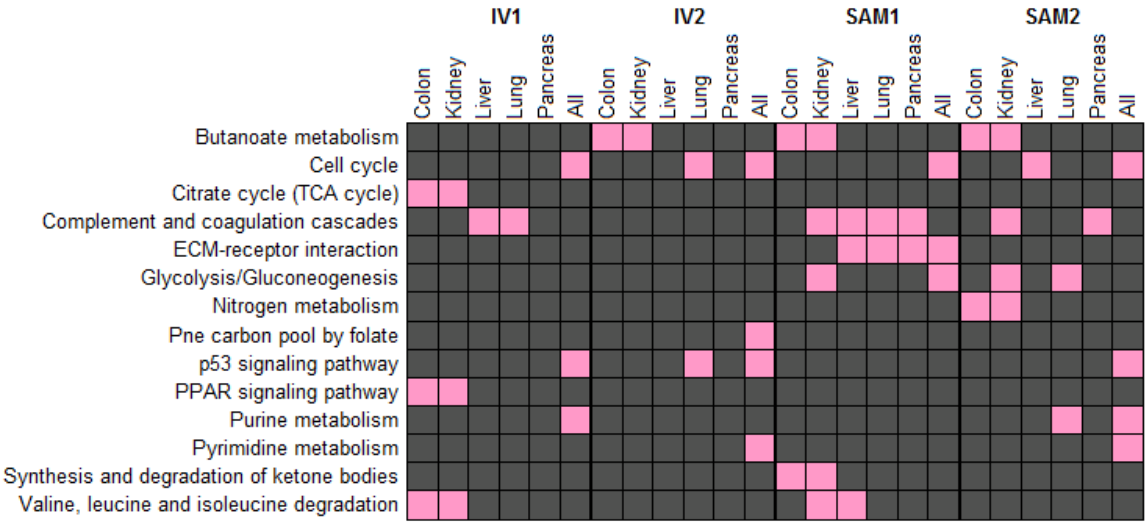

Figure 10 - Enriched KEGG Pathways: A list of KEGG pathways, shown in pink, that appear to be statistically enriched according to the top 400 genes from IV1, IV2, SAM1 and SAM2 at a p-value cutoff of 0.05. Results are limited to pathways independently enriched in at least two of the or in the combined test including all tissues.

In addition, the p53 signaling pathway and purine metabolism were significantly enriched in all-tissue analyses of both IV tests and SAM2. Pyrimidine metabolism is also enriched for the merged SAM2 significant genes while SAM1 genes are associated with ECM-receptor interaction and glycolysis/gluconeogenesis pathways. At the tissue level, some of the metabolic pathways were common to both kidney and colon cancers (butanoate and nitrogen metabolism). Complement and coagulation cascades were enriched in four out of the five tissues under study. These results show that both methods of integration are capable of reproducing a significant portion of the research literature on cellular pathways activated in cancer.

Table 2 - Overlap in Top-Ranked Genes: The overlap among n top-ranked genes between the IV1 and SAM1/SAM2 tests are shown as well as the corresponding p-values of the intersection. Overlaps of top 400 genes between the similar approaches (IV1/IV2 and SAM1/SAM2) are also shown.

IV1 ∩ SAM1

n	Colon		Kidney		Liver		Lung		Pancreas		All	
	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value
10	3	1.01E-07	0	0.989487	5	9.98E-14	2	4.49E-05	0	0.989487	0	0.989487
50	11	8.67E-16	5	5.71E-06	17	7.85E-28	14	1.44E-21	8	1.49E-10	6	2.05E-07
100	26	4.68E-30	23	3.04E-25	24	8.09E-27	34	4.21E-44	17	1.93E-16	18	7.94E-18
200	62	1.88E-57	68	1.57E-66	53	9.78E-45	93	2.56E-109	34	8.96E-22	64	2.00E-60
300	109	2.48E-91	106	4.38E-87	89	1.69E-64	146	5.40E-150	51	7.65E-24	103	6.46E-83
400	132*	3.74E-98	146	1.41E-104	119	7.44E-72	198	8.42E-181	71	2.66E-26	140	6.82E-97

IV1 ∩ SAM2

n	Colon	Kidney	Liver	Lung	Pancreas	All						
	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value
10	3	1.01E-07	0	0.989487	4	1.31E-10	2	4.49E-05	0	0.989487	0	0.989487
50	12	1.17E-17	5	5.71E-06	12	1.17E-17	8	1.49E-10	8	1.49E-10	5	5.71E-06
100	32	1.97E-40	23	3.04E-25	24	8.09E-27	28	2.09E-33	17	1.93E-16	21	3.50E-22
200	67	5.51E-65	66	1.88E-63	43	5.97E-32	69	4.34E-68	34	8.96E-22	65	6.22E-62
300	111	3.32E-94	116	1.54E-101	60	4.00E-32	101	3.52E-80	51	7.65E-24	101	3.52E-80
400	124*	9.02E-88	168	1.02E-134	86	1.19E-38	149	1.67E-108	71	2.66E-26	145	2.80E-103

IV1 ∩ IV2

n	Colon	Kidney	Liver	Lung	Pancreas	All						
	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value
400	163*	1.39E-186	355	0	No data	-	144	3.97E-140	No data	-	280	0

SAM1 ∩ SAM2

n	Colon	Kidney	Liver	Lung	Pancreas	All						
	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value	Overlap	P-Value
400	176	1.92E-146	284	0	253	6.86E-281	241	3.15E-257	No data	-	262	2.34E-299

* Only 338 genes are used for colon IV1

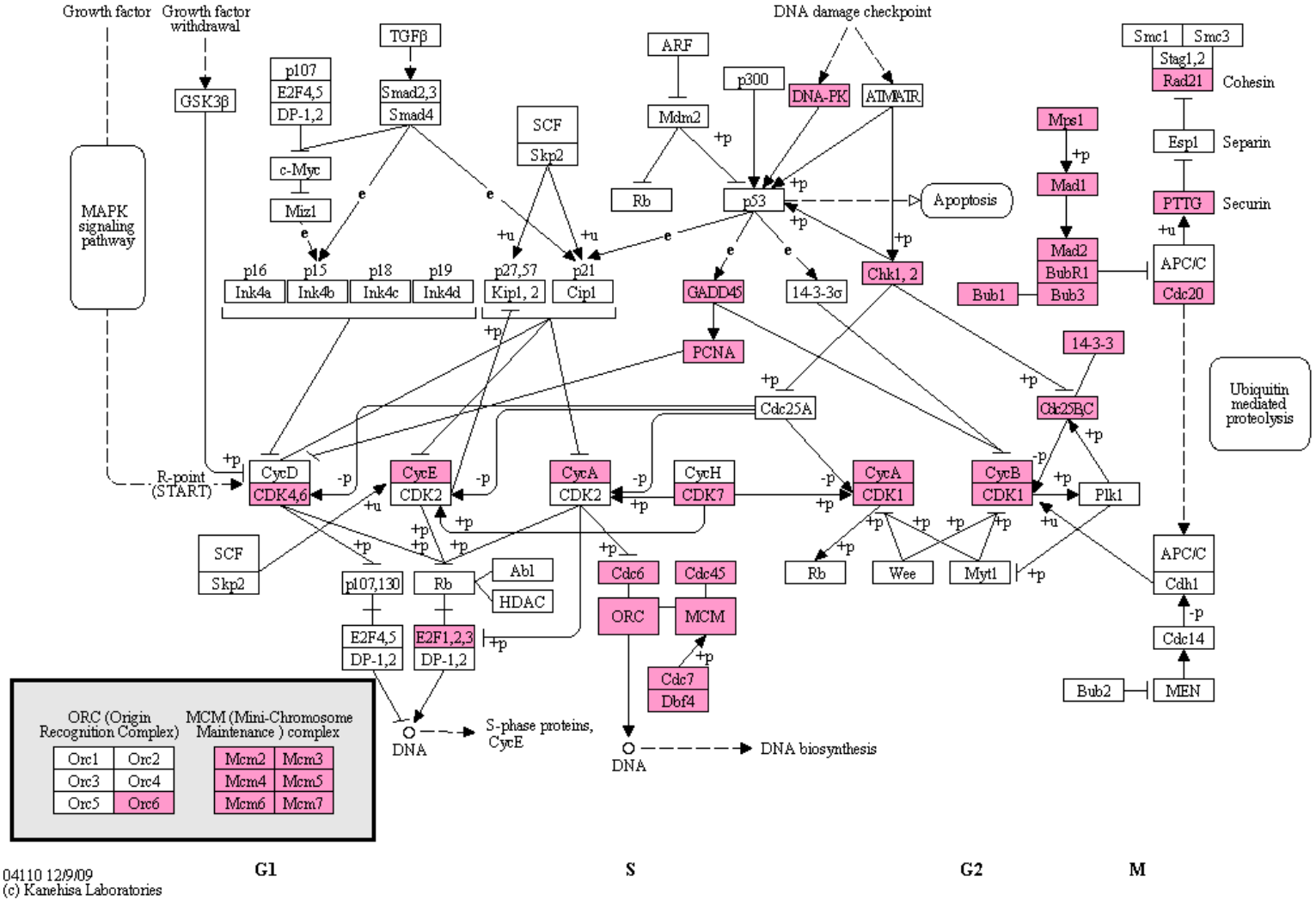

Figure 11 - Cell Cycle Pathway: Differentially expressed genes involved in the cell cycle are shown in pink. Genes are ranked among the top 400 genes by at least one of the statistical approaches used (IV1, IV2, SAM1 and/or SAM2), based on analyses of all five tissues together.

3.4.4 Microarray results match cancer research literature with low p-values

Next, the SAM1, SAM2, IV1, and IV2 top four hundred gene lists were tested for PubMed hits associated with cancer. An automated PubMed abstract search was conducted for the genes in the aforementioned lists, excluding those abstracts that belonged to microarray-based research. Also excluded were abstracts that did not contain the word "cancer". A gene had to have at least one such PubMed abstract match to be considered as a literature search hit. The number of successful hits produced from the merged SAM methods and the IV tests intersected the research literature with significantly higher coverage than would be expected for randomly generated gene lists (Figure 12). The p-values shown in Figure 12 for the top 300 and 400 genes for all three methods were computed by using control gene lists obtained from the same Affymetrix platforms by randomly selecting lists of equal size (300 or 400) and averaging the number of hits over 100 iterations. The p-values for each tissue were then calculated using a normal distribution given the mean and standard deviation parameters of the randomly generated data. The p-value for the colon IV1 in the top 400 gene list was adjusted to a hundred iterations of 338 randomly chosen genes to account for the maximum available number of genes. The merged SAM methods produced gene lists that matched the research literature more accurately than the gene lists produced by the IV tests in four out of the five tissues under consideration. Both SAM1 and SAM2 also produced more significant p-values per tissue than the average p-value obtained from the SAM tests performed on the individual datasets for a given tissue (data not shown). The addition of single sample-type datasets resulted in fewer literature-associated gene lists than the SAM1 approach; however, the results improved when considering the top 400 genes as opposed to the top 300. Note also that PubMed hits on gene lists presented by meta-analysis and merged SAM approaches fell inside and outside the intersections. For example, the case of colon cancer in IV1 and SAM1 gene lists. There were 93 hits on IV1 ∩ SAM1 ($p = 1.19E-07$), 103 hits on IV1 - IV1 ∩ SAM1 ($p = 5.09E-02$); and 205 hits on SAM1 - IV1 ∩ SAM1 ($p = 2.32E-23$).

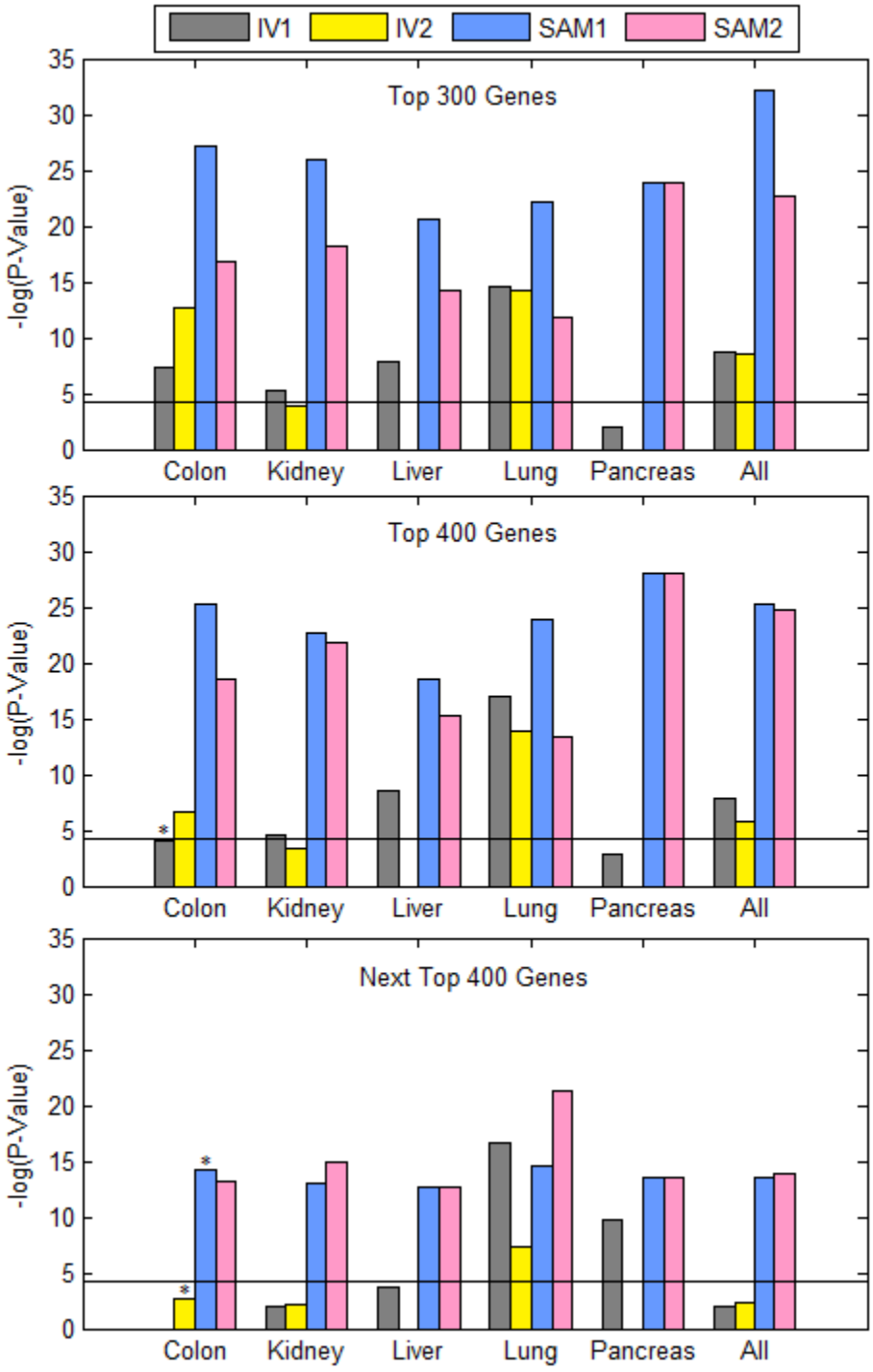

Figure 12 - Literature Search Results: Histogram representing p-values of the number of top-ranked genes with at least 1 PubMed abstract relating the genes to cancer research from a non-microarray study according to each of the four test procedures: IV1 (gray), IV2 (yellow), SAM1 (blue) and SAM2 (pink). P-values are calculated based on expected data from a hundred random gene lists obtained from the platform and similarly related to non-microarray cancer literature. . The horizontal line represents a p-value cutoff of 0.0001.
* P-values adjusted to maximum number of available top genes.

As an additional control, the next top 400 genes (ranks 401-800) in each list, if available, were subjected to a similar PubMed abstract search. The p-values representing the results revealed decreased literature coverage of these genes compared to the first top 400 genes in all cases except for SAM2 results in lung tissue. In this test, the majority of the IV results (except for lung and pancreas) dropped below the 0.0001 p-value threshold marked by a horizontal black line in Figure 12.

3.5 Discussion

Meta-analysis approaches applied to microarray data aim to increase the statistical power of the results as well as to increase the reproducibility of individual studies [46]. Typical meta-analysis approaches combine results of independent datasets to produce a generalized outcome across these datasets. Meta-analysis approaches require both perturbed and control data within the same microarray datasets under consideration. However, the recent dramatic increase in publicly accessible microarray samples is mainly due to datasets containing no data on normal tissue. Noting that microarray samples on normal tissue are available in other public datasets, the idea of picking samples from different datasets obtained with same/similar microarray chips and normalizing them together before the identification of significantly altered genes in normal to cancer comparison was explored. The resulting merged SAM sacrifices the use of data from other platforms. However, it could be potentially useful for integrated analysis of cancer microarray datasets for which much of the available data is highly asymmetric.

One reason for asymmetry in the current public access microarray data is that the goals of global gene expression quantification in cancer research shifted towards identifying significant genes associated with cancer subtypes [39, 129-132]. The merged SAM analysis presented here is applicable to any microarray inquiry where there is a perturbed state (say cancer subtype 1) and

control state (cancer subtype 2). The method of integration was applied to cases where there was plenty of data for both meta-analysis and merged data approaches. Even when one aims to uncover differences in gene expression profiles between two cancer subtypes, it is often useful to consider such differences between subtypes and control normal tissue samples [122]. Such triple comparisons reveal the original basis for the subtype differences that stem from normal to cancer transformations.

A quick study of the GEO database clearly shows that microarray data for hormone-associated solid cancers such as breast, prostate and ovarian cancers are highly asymmetric. The more recent datasets increasingly come from studies for which one cancer subtype is compared to another cancer subtype and as a result contain no data from normal samples. The five tissue types presented were chosen in this study because of the availability of data that could be used for both merged SAM and meta-analysis approaches. Previous studies have addressed the possible problems that arise from combining data across different technologies [133-134]. We have used the datasets obtained with similar chips to compare the performance of meta-analysis and merged the SAM approaches. The direct integration of data preceding the analysis as in the case of the merged SAM overcomes the problems associated with small sample sizes in individual studies. While data merging across similar chips sacrifices the inclusion of some of the genes not common to all platforms, it provides additional robustness since all samples are normalized together as opposed to being normalized separately per dataset [135].

The meta-analysis and merged SAM approaches yielded significant gene lists with intersecting common gene subsets that could not be plausibly obtained by chance. Both approaches matched automated PubMed abstract searches of research literature (excluding microarray studies) with very low p-values for random occurrence. However, the merged SAM approach replicated the

existing literature much more accurately than the meta-analysis approach in five of the six cases under study. Addition of cDNA arrays into meta-analysis resulted in reduced overlap with the cancer literature. Meanwhile, the inclusion of asymmetric datasets also produced slightly less statistically significant results in merged SAM analyses; nevertheless, the approach still generated results that were at least as significant as the meta-analyses, again surpassing meta-analysis in five out of the six cases. Despite the addition of hundreds of samples from asymmetric sets, the merged SAM continued to perform well, matching literature as well as results of symmetric microarray data. Moreover, the match between microarray lists and the literature became less pronounced as lesser-ranked significant genes (401 – 800) were used in the comparison. The gene lists obtained in all the tests were further validated by associating them with functional annotation through KEGG pathways. While individually each tissue possessed a unique list of pathways and processes with which it was associated, overall, cell division appeared to be the common driving factor to all tissues, as would be expected.

The automated text searches was used as an instrument for validation of the prediction value of the two different approaches to integrating microarray data associated with cancer. Typical validation used in microarray analysis for illustrating relevance of gene list to disease state under consideration is usually via partitioning the dataset into learning, testing/validation subsets in a supervised learning approach [136-138]. However, it is relatively easy to differentiate between cancer and normal tissue with a variety of gene sets, but in many cases, such sets are laboratory specific [139]. Research literature in cancer is rich with data on genes associated with this disease and the bulk of such data was collected by using research tools other than microarrays, and therefore, automated text search constituted an independent means of validating the microarray results.

PubMed hits on gene lists produced by meta-analysis and merged SAM approaches fall on the intersections of such lists as well as outside the intersections, suggesting the use of both approaches whenever data is available. The top ranked 400 genes in both cases are highly statistically enriched with PubMed hits and for which the intersection between the two approaches had typically the lowest p-value. When considering the role of well studied genes such as hub genes or genes in public access cellular pathways, projecting both gene lists onto known pathways to generate new hypothesis for experimental verification is a straightforward process. The merged SAM technique provides a unique opportunity to obtain a candidate list for genes associated with a perturbed state in cases where the public microarray data is largely asymmetric.

3.6 Conclusion

Typical meta-analysis approaches allow for the use of various platforms at the expense of utilizing large amounts of data that exist in datasets containing either normal or cancer tissues only. The merged SAM approaches in the study were shown to reproduce much of the known cancer literature while effectively being applied to asymmetrical microarray datasets. Hence, this approach can be extended and applied to various other diseases. While many of the genes in these lists have already been associated with cancer, the merged SAM approach sheds light on new genes that could play a pivotal role in cancer pathogenesis.

Chapter 4: Large-scale integration of microarray data reveals genes and pathways common to multiple cancer types

4.1 Summary

This chapter discusses the commonalities in aberrant gene expression that are shared by cancers arising in different tissue types through the use of microarray data. The global gene expression analysis of cancer and healthy tissues typically results in large numbers of significantly altered SAM genes. Such data, however, has been difficult to interpret due to the high level of variation of gene lists across laboratories and the small sample sizes used in individual studies. For this research, the compiled microarray data was obtained from 84 laboratories using samples that were hybridized on the same platform family, resulting in a database containing 1043 healthy tissue samples and 4900 cancer samples for 13 different tissue types. The primary cancers considered were adrenal gland, brain, breast, cervix, colon, kidney, liver, lung, ovary, pancreas, prostate and skin and stomach tissues. The data was normalized together and analyzed in subsets for the discovery of genes involved in normal to cancer transformations. This integrated approach produced top 400 ranked SAM gene lists for each of the thirteen cancer types. These lists were highly statistically enriched with genes already associated with cancer in research publications excluding microarray studies ($p < 1.31 E - 12$). The genes *MTIM* and *RRM2* appeared in nine and *TOP2A* in eight lists of significantly altered genes in cancer. In total, there were 132 genes present in at least four gene lists, eleven of which had not been previously associated with cancer. The list contains 17 metal ion and 15 adenyl ribonucleotide binding proteins, 6 kinases and 6 transcription factors. These results point to the value of integrating microarray data in the study of combination drug therapies targeting metastasis.

4.2 Background

Tens of thousands of microarray samples have accumulated in public access databases in the last decade [96-98]. A large portion of such data is cancer-specific and therefore holds the promise of cancer-associated gene discovery based on thousands of samples (not tens or hundreds). Much of the cancer-associated microarray data in public domains comes without control samples. In fact, the data in GEO is highly asymmetric, containing datasets with cancer microarray samples only and other datasets containing samples for healthy tissues but not cancer tissues. Conventional meta-analysis approaches of integrating data, where laboratory results are combined after the datasets are independently analyzed, would not be useful in drastically increasing the sample sizes in microarray analysis of cancer. Such analyses require the presence of both cancer and normal tissue samples in the same microarray dataset.

In this study, a large-scale approach was used to integrate microarray data from multiple laboratories by normalizing them together and then using the Significance Analysis of Microarray (SAM) method [75]. This allowed for the identification of list of genes that are significantly altered in cancer compared to normal, specific for thirteen distinct tissues. This methodology is grounded on previous studies that revealed the predictive potential of integrated microarray data. Large-scale meta-analysis techniques applied to cancer have already been adopted by a few groups [47, 84, 113-114], focusing on a single tissue type. Other studies merged all cancer microarray data regardless of tissue type into one group and controls into another [48, 116] to identify gene sets associated with common cancer mechanisms. The merge SAM approach is unusual when compared to the typical meta-analysis methods but it allows for the integration of asymmetric microarray data for global gene expression. The various previously used methods reflect the purpose of the study undertaken, and despite the paper trail on the general methodology used, the question still arises as to the validity of the results from this research.

This study addresses the question regarding the extent to which the currently available microarray data has the potential to replicate the research literature on the molecular mechanisms of cancer. The automated text search algorithms utilized point to high-level coincidence between the generated gene lists and the cancer-associated genes determined from the non-microarray research literature.

Using nearly 6,000 microarray samples, this study identifies 132 genes that are highly significantly associated with at least four distinct cancer types. This research also presents a set of 270 genes that appear to be highly significant in comparisons of datasets consisting of cancer and normal tissues independent of tissue type. These two sets have 74 genes in common and will potentially contribute to a more detailed annotation of the genes in the cancer bioinformatics databases. This study points to the value of large-scale compilation of microarray data in cancer research, as the inclusion of large amounts of microarray data from different labs helps eliminate the effects of lab-specific noise to increase the reliability of the results [107].

4.3 Materials and Methods

4.3.1 Microarray dataset selection and normalization

An Affymetrix microarray database was constructed for normal and cancer samples obtained from 13 different solid tissues. The tissues considered were: adrenal gland, brain, breast, cervix, colon, kidney, liver, lung, ovary, pancreas, prostate, skin and stomach. The microarray data contained a total of 4,900 cancer and 1,043 normal tissue samples acquired from 84 labs. All the data was obtained from the publically accessible Gene Expression Omnibus [96-97] and Array Express [98] online repositories. The inclusion criteria restricted the use of datasets hybridized specifically on one of the three comparable Affymetrix platforms (HG-U133A, HG-U133A 2.0, and the HG-U133 Plus 2.0), where raw data CEL files were available, with at least 20 usable

microarray samples. In addition, the results from the datasets should have been previously published in a peer-reviewed study. No differentiation was made with respect to the different malignancies obtained from the same tissue.

The data was normalized using the refRMA algorithm [63], utilizing the platform-compatible custom ENTREZG CDF files (version 12) [117] in order to obtain Entrez gene intensities. Background adjustment was applied and quantile normalization was performed on a 909-array training set that was then applied to compute the probe-level quantiles for the remaining data. Median polishing of the training set was finally used to adjust the normalized probe intensities of the remaining data. Data was then filtered to remove the genes not shared by the three platforms. Finally, the gene intensities of replicate samples obtained from the same source were averaged across replicates. All data pre-processing was performed in MATLAB [62].

4.3.2 Differential gene expression

The differential expression of genes between cancer tissues and the corresponding controls was investigated using the Significance Analysis of Microarrays by utilizing the *samr* package [119] in R [118]. The SAM test was applied individually to the microarray datasets specific to each of the thirteen tissues under consideration. For each SAM test, a hundred random iterations were performed and the false discovery rate (FDR) was constrained to zero.

In addition, a general normal versus cancer test was conducted. In order to avoid over-representation and dominance of certain tissues, ten arrays were randomly chosen from both the normal and tumor samples of each tissue to produce two datasets (cancer, control) for SAM analysis. The number ten was determined by the smallest sample size available for any tissue

(adrenal normal tissue). SAM genes were then identified following the aforementioned criteria. Again, the random selection and differential expression process was repeated a hundred times.

4.3.3 Functional annotation of top ranked and conserved genes

The lists of top 400 SAM genes were obtained for each of the 13 tissues. The cutoff (top 400) was chosen in order to optimize the match between the predicted SAM lists and lists of cancer associated genes obtained by via automated text search of non-microarray PubMed abstracts. An enriched KEGG pathway profile was produced for each of the 13 tissues individually, at a p-value ≤ 0.05 using DAVID Bioinformatics resources.

4.3.4 Consistent differential expression across tissues

Among the top 400 gene lists provided for each tissue, a subset of genes that were consistently differentially expressed was determined. These genes were selected provided they appeared to be significantly altered in at least 4 of the 13 tissues. Moreover, the top 400 genes from the general normal-cancer comparisons were obtained for each of the 100 iterations. The frequency of occurrence of each of the genes appearing in any of the lists was calculated to determine those genes whose changes in expression were most concordant. The results of the two approaches were then compared.

4.3.5 Cancer literature annotation of identified significant SAM genes

To determine which genes from the SAM lists were known to be associated with cancer, an automated text search was performed. For all the genes in the microarray platform, a search of the gene symbol and cancer was conducted in PubMed abstracts. The results were limited to non-microarray literature. In addition, all literature papers associated with each of these genes as provided by the NCBI ftp site were obtained. A list of PubMed IDs of all cancer non-microarray

literature were then acquired and was used to further determine which genes had been previously associated with cancer. Results from the two approaches were combined to provide a comprehensive coverage of the known cancer literature. The SAM lists were then annotated with these results, identifying those genes that were cited in relation to cancer at least once from those that had no cancer association. As a control, a hundred random gene lists from the same platform of equal size to the SAM lists under consideration were obtained. The number of cancer-related genes in each iteration was determined, and the mean and standard deviation were calculated from these values to obtain the parameters of a normal distribution. The expected value and the standard deviation were then used to compute the p-values for the significant association of each of our cancer gene lists with the known non-microarray literature.

4.4 Results

4.4.1 Dataset

Nearly six thousand microarray samples were used to identify significant gene lists involving normal to cancer transformations in 13 distinct human tissue types. The distribution of samples across each tissue is shown in Table 3. Overall, there were 4900 cancer samples and 1043 normal tissue microarray samples. The largest sample sets in the database belonged to breast, brain, colon, and kidney. Sample distributions were asymmetric, with many more cancer samples than normal tissue samples. Moreover, in order to increase sample sizes, datasets with only cancer or only normal tissue samples were added to the large-scale datasets. This approach eliminated the use of microarray meta-analysis where each dataset is normalized and analyzed separately. On the other hand, the merged SAM analysis used here best fits the recent trend of asymmetric growth in cancer samples in public-access microarray data. Restriction of analysis to comparable microarray chips allowed for the normalization and analysis of samples in an integrated fashion without significantly reducing the number of samples that could be used in the analysis.

Table 3 - Data Summary: Dataset accessions and number of normal and cancer microarray samples used for each of the 13 tissue analyses

Tissue	Accession #	Normal	Cancer	Total
Adrenal Gland	GSE10927	10	33	43
	E-TABM-311	0	34	34
	Total:	**10**	**67**	**77**
Brain	GSE12907	3	21	24
	GSE13041	0	175	175
	GSE11882	173	0	173
	GSE4271	0	100	100
	GSE3790	87	0	87
	GSE4412	0	85	85
	GSE5675	0	41	41
	GSE2817	0	30	30
	GSE17612	23	0	23
	Total:	**286**	**452**	**738**
Breast	GSE10780	143	42	185
	GSE10797	10	56	66
	GSE3744	7	40	47
	E-TABM-276	13	18	31
	GSE5764	20	10	30
	GSE16873	12	12	24
	E-MEXP-882	4	19	23
	GSE8977	15	7	22
	GSE4922	0	289	289
	GSE2034	0	286	286
	GSE11121	0	200	200
	GSE7390	0	198	198
	GSE1456	0	159	159
	GSE2603	0	99	99
	GSE6532	0	73	73
	GSE5327	0	58	58
	GSE5847	0	55	55
	GSE1561	0	49	49
	GSE12276	0	48	48
	GSE12763	0	30	30
	GSE6596	0	24	24
	GSE13787	0	23	23
	Total:	**224**	**1795**	**2019**
Cervix	GSE9750	21	33	54
	GSE7803	10	21	31
	GSE6791	8	20	28
	GSE5787	0	33	33
	Total:	**39**	**107**	**146**
Colon	E-MTAB-57	22	25	47
	GSE4183	8	15	23
	GSE4107	10	12	22
	GSE17538	0	232	232
	E-TABM-176	55	0	55
	E-MEXP-1224	0	55	55
	GSE12945	0	36	36
	E-MEXP-383	0	36	36
	Total:	**95**	**411**	**506**

Table 3 (continued)

Kidney	GSE15641	23	57	80
	GSE11151	3	57	60
	E-TABM-282	11	16	27
	GSE14762	12	10	22
	GSE6344	10	10	20
	GSE10320	0	144	144
	GSE11024	0	60	60
	GSE7023	0	35	35
	GSE11904	0	21	21
	Total:	**59**	**410**	**469**
Liver	GSE14323	19	47	66
	GSE6764	10	35	45
	GSE9843	0	69	69
	E-TABM-36	0	57	57
	E-TABM-292	0	32	32
	Total:	**29**	**240**	**269**
Lung	GSE10072	49	58	107
	E-MEXP-231	9	49	58
	GSE7670	27	27	54
	GSE12667	0	75	75
	GSE10445	0	72	72
	Total:	**85**	**281**	**366**
Ovary	GSE6008	4	99	103
	GSE18520	10	53	63
	GSE9891	0	189	189
	GSE14764	0	80	80
	E-MEXP-935	0	27	27
	GSE9455	0	20	20
	Total:	**14**	**468**	**482**
Pancreas	GSE15471	39	39	78
	GSE16515	15	36	51
	E-MEXP-950	11	14	25
	E-MEXP-1121	6	17	23
	Total:	**71**	**106**	**177**
Prostate	GSE6956	18	69	87
	E-TABM-26	13	44	57
	GSE17356	0	27	27
	GSE2443	0	20	20
	Total:	**31**	**160**	**191**
Skin	GSE7553	5	82	87
	GSE13355	64	0	64
	GSE8401	0	31	31
	Total:	**69**	**113**	**182**
Stomach	GSE13911	31	38	69
	GSE15460	0	229	229
	GSE8167	0	23	23
	Total:	**31**	**290**	**321**
Overall	**Total:**	**1043**	**4900**	**5943**

4.4.2 SAM genes and their match with research literature

The SAM gene lists obtained for the thirteen distinct human tissues by setting the false discovery rate to zero varied in length depending on the tissue. However, the top 400 genes in each list matched well with the cancer-associated gene literature obtained from experiments excluding microarrays (Table 4). The automated text search algorithm described in the methods section showed that nearly 80% of the genes in these lists were previously associated with cancer in non-microarray studies. P-values for the occurrence of these matches by chance were estimated by generating randomly chosen gene lists from the microarray chip, and varied from a low of 2.86 E-33 for adrenal tissue to 6.99 E-12 in brain tissue. Next, genes that occurred in multiple tissue-specific lists were selected and their match with the literature was similarly difficult to explain by chance events. These results indicate the potential of microarray studies based on large sample sizes to regenerate much of the known literature associated with cancer. The choice of top 400 as a cut off is somehow arbitrary, however, results indicated that the match between microarray predictions and literature was nearly optimal at this particular cutoff value (data not shown).

4.4.3 Cellular pathways enriched for top 400 SAM genes

The top 400 SAM gene lists from all tissue types were projected onto KEGG [101] cellular pathways to evaluate their statistical enrichment using DAVID [120-121]. Results shown in Figure 13 indicate the statistically enriched cellular pathways previously associated with cancer such as the glycine, serine, and threonine metabolism, PPAR signaling pathway, DNA replication, and ECM-receptor interaction. The variation in the catalog of enriched pathways from tissue to tissue is a reflection of the tissue-specific dimensions of cancer. It must also be noted that pathways not enriched for some tissues and enriched for others still included considerable amounts of SAM genes even for those tissues in which the pathway p-values were not significant.

Table 4 - Overview of Results: Number of significant genes among the top 400 genes for the 13 tissues appearing at least in one (T 400), two (T2 400) or three (T3 400) tissues. Also shown are the number and corresponding percentages and p-values of these gene that have been associated with cancer in the non-microarray literature found in PubMed (PM) abstracts

	T 400	PM 400	(%)	P-Value	T2 400	PM2 400	(%)	P-Value	T3 400	PM3 400	(%)	P-Value
Adrenal	400	343	85.8	2.86E-33	235	203	86.4	8.26E-15	132	119	90.2	4.26E-14
Brain	400	300	75.0	6.99E-12	277	237	85.6	8.54E-19	154	135	87.7	1.63E-12
Breast	400	344	86.0	6.53E-34	153	117	76.5	2.76E-05	56	45	80.4	3.26E-03
Cervix	400	335	83.8	2.21E-28	234	202	86.3	6.01E-14	106	96	90.6	1.39E-10
Colon	400	324	81.0	2.45E-22	243	208	85.6	3.91E-15	130	117	90.0	1.74E-11
Kidney	400	333	83.3	3.18E-27	230	194	84.3	4.86E-12	126	112	88.9	1.57E-10
Liver	400	328	82.0	1.91E-24	194	161	83.0	1.73E-11	91	83	91.2	5.21E-09
Lung	400	325	81.3	7.44E-23	224	193	86.2	2.02E-14	119	105	88.2	2.41E-08
Ovary	400	323	80.8	7.92E-22	231	198	85.7	3.24E-15	115	106	92.2	3.81E-15
Pancreas	400	337	84.3	1.45E-29	229	202	88.2	6.28E-16	110	99	90.0	7.26E-11
Prostate	400	307	76.8	1.52E-14	197	148	75.1	8.67E-06	76	60	78.9	6.77E-04
Skin	400	324	81.0	2.45E-22	247	206	83.4	3.78E-15	127	107	84.3	1.91E-07
Stomach	400	302	75.5	1.31E-12	184	144	78.3	4.81E-09	72	59	81.9	5.80E-05

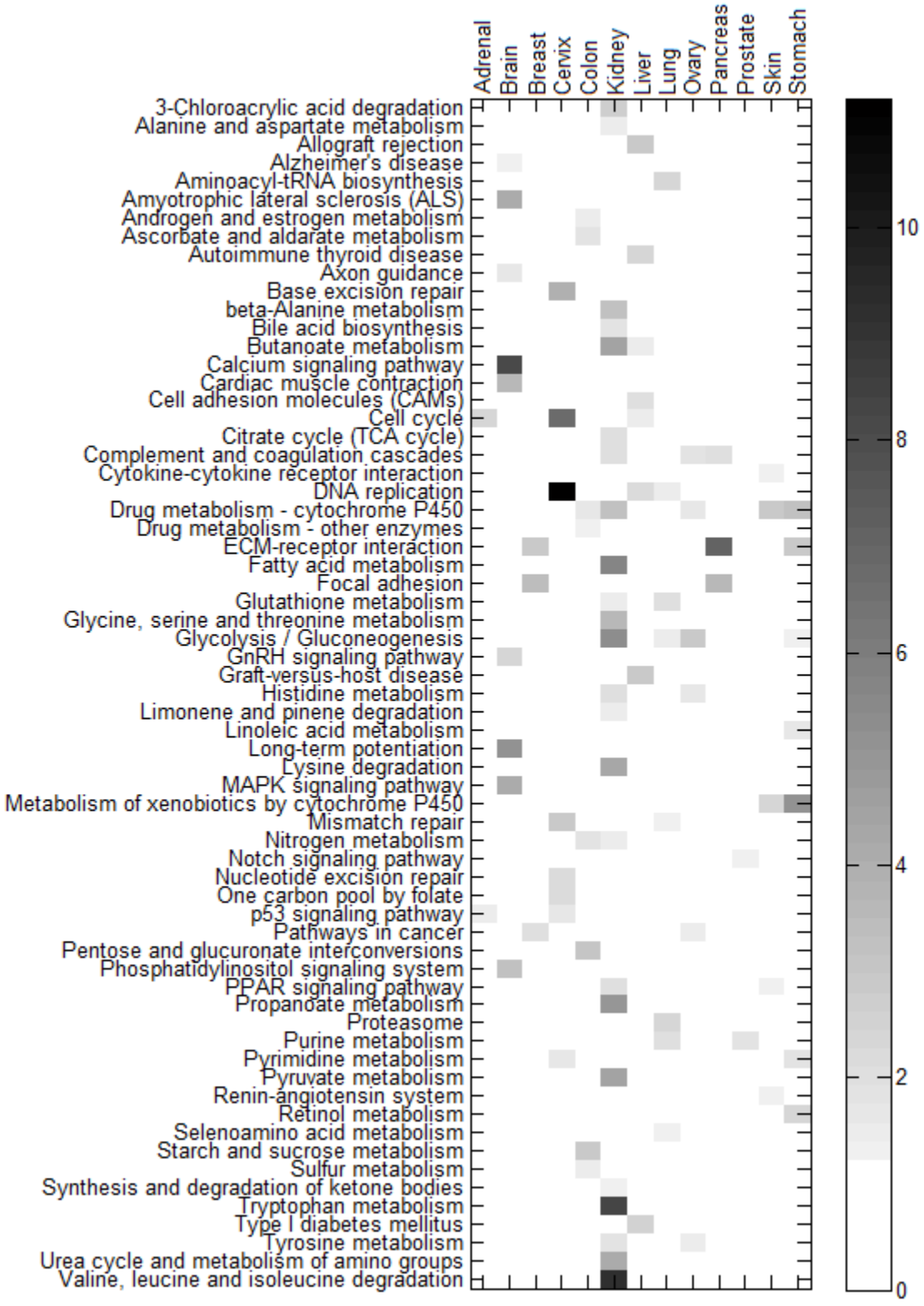

Figure 13 - Pathway Profiles of Different Cancer Tissues: Heat map showing the significant pathway profiles for each of the thirteen cancer tissues considered. The color-scale represents the –log of the p-value for the pathway enrichment using a p-value cutoff of 0.05

4.4.4 SAM genes in multiple gene lists

A total of 132 genes appeared in at least four of the top 400 SAM genes out of the 13 total tissue types considered. All, with the exception of 11 genes were previously affiliated with cancer in the non-microarray based research literature. These genes are listed in Table 5a along with the affiliated tissue types in which they appeared among the top 400 SAM genes. The table also identifies the up- and downregulation of the genes in each cancer tissue and annotates approved and experimental drugs targeting some of these genes as obtained from DrugBank [140-141]. The genes *MTIM* and *RRM2* appear in nine and *TOP2A* appears in eight out of the thirteen tissues. These genes are followed in the list by genes that appear in at least seven cancer types: *ADH1B*, *CDC20*, *CFD*, *GSTM5*, *CLEC3B*, *PRC1*, and *MELK*, *ABCA8*, *UBE2C*, *KIF4A*, and *RACGAP1*. Among this list, *TOP2A* is currently targeted by seven approved drugs (Table 5b). The gene *EPHX2* is targeted by tamoxifen in the treatment of breast cancer, and *ESSRG* by Diethylstilbestrol for prostate cancer. Meanwhile, experimental drugs targeting *CDC2* and *TUBA1B* are going through approval processes.

Shown in Table 6a are those top 400 SAM genes found in at least four lists but have not been previously associated with cancer. The gene *LPCAT1* appears in the lists for cervix, colon, kidney, pancreas and stomach. This enzyme mediates conversion of LPC to PC, thereby playing a pivotal role in respiratory physiology. Among, the genes in Table 6a are found in four SAM gene lists out of the thirteen tissue types under study, only *BBOX1* was associated with approved drug targets. Further studies are needed to annotate the potential roles of these genes in the progression of cancer.

Table 5 - Annotation of Commonly Altered Genes: a) List of genes differentially expressed in at least 4 tissues and have been previously associated with cancer in non-microarray literature. The tissues in which the genes are altered are shown where regular font indicated upregulation and italics represents downregulation in cancer compared to normal tissue. Entrez IDs shown in bold represent genes that appeared to be significant in the general normal/cancer comparisons. b) List of approved and experimental cancer drugs targeting commonly altered genes

a)

Entrez ID	Gene Symbol	Gene Name	Tissues
4499	*MT1M*	Metallothionein 1M	*Adrenal*, *Breast*, *Colon*, *Kidney*, *Liver*, *Lung*, *Pancreas*, *Prostate*, *Stomach*
6241	*RRM2*	Ribonucleotide reductase M2 polypeptide	Adrenal, Breast, Cervix, Colon, Kidney, Liver, Lung, Pancreas, Prostate
7153	*TOP2A*	Topoisomerase (DNA) II alpha 170kDa	Adrenal, Breast, Cervix, Kidney, Liver, Lung, Ovary, Pancreas
125	*ADH1B*	Alcohol dehydrogenase 1B (class I), beta polypeptide	*Adrenal*, *Breast*, *Colon*, *Kidney*, *Lung*, *Ovary*, *Skin*
991	*CDC20*	Cell division cycle 20 homolog (S. cerevisiae)	Adrenal, Breast, Cervix, Liver, Lung, Ovary, Pancreas
1675	*CFD*	Complement factor D (adipsin)	*Adrenal*, *Breast*, *Cervix*, *Colon*, *Ovary*, *Prostate*, *Skin*
2949	*GSTM5*	Glutathione S-transferase M5	*Adrenal*, *Breast*, *Cervix*, *Lung*, *Ovary*, *Prostate*, *Skin*
7123	*CLEC3B*	C-type lectin domain family 3, member B	*Adrenal*, *Breast*, *Colon*, *Kidney*, *Lung*, *Prostate*, *Skin*
9055	*PRC1*	Protein regulator of cytokinesis 1	Adrenal, Cervix, Kidney, Liver, Lung, Pancreas, Prostate
9833	*MELK*	Maternal embryonic leucine zipper kinase	Adrenal, Breast, Cervix, Colon, Lung, Ovary, Pancreas
10351	*ABCA8*	ATP-binding cassette, sub-family A (ABC1), member 8	*Breast*, *Cervix*, *Colon*, *Kidney*, *Lung*, *Ovary*, *Skin*
11065	*UBE2C*	Ubiquitin-conjugating enzyme E2C	Adrenal, Breast, Cervix, Colon, Liver, Ovary, Skin
24137	*KIF4A*	Kinesin family member 4A	Adrenal, Breast, Cervix, Colon, Liver, Ovary, Skin
29127	*RACGAP1*	Rac GTPase activating protein 1	Adrenal, Cervix, Kidney, Liver, Lung, Ovary, Pancreas
316	*AOX1*	Aldehyde oxidase 1	*Adrenal*, *Kidney*, *Ovary*, *Pancreas*, *Prostate*, *Skin*
701	*BUB1B*	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)	Adrenal, Cervix, Kidney, Liver, Lung, Pancreas
4306	*NR3C2*	Nuclear receptor subfamily 3, group C, member 2	*Breast*, *Colon*, *Kidney*, *Ovary*, *Pancreas*, *Skin*
4674	*NAP1L2*	Nucleosome assembly protein 1-like 2	*Brain*, *Breast*, *Colon*, *Kidney*, *Ovary*, *Skin*
4886	*NPY1R*	Neuropeptide Y receptor Y1	*Adrenal*, *Colon*, *Kidney*, *Liver*, *Ovary*, *Skin*
6696	*SPP1*	Secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1)	Adrenal, Cervix, Colon, Lung, Skin, Stomach
6790	*AURKA*	Aurora kinase A	Adrenal, Cervix, Colon, Kidney, Liver, Lung
9133	*CCNB2*	Cyclin B2	Adrenal, Breast, Cervix, Liver, Lung, Ovary
9232	*PTTG1*	Pituitary tumor-transforming 1	Adrenal, Cervix, Colon, Ovary, Pancreas, Skin
9314	*KLF4*	Kruppel-like factor 4 (gut)	*Adrenal*, *Breast*, *Cervix*, *Colon*, *Lung*, *Skin*

Table 5a (continued)

Entrez ID	Gene Symbol	Gene Name	Tissues
11130	*ZWINT*	ZW10 interactor	Adrenal, Cervix, Liver, Lung, Pancreas, Prostate
22974	*TPX2*	TPX2, microtubule-associated, homolog (Xenopus laevis)	Adrenal, Breast, Cervix, Colon, Ovary, Skin
51203	*NUSAP1*	Nucleolar and spindle associated protein 1	Breast, Cervix, Kidney, Liver, Lung, Pancreas
54810	*GIPC2*	GIPC PDZ domain containing family, member 2	*Adrenal*, *Breast*, *Kidney*, *Liver*, *Ovary*, *Skin*
84981	*MGC14376*	Hypothetical protein MGC14376	*Adrenal*, *Breast*, *Colon*, *Kidney*, *Liver*, *Ovary*
38	*ACAT1*	Acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase)	*Adrenal*, *Colon*, *Kidney*, *Liver*, *Pancreas*
762	*CA4*	Carbonic anhydrase IV	*Breast*, *Colon*, *Kidney*, *Lung*, *Pancreas*
983	*CDC2*	Cell division cycle 2, G1 to S and G2 to M	Adrenal, Breast, Cervix, Liver, Ovary
2053	*EPHX2*	Epoxide hydrolase 2, cytoplasmic	*Adrenal*, *Colon*, *Kidney*, *Pancreas*, *Skin*
2348	*FOLR1*	Folate receptor 1 (adult)	*Adrenal*, *Breast*, *Kidney*, Ovary, *Stomach*
3075	*CFH*	Complement factor H	*Adrenal*, *Breast*, Cervix, *Ovary*, *Skin*
3248	*HPGD*	Hydroxyprostaglandin dehydrogenase 15-(NAD)	*Cervix*, *Colon*, *Kidney*, *Liver*, *Stomach*
4128	*MAOA*	Monoamine oxidase A	*Breast*, *Colon*, *Kidney*, *Ovary*, *Skin*
4171	*MCM2*	Minichromosome maintenance complex component 2	Cervix, Colon, Liver, Pancreas, Prostate
4246	*SCGB2A1*	Secretoglobin, family 2A, member 1	*Colon*, Ovary, *Prostate*, *Skin*, *Stomach*
4494	*MT1F*	Metallothionein 1F	*Colon*, *Kidney*, *Liver*, *Pancreas*, *Stomach*
4495	*MT1G*	Metallothionein 1G	*Colon*, *Kidney*, *Liver*, *Pancreas*, *Stomach*
5950	*RBP4*	Retinol binding protein 4, plasma	*Adrenal*, *Brain*, *Breast*, *Kidney*, *Skin*
6776	*STAT5A*	Signal transducer and activator of transcription 5A	*Adrenal*, *Breast*, *Ovary*, *Prostate*, Stomach
7102	*TSPAN7*	Tetraspanin 7	*Breast*, *Colon*, *Kidney*, *Liver*, *Lung*
9073	*CLDN8*	Claudin 8	*Breast*, *Cervix*, *Colon*, *Kidney*, *Skin*
9173	*IL1RL1*	Interleukin 1 receptor-like 1	*Adrenal*, *Kidney*, *Liver*, *Lung*, Skin
9768	*KIAA0101*	KIAA0101	Adrenal, Breast, Cervix, Liver, Lung
9936	*CD302*	CD302 molecule	*Adrenal*, *Breast*, *Liver*, *Ovary*, *Skin*
10051	*SMC4*	Structural maintenance of chromosomes 4	Adrenal, Brain, Cervix, Kidney, Pancreas
10894	*LYVE1*	Lymphatic vessel endothelial hyaluronan receptor 1	*Breast*, *Liver*, *Lung*, *Ovary*, *Skin*
23492	*CBX7*	Chromobox homolog 7	*Brain*, *Breast*, *Lung*, *Ovary*, *Skin*
35	*ACADS*	Acyl-Coenzyme A dehydrogenase, C-2 to C-3 short chain	*Colon*, *Liver*, *Pancreas*, *Prostate*
290	*ANPEP*	Alanyl (membrane) aminopeptidase (aminopeptidase N, aminopeptidase M, microsomal aminopeptidase, CD13, p150)	*Breast*, *Colon*, *Kidney*, *Pancreas*

Table 5a (continued)

Entrez ID	Gene Symbol	Gene Name	Tissues
994	*CDC25B*	Cell division cycle 25 homolog B (S. pombe)	Cervix, Colon, Pancreas, Skin
1012	*CDH13*	Cadherin 13, H-cadherin (heart)	Adrenal, Liver, *Lung*, Stomach
1113	*CHGA*	Chromogranin A (parathyroid secretory protein 1)	*Adrenal*, *Brain*, *Colon*, *Stomach*
1164	*CKS2*	CDC28 protein kinase regulatory subunit 2	Breast, Cervix, Ovary, Pancreas
1282	*COL4A1*	Collagen, type IV, alpha 1	Liver, Ovary, Pancreas, Stomach
1410	*CRYAB*	Crystallin, alpha B	*Breast*, *Cervix*, *Lung*, *Prostate*
1776	*DNASE1L3*	Deoxyribonuclease I-like 3	*Adrenal*, *Colon*, *Kidney*, *Liver*
1805	*DPT*	Dermatopontin	*Adrenal*, *Breast*, *Prostate*, *Skin*
1827	*RCAN1*	Regulator of calcineurin 1	*Breast*, *Colon*, *Kidney*, *Liver*
2023	*ENO1*	Enolase 1, (alpha)	Adrenal, Lung, Ovary, Stomach
2104	*ESRRG*	Estrogen-related receptor gamma	*Kidney*, *Pancreas*, *Skin*, *Stomach*
2146	*EZH2*	Enhancer of zeste homolog 2 (Drosophila)	Cervix, Kidney, Lung, Prostate
2273	*FHL1*	Four and a half LIM domains 1	*Breast*, *Colon*, *Lung*, *Skin*
2305	*FOXM1*	Forkhead box M1	Cervix, Colon, Ovary, Skin
2690	*GHR*	Growth hormone receptor	*Breast*, *Liver*, *Ovary*, *Skin*
2819	*GPD1*	Glycerol-3-phosphate dehydrogenase 1 (soluble)	*Breast*, *Kidney*, *Lung*, *Pancreas*
3131	*HLF*	Hepatic leukemia factor	*Brain*, *Breast*, *Ovary*, *Skin*
3223	*HOXC6*	Homeobox C6	Cervix, *Ovary*, Pancreas, Stomach
3479	*IGF1*	Insulin-like growth factor 1 (somatomedin C)	*Breast*, *Liver*, *Prostate*, *Skin*
3489	*IGFBP6*	Insulin-like growth factor binding protein 6	*Adrenal*, *Breast*, *Ovary*, *Skin*
3815	*KIT*	V-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	*Breast*, *Colon*, *Ovary*, *Prostate*
3957	*LGALS2*	Lectin, galactoside-binding, soluble, 2	*Colon*, *Ovary*, *Pancreas*, *Prostate*
4147	*MATN2*	Matrilin 2	*Adrenal*, *Breast*, *Prostate*, *Skin*
4501	*MT1X*	Metallothionein 1X	*Colon*, *Kidney*, *Liver*, *Skin*
4692	*NDN*	Necdin homolog (mouse)	*Breast*, *Cervix*, *Ovary*, *Prostate*
4830	*NME1*	Non-metastatic cells 1, protein (NM23A) expressed in	Colon, Kidney, Lung, Stomach
5050	*PAFAH1B3*	Platelet-activating factor acetylhydrolase, isoform Ib, gamma subunit 29kDa	Adrenal, Breast, Lung, Skin
5101	*PCDH9*	Protocadherin 9	*Adrenal*, *Breast*, *Ovary*, *Prostate*
5121	*PCP4*	Purkinje cell protein 4	*Brain*, *Kidney*, *Prostate*, *Skin*
5348	*FXYD1*	FXYD domain containing ion transport regulator 1 (phospholemman)	*Breast*, *Colon*, *Lung*, *Skin*
5577	*PRKAR2B*	Protein kinase, cAMP-dependent, regulatory, type II, beta	*Colon*, *Liver*, *Ovary*, *Skin*
5734	*PTGER4*	Prostaglandin E receptor 4 (subtype EP4)	*Adrenal*, *Breast*, *Colon*, *Pancreas*
5984	*RFC4*	Replication factor C (activator 1) 4, 37kDa	Adrenal, Cervix, Kidney, Lung
6338	*SCNN1B*	Sodium channel, nonvoltage-gated 1, beta (Liddle syndrome)	*Cervix*, *Colon*, *Kidney*, *Stomach*
6456	*SH3GL2*	SH3-domain GRB2-like 2	*Brain*, *Kidney*, *Ovary*, *Stomach*
6659	*SOX4*	SRY (sex determining region Y)-box 4	Brain, Cervix, Liver, Lung
7045	*TGFBI*	Transforming growth factor, beta-induced, 68kDa	Cervix, Colon, Liver, Pancreas
7049	*TGFBR3*	Transforming growth factor, beta receptor III	*Breast*, *Kidney*, *Lung*, *Skin*
7058	*THBS2*	Thrombospondin 2	Colon, Lung, Pancreas, Stomach
7070	*THY1*	Thy-1 cell surface antigen	Colon, Liver, Pancreas, Stomach
7122	*CLDN5*	Claudin 5 (transmembrane protein deleted in velocardiofacial syndrome)	*Breast*, *Lung*, *Prostate*, *Skin*
7433	*VIPR1*	Vasoactive intestinal peptide receptor 1	*Brain*, *Colon*, *Liver*, *Lung*
7704	*ZBTB16*	Zinc finger and BTB domain containing 16	*Breast*, *Lung*, *Ovary*, *Skin*
9104	*RGN*	Regucalcin (senescence marker protein-30)	*Breast*, *Kidney*, *Ovary*, *Pancreas*
9413	*C9orf61*	Chromosome 9 open reading frame 61	*Breast*, *Cervix*, *Lung*, *Stomach*

Table 5a (continued)

Entrez ID	Gene Symbol	Gene Name	Tissues
8418	*CMAH*	Cytidine monophosphate-N-acetylneuraminic acid hydroxylase (CMP-N-acetylneuraminate monooxygenase)	*Adrenal*, *Colon*, *Ovary*, *Skin*
8611	*PPAP2A*	Phosphatidic acid phosphatase type 2A	*Breast*, *Colon*, *Ovary*, *Skin*
9104	*RGN*	Regucalcin (senescence marker protein-30)	*Breast*, *Kidney*, *Ovary*, *Pancreas*
9413	*C9orf61*	Chromosome 9 open reading frame 61	*Breast*, *Cervix*, *Lung*, *Stomach*
9601	*PDIA4*	Protein disulfide isomerase family A, member 4	Brain, Cervix, Lung, Ovary
9636	*ISG15*	ISG15 ubiquitin-like modifier	Breast, Cervix, Liver, Pancreas
9837	*GINS1*	GINS complex subunit 1 (Psf1 homolog)	Adrenal, Cervix, Liver, Lung
10376	*TUBA1B*	Tubulin, alpha 1b	Cervix, Kidney, Pancreas, Stomach
10417	*SPON2*	Spondin 2, extracellular matrix protein	*Adrenal*, Colon, Liver, Pancreas
10797	*MTHFD2*	Methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase	Brain, Cervix, Lung, Ovary
11170	*FAM107A*	Family with sequence similarity 107, member A	*Breast*, *Lung*, *Pancreas*, *Prostate*
11335	*CBX3*	Chromobox homolog 3 (HP1 gamma homolog, Drosophila)	Cervix, Colon, Kidney, Lung
23213	*SULF1*	Sulfatase 1	Colon, Lung, Pancreas, Stomach
25802	*LMOD1*	Leiomodin 1 (smooth muscle)	*Adrenal*, *Breast*, *Lung*, *Skin*
25928	*SOSTDC1*	Sclerostin domain containing 1	*Breast*, *Cervix*, *Lung*, *Stomach*
26586	*CKAP2*	Cytoskeleton associated protein 2	Colon, Kidney, Liver, Pancreas
27284	*SULT1B1*	Sulfotransferase family, cytosolic, 1B, member 1	*Cervix*, *Colon*, Skin, *Stomach*
51053	*GMNN*	Geminin, DNA replication inhibitor	Adrenal, Cervix, Liver, Lung
51659	*GINS2*	GINS complex subunit 2 (Psf2 homolog)	Adrenal, Cervix, Kidney, Lung
53405	*CLIC5*	Chloride intracellular channel 5	*Breast*, *Colon*, *Kidney*, *Lung*
55165	*CEP55*	Centrosomal protein 55kDa	Cervix, Colon, Lung, Pancreas
57088	*PLSCR4*	Phospholipid scramblase 4	*Breast*, *Liver*, *Ovary*, *Skin*
79728	*PALB2*	Partner and localizer of BRCA2	Adrenal, Lung, Pancreas, Stomach
84560	*MT4*	Metallothionein 4	*Adrenal*, *Kidney*, *Liver*, *Pancreas*
283298	*OLFML1*	Olfactomedin-like 1	*Adrenal*, *Breast*, *Ovary*, *Skin*

b)

Entrez ID	Gene Symbol	Status	Drug Name	Indication
7153	*TOP2A*	Approved	Dexrazoxane	For reducing the incidence and severity of cardiomyopathy associated with doxorubicin administration in women with metastatic breast cancer
		Approved	Valrubicin	For the treatment of cancer of the bladder.
		Approved	Teniposide	Treatment of refractory acute lymphoblastic leukaemia
		Approved	Epirubicin	For use as a component of adjuvant therapy in patients with evidence of axillary node tumor involvement following resection of primary breast cancer

Table 5b (continued)

Entrez ID	Gene Symbol	Status	Drug Name	Indication
		Approved	Etoposide	For use in combination with other chemotherapeutic agents in the treatment of refractory testicular tumors and as first line treatment in patients with small cell lung cancer. Also used to treat other malignancies such as lymphoma, non-lymphocytic leukemia, and glioblastoma multiforme.
		Approved	Idarubicin	For the treatment of acute myeloid leukemia (AML) in adults. This includes French-American-British (FAB) classifications M1 through M7.
		Approved	Lucanthone	Intended for use as a radiation sensitizer in the treatment of brain cancer.
2053	*EPHX2*	Approved	Tamoxifen	For the treatment of breast cancer
2104	*ESRRG*	Approved	Diethylstilbestrol	Used in the treatment of prostate cancer
983	*CDC2*	Experimental	Flavopiridol	n/a
10376	*TUBA1B*	Experimental	Epothilone B	n/a
		Experimental	Epothilone D	n/a

Table 6 - Annotation of New Cancer Genes: a) List of genes that are differentially expressed in at least 4 tissues and have not been previously associated with cancer in non-microarray literature. The tissues in which the genes are altered are shown where regular font indicated upregulation and italics represents downregulation in cancer compared to normal tissue. Entrez IDs shown in bold represent genes that appeared to be significant in the general normal/cancer comparisons. b) List of approved cancer drugs targeting commonly altered genes that have not been previously associated with cancer

a)

Entrez ID	Gene Symbol	Gene Name	Tissues
79888	*LPCAT1*	Lysophosphatidylcholine acyltransferase 1	Cervix, Colon, Kidney, Pancreas, Stomach
33	*ACADL*	Acyl-Coenzyme A dehydrogenase, long chain	*Lung*, *Ovary*, *Pancreas*, *Skin*
2824	*GPM6B*	Glycoprotein M6B	*Adrenal*, *Breast*, *Lung*, *Prostate*
8424	*BBOX1*	Butyrobetaine (gamma), 2-oxoglutarate dioxygenase (gamma-butyrobetaine hydroxylase) 1	*Breast*, *Cervix*, *Kidney*, *Skin*
9452	*ITM2A*	Integral membrane protein 2A	*Breast*, *Colon*, *Ovary*, *Skin*
9631	*NUP155*	Nucleoporin 155kDa	Adrenal, Cervix, Lung, Stomach
10391	*CORO2B*	Coronin, actin binding protein, 2B	*Adrenal*, *Breast*, *Lung*, *Ovary*
27147	*DENND2A*	DENN/MADD domain containing 2A	*Breast*, *Colon*, *Lung*, *Prostate*
51660	*BRP44L*	Brain protein 44-like	*Kidney*, *Liver*, *Skin*, *Stomach*
51751	*HIGD1B*	HIG1 domain family, member 1B	Adrenal, Liver, *Lung*, *Prostate*
65983	*GRAMD3*	GRAM domain containing 3	*Adrenal*, *Breast*, *Colon*, *Skin*

b)

Entrez ID	Gene Symbol	Status	Drug Name	Indication
8424	*BBOX1*	Approved	Vitamin C	Used to treat vitamin C deficiency, scurvy, delayed wound and bone healing, urine acidification, and in general as an antioxidant. It has also been suggested to be an effective antiviral agent.
8424	*BBOX1*	Approved	Succinic acid	For nutritional supplementation, also for treating dietary shortage or imbalance

A second, alternative method was used to identify those genes that are common in the general pathway of cancer. A cancer microarray database and a control database were generated randomly selecting ten samples from each tissue type, resulting in a set of 130 cancer and 130 control samples. SAM analysis was then used to identify the top 400 significant genes, repeating this operation a hundred times. The union of these hundred gene lists each containing 400 genes produced 1,411 genes of which 44 are in the KEGG's pathways in cancer. The union of genes from the first 50 iterations produces a list of 1,235 genes indicating that additional iterations produce few new SAM genes. The p-value associated with the intersection with the pathways in cancer using the hypergeometric test using the platform genes as the background is 0.0196. Of the 1411, 271 genes are found in at least 70% of the iterations, of which 12 are found in the pathways of cancer with corresponding p-values of 0.0208. Moreover, 74 genes out of the 271 appeared among the 132 genes listed in Table 5 and Table 6. The p-value for this overlap is 9.0763E-082. The list of 271 genes is provided as Additional File 2. Taken together with genes in Table 5 and Table 6, they can be used to extend and further annotate the general pathways of cancer.

4.5 Discussion

In this study, nearly six thousand microarray samples were obtained from comparable Affymetrix platforms to investigate the commonalities as well as the tissue specific components of normal to cancer transformation in thirteen distinct tissue types. It was possible to obtain such a large sample size through the addition of highly asymmetric datasets into the microarray sample pool. Mainly, those datasets with large numbers of cancer samples and small numbers (including zero) of control samples and vice versa, were considered. Otherwise, out of the thirteen tissue types under study, only the breast, colon, kidney, and pancreas tissues had three or more different datasets that included at least ten cancer and ten control samples.

This approach is unusual in the sense that it does not fit typical meta-scale analyses [46-48, 86, 105-106] where each dataset needs to have both disease and control samples in sufficient numbers and datasets are normalized and analyzed separately for significant genes. Using the meta-analysis approach, Ramasamy et al. [46] analyzed 21 distinct microarray datasets from 14 different cancer types comprising 419 control and 973 samples. The minimum sample size for cancer and control in their study was seven and some of the tissue types such as renal tissue appeared only in one dataset in their collection. The advantage of this method is the flexibility concerning the multiple platforms that can be incorporated and thereby increasing sample size through acceptance of several platforms. Because this research focuses on a set of comparable platforms, the results are not directly comparable. Nevertheless, Ramasamy et al. [46] published five upregulated and five downregulated genes as most significantly associated with cancer. Among this list of ten, four genes (*TMEM136*, *RBM15*, *FGD4* and *KIAA1881*) are not part of the minimal platform considered in this study, suggesting that as the data in public-access microarray repositories grow, datasets used in the proposed approach will be restricted to the latest version of platforms containing many more probes. Of the remaining six genes, the top 400 lists from this research confirmed the downregulation of *PRKAR2B* and *GPM6B* in four different tissues. Genes *MYOM2* and *RBCK1* in their ten gene list were SAM genes in multiple lists in this study but were in the top 400 only in the liver gene list. Similarly, *ALG3* did not appear in any of the top 400 gene lists but was significantly upregulated in six of the thirteen tissues in the complete SAM lists. The last gene in their list, *IRAK1* was a top ten ranking gene in the pancreas SAM gene list, however, this gene was downregulated in pancreas as well as five more tissues in this study, as opposed to the upregulated notation presented to the gene by Ramasamy et al. [46]. Note that this research contained 106 cancer and 71 normal pancreatic tissue microarray samples as opposed to the 12 tumor and 7 normal microarray samples in [46]. It is not feasible to summarize the comparison with a p-value because the gene list presented in [46] contains only ten genes whereas the various gene lists produce by this research contain hundreds of genes. Nevertheless,

it is clear that the two approaches could potentially produce gene lists whose intersection is unlikely to be a random event.

The proposed approach takes advantage of the rapid increase of asymmetric data in public-access microarray repositories. Moreover, gene lists predicted using this large asymmetric data reproduces much of the research literature on cancer-associated genes obtained by experimental methods other than microarray. This analysis predicts 132 genes as significantly altered in normal to cancer transformation in at least four tissue types and out of this list, 121 were previously annotated in the literature as cancer-related. The remaining eleven genes comprise potential targets for further studies in cancer research. Note also that 74 out of the 132 genes in the list also appear in 70% of the SAM gene lists generated by comparing normal and cancer datasets comprising of randomly chosen ten samples from each tissue type. The two gene lists presented in this study for cancer-associated genes with multiple tissue specificity will further contribute to the annotation of pathways of cancer. Recently emerging annotation-based microarray data tools such as A-MADMAN [142] will help in the compilation process of large-scale microarray data for studying complex diseases, and for biomarker and drug development.

4.6 Conclusion

In this study, almost 6,000 microarray samples were obtained and a total of 329 genes were identified that appeared as highly significant in normal to cancer transformation with regards to multiple cancer types. The gene list consists largely of genes that have already been associated with cancer in research literature excluding microarray studies. The list can be used in the detailed annotation of cancer pathways. In addition due to the inclusion of numerous subtypes and cancer grades, the genes in this list can serve as potential targets for new drug development.

Chapter 5: Virus and host iron binding protein interactions

5.1 Summary

The intricate relationships that evolve between viruses, host iron binding proteins, cellular iron supplies and the immune response are the core of this chapter. The regulation and utilization of iron in humans has evolved to a high degree of complexity. Many of the basal pathways and cellular functions depend on iron as a component of iron binding proteins. These proteins are therefore involved in a wide range of functions varying from energy metabolism to DNA synthesis to oxygen transport. More importantly, iron binding proteins are also part of the human immune system. Iron redox properties render it crucial yet potentially toxic. As a result, iron homeostasis is necessary since iron is needed for maintaining a healthy system as well as a diseased one. In general, viruses rely on host cellular machinery for their own replication. Consequently, sufficient iron quantities are essential to allow for efficient viral propagation. Iron overloads have been observed in viral infections including HIV and hepatitis C and are generally associated with poor prognosis. By using publicly accessible databases, human iron binding proteins were identified. Microarray data on three viral infections: HIV, hepatitis C and influenza A were collected and analyzed to identify direct and indirect targets of these viral infections that are dependent on iron and therefore important for iron homeostasis. Results revealed significant changes in the transcript levels of 101, 122, and 107 iron binding proteins in HIV, hepatitis C and influenza A, respectively. These proteins appeared to be involved in biological processes related to cellular metabolism, oxidative stress response and immune system processes. Moreover, the microarray results captured some of the known imperative changes induced by HIV-1 viruses that have been documented in the literature. These outcomes emphasize the vitality of iron for sustaining viral demands as well as the critical role that iron recruited by the virus could potentially play in helping the virus escape the host's immunes system.

5.2 Background

Several fundamental cellular operations in living systems require the presence of iron ion binding proteins [143] and rely on the redox abilities of ferrous (Fe^{2+}) and ferric (Fe^{3+}) iron [144]. Iron ions act as a cofactor for enzymes involved in energy metabolism, DNA synthesis, replication and repair, transcription, and mRNA translation, rendering it essential for cells. In addition, iron in hemoglobin and myoglobin binds oxygen allowing for its transport [143, 145]. The primary function of iron in living systems is therefore greatly dependent on its role in shuttling electrons between proteins and its flexibility for binding ligands in diverse orientations [143].

Iron also has a crucial role in immunity and immunosurveillance. This is achieved through iron's involvement in cell-mediated immune effector pathways and cytokine activities as well as its role in promoting immune cells' growth [145-147], which can then affect the cells' response to an invading pathogen. In return, cytokines and radicals produced and released by the immune cells can control and regulate iron homeostasis, through transcriptional and post-transcriptional methods [145]. Hence, iron metabolism and the immune system possess a delicate relationship through which they can regulate one another.

The link between the immune defense and iron metabolism is often targeted by infectious agents including the human immunodeficiency virus and hepatitis C [145, 148-149]. Viruses depend on host cells for their survival, and viral replication requires enhanced cellular metabolism for transcribing and translating viral genomes and proteins. Since these processes depend on and require iron, the host cells have to contain a sufficient supply of iron to meet the demands [143]. Iron accumulation can accompany the more advanced stages of HIV infection [148, 150], while increased iron storage in bone marrow macrophages could be associated with shorter survival times [151-152]. After an HIV infection, the virus reverse-transcribes its RNA into double-stranded DNA which is then integrated into the host's genome. HIV kills target cells and alters

gene expression through the involvement of viral regulatory proteins (*Tat*, *Rev*, *Nef*), and many of the activities targeted by HIV are iron-dependent [148].

A number of recent studies focused on the role of iron in clinical progression of HIV infection. Elevated iron stores have been detected in HIV patients including those in brain, liver and muscles [148]. The effects of iron supplementation on HIV infection have also been studied [153], with research on pregnant women from Zimbabwe revealed that receiving iron supplementation was an independent predictor of higher viral load [154]. In a study on thalassemic HIV patients, the rate of progression of the disease was associated with desferroxamine and higher serum ferritin concentrations [155]. Meanwhile research on Belgian HIV cohort reported that haptoglobin 2-2 was related to increased iron storage, higher rates of viral replication and shortened survival [156]. Another study in Kenya found that iron supplementation reduced the rates of post-treatment reinfection, and that viral load was higher in patients receiving iron compared to the placebo group [157], while several other studies have agreed that iron overload resulted in decreased survival of HIV patients [158-160].

Similar iron overload has been observed in hepatitis C patients. Hepatitis C is known to cause liver injury and cancer, and while the process is not fully understood, the pathology seems to be driven by chronic inflammation. Increased morbidity and mortality in hepatitis C patients has been associated with elevated levels of cellular iron, which behaves as a pro-inflammatory agent. Haemochromatosis can also result in chronic iron deposition in the liver, and has been associated with cirrhosis and injury that could lead to hepatocellular carcinoma. While iron deposition can be the result of inherited defects in host genes involved in iron metabolism, the virus itself can also induce similar iron overloads. Elevated levels of hydroxyl radicals are generated in the presence of excess iron. Such radicals are highly reactive and can therefore cause damage to proteins, DNA and lipids within the cell [143, 149], thereby inflicting damage to the cell

membranes and the genome [149]. The oxidative stress can also result in mitochondrial dysfunction and cause liver satellite cells to produce collagen, contributing to the development of fibrosis. Therefore, excess iron in hepatitis C patients can trigger an inflammatory environment that disturbs the liver's normal function [143].

This research uses bioinformatics methods and publicly accessible molecular and functional genomics databases to identify the components of virus-host crosstalk involving host iron binding proteins. The National Institute of Allergy & Infectious Diseases' (NIAID) HIV-1, Human Protein Interaction Database [161-163] and molecular function annotation have provided knowledge for the identification of nodes in pathways leading to HIV viral replication associated with iron binding. Results from HIV microarray analysis have confirmed viral effects on several key proteins including NADPH oxidase complex, *ABCE1*, *IDO1* and *ALOX5*. Differential gene expression analysis conducted on hepatitis C microarray data has also revealed significant overlap with HIV induced alterations to iron binding proteins. Data obtained on influenza A virus was also tested, providing a control non-persistent infection to facilitate the understanding of how persistent viral infections influence iron homeostasis and evade the host's immune system.

5.3 Methods

5.3.1 Identification of iron-associated proteins

To determine the human proteins that are associated with iron binding a list of proteins annotated with the GO [102-103] molecular functions "iron ion binding" and "iron-sulfur cluster binding" were retrieved from the GO Consortium and DAVID Bioinformatics [120-121] databases. The summation list was checked against the literature and UniProtKB [164-165] database to confirm the functional association of these proteins with iron. Only genes with RefSeq status of "REVIEWED" or "VALIDATED" were retained. Moreover, 6 proteins were added as iron

binding based on a recent review on the role of iron-ion binding proteins in viral infections [143]. The final list contained 299 iron binding proteins.

5.3.2 Identifying direct HIV-1 iron binding protein targets

To understand the interplay between HIV proteins and host proteins mediated by the function of iron within the cell, a list of proteins that are known to interact with HIV-1 proteins, either directly or indirectly, was obtained from the NIAID HIV-1, Human Protein Interaction Database [161-163] (Version: December 2009). The list contained a total of 1433 human proteins corresponding to 68 types of different interactions with the viral proteome. Proteins that were annotated as iron binding were identified within this list, to determine the iron-dependent proteins that are targeted by HIV-1.

5.3.3 Microarray dataset selection on viral infections

Changes in gene expression levels induced by HIV-1 infection were then investigated. Microarray datasets for healthy and HIV-1 infected CD4+ T-cells were collected from the Gene Expression Omnibus [96-97]. The details of the datasets used are shown in Table 7. In summary, 130 healthy and 21 HIV-1 infected CD4+ T-cell samples were collected. In order to identify commonalities in interactions with iron ion binding host proteins, similar data were obtained for hepatitis C and influenza A infected PBMC cells as well as healthy controls. The summary list of the datasets and the sample distribution is also presented in Table 7. The microarray data was confined to datasets hybridized on the Affymetrix human microarray platforms HG-U133A, HG-U133A2 and HG-U133 Plus 2.0, to allow for data merger due to the large overlap between these platforms.

Table 7 - Microarray Datasets: Microarray samples utilized in analysis of changes induced by viral infections.

GEO Accession #	Platform	Healthy Samples	Infected Samples
HIV Infection:			
GSE6740	HG-U133A	5	10
GSE9927	HG-U133 Plus 2.0	9	11
GSE6338	HG-U133 Plus 2.0	5	0
GSE7497	HG-U133A	16	0
GSE8835	HG-U133A	12	0
GSE10586	HG-U133 Plus 2.0	15	0
GSE12079	HG-U133 Plus 2.0	4	0
GSE13732	HG-U133 Plus 2.0	40	0
GSE14879	HG-U133 Plus 2.0	10	0
GSE14924	HG-U133 Plus 2.0	10	0
GSE17354	HG-U133A	4	0
Total		**130**	**21**
Hepatitis C Infection:			
GSE7123	HG-U133A	0	59
GSE11190	HG-U133 Plus 2.0	0	19
GSE11342	HG-U133A	0	20
Total		**0**	**98**
Influenza A Infection:			
GSE6269	HG-U133A & Plus 2.0	6	25
GSE17156	HG-U133A2	17	17
Total		**23**	**42**
Healthy PBMC:			
GSE8507	HG-U133 Plus 2.0	34	0
GSE8650	HG-U133A	21	0
GSE12839	HG-U133A	7	0
GSE14895	HG-U133A2	11	0
GSE15072	HG-U133A	8	0
GSE16728	HG-U133 Plus 2.0	5	0
Total		**86**	**0**

5.3.4 Microarray data normalization and differential gene expression

Raw .CEL files for all the samples were obtained and normalized using the refRMA [63] conducted in MATLAB . A total of 909 diverse microarray samples from the HG-U133 Plus 2.0 chip were used to train the data. The normalization process included background adjustment, quantile normalization, and median polishing. In addition, the custom ENTREZ CDF files (version 12) [117] were used in the normalization process in order to obtain Entrez gene intensities. The outputs from the training set were then used to adjust the normalized gene intensities of the data utilized in this analysis. The data from the common genes between the

three platforms were filtered and used for differential gene analysis. Finally, where multiple samples were obtained from the same individual, average intensities were calculated from these samples prior to differential gene expression analysis.

To identify the genes that exhibited changes in expression due to viral infection, the data was imported into R [118] and the Significance Analysis of Microarrays (SAM; [75]) test was applied using the *samr* package [119]. The test parameters were set to a hundred permutations of the analysis and the false discovery rate (FDR) was not allowed to exceed 6%. Each viral infection data was compared to the corresponding healthy control data separately. Among the significant genes satisfying these conditions, iron binding host proteins that appeared to be affected, directly or indirectly, by each of the viral infections at the transcriptional level were determined.

5.3.5 Distribution of gene expression levels of iron binding proteins

The inherent range of expression values for iron binding proteins at the transcript level was determined to understand their behavior within the normal state. Intersecting the three microarray platforms results in fewer iron binding genes that can be studied. Therefore, to obtain a full representation of the levels of gene expression in healthy tissues, only data hybridized on the HG-U133 Plus 2 chip for uninfected CD4+ T-cells were utilized, as it is the largest of the three platforms used in this study. Average intensity values were computed for each of the 17,726 genes on the chip. The data from the entire platform were clustered using the K-means clustering algorithm [166] into four groups depicting the genes' level of expression: low, medium-low, medium-high and high. The iron binding proteins were then mapped to their location within these clusters (Figure 14).

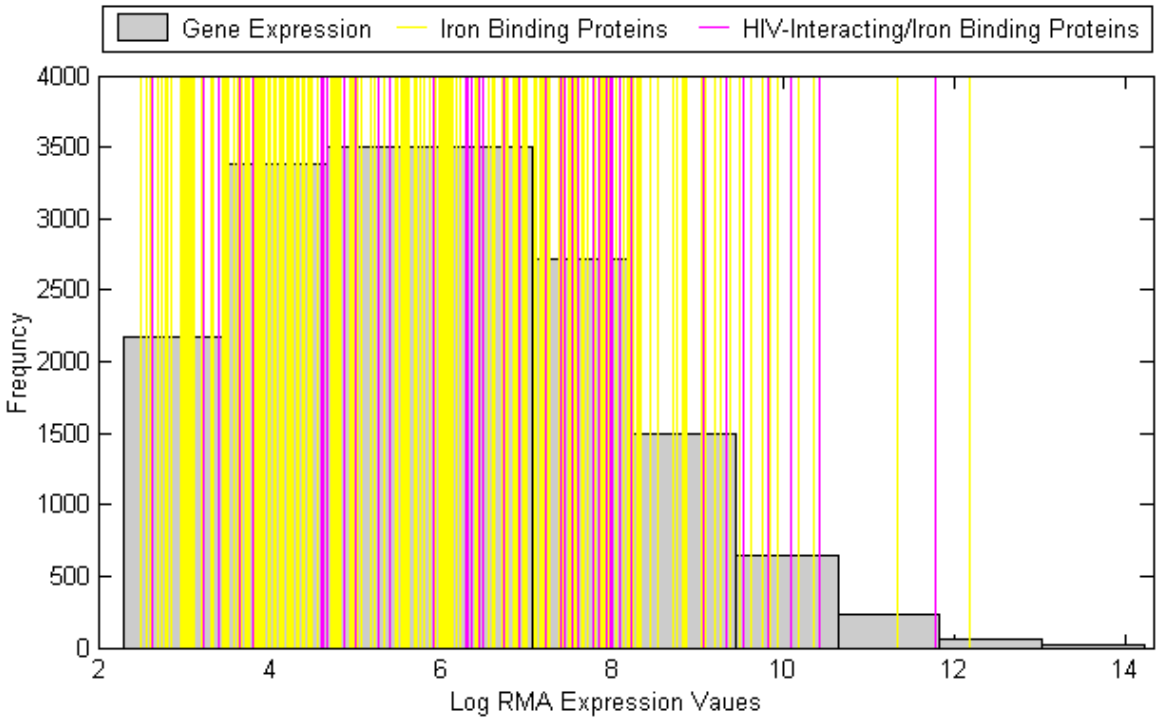

Figure 14 - Gene Expression Histogram: Distribution of Iron binding proteins (yellow) and HIV-interacting iron binding proteins (pink) with respect to all genes represented on the HG-U133 Plus 2.0 microarray platform based on gene intenisties in normal CD4+ T-cells.

5.4 Results

5.4.1 Iron binding proteins are statistically enriched among HIV targeted host proteins.

A total of 299 host proteins were identified as iron binding and among them 40 were previously annotated as HIV-1 interacting proteins (Figure 15a). A hypergeometric test based on all proteins from NCBI as the background resulted in a p-value of 1.80E-12 for this overlap. Iron-ion binding proteins are therefore statistically enriched among known HIV-1 interacting proteins, indicating the important role iron ions play in viral replication [143]. A list of those 40 proteins is provided in Table 9 depicting the HIV-1 proteins that target them and the types of interactions that occur between them.

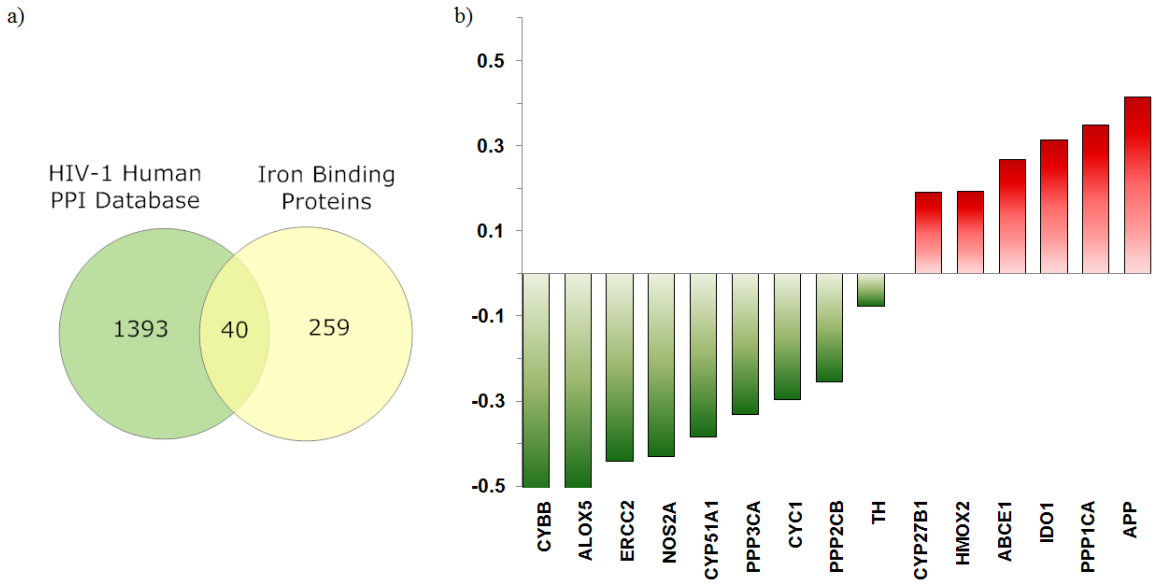

Figure 15 - HIV-1/Iron Binding Proteins and Differential Gene Expression: a) Venn diagram representing the overlap between the iron-binding proteins used in this study and proteins in the HIV-1 Human PPI Database, and b) Fold change values for the 15 genes whose expression is altered in HIV-1 infection

5.4.2 Gene expression analysis confirms the effect of HIV-1 infection on CD4+ T cells

Among the 40 iron-associated HIV-interacting proteins identified, 15 appeared to be significantly enriched according to SAM analysis conducted on CD4+ T cells from HIV patients compared with healthy T cells. Among those differentially expressed were 9 downregulated and 6 upregulated genes as shown in Figure 15b. The absolute fold change is shown after deducting 1.0 from all values. The behavior of these genes in the other viral conditions was also considered (Table 9). Eight of the 15 differentially expressed genes in HIV-1 infections exhibited similar significant alterations in hepatitis C, compared to only three concurrent alterations in gene expression inflicted by influenza A infection.

5.4.3 Significant commonalities in alteration induced by persistent viral infections on iron binding proteins

The microarray platform used for analysis of differential gene expression contained information on transcript levels of 191 out of the 299 iron binding proteins. Of these, 172 (90%) genes were

significantly altered in at least one type of viral infection, and a total of 43 genes appeared in all three SAM lists (Figure 16). Moreover, the overlap between the different infections was as follows: 73 genes in HIV ∩ Hepatitis C, 60 genes in HIV ∩ Influenza A, and 68 genes in Hepatitis C ∩ Influenza A. Using a hypergeometric test with a 0.05 threshold, only the intersection between HIV and hepatitis C was significant with a p-value of 4.6E-03. This highlights the similarities in HIV and hepatitis C as persistent infections, which is suggestive of their subsequent influences on the cellular machinery and immune system.

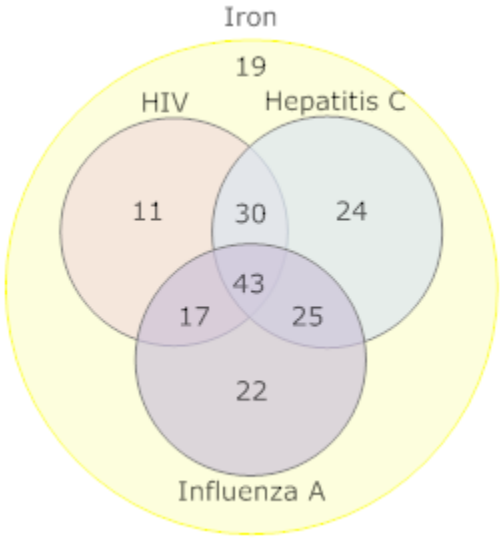

Figure 16 – Differential Expression of Iron Binding Proteins Induced by Viral Infections: Venn diagram depicting the distribution of the 191 iron binding proteins present on the microarray platform according to differential expression in HIV, Hepatitis C, and Influenza A.

To further investigate the roles of these proteins, all iron binding genes that were differentially expressed in at least one of the infection types considered were merged together. Enriched KEGG pathways were identified using a p-value cutoff of 0.05. Many of the common iron binding genes belong to the cytochrome P450 family. Therefore, as the results in Table 8 reflect, several metabolic pathways are affected by these infections. Retinol metabolism (Figure 17) contains the largest number of significantly altered genes, most of which are P450 enzymes. Iron

is known to interact with several dietary components including retinol (vitamin A). Retinol is essential for normal haematopoiesis and deficiencies have been associated with mild anaemia, poor immune response and delayed repair of damaged epithelial cells. Such deficiencies can also affect the severity of some infectious diseases. In addition, iron is necessary for retinol mobilization, and thus retinol and iron metabolism are closely interrelated [167].

Table 8 – Pathways Affected by Iron Binding Proteins: Enriched KEGG pathways associated with iron binding genes that were differentially expressed due to viral infection by HIV, Hepatitis C and/or Influenza A

KEGG Metabolic Pathway	Genes	P-Value
Retinol metabolism	17	3.80E-16
Drug metabolism – Cytochrome P450	15	7.67E-12
Linoleic acid metabolism	12	1.45E-11
Arachidonic acid metabolism	14	4.81E-11
Metabolism of xenobiotics by cytochrome P450	13	8.12E-10
Steroid hormone biosynthesis	12	1.85E-09
Drug metabolism	8	1.09E-05
Tryptophan metabolism	7	1.52E-04
Caffeine metabolism	4	4.62E-04
Primary bile acid biosynthesis	5	5.60E-04
Alzheimer's disease	12	7.47E-04
Parkinson's disease	9	2.28E-03
Porphyrin and chlorophyll metabolism	5	3.41E-03
Steroid biosynthesis	4	8.66E-03
Tyrosine metabolism	5	2.01E-02
Citrate cycle (TCA cycle)	4	2.85E-02
Arginine and proline metabolism	5	2.85E-02
Oxidative phosphorylation	7	3.37E-02
Biosynthesis of unsaturated fatty acids	3	4.66E-02

Cytochrome P450 proteins are also the major enzymes involved in drug metabolism contributing to the metabolism of approximately 75% of drugs [168], including tamoxifen, cyclophosphamide, ifosfamide and methadone (Figure 18). Drug doses are adjusted such that they can be cleared by the body at a reasonable rate. Hence, alterations in the availability of P450 enzymes, directly affect the body's ability to metabolize and clear drugs. While inhibition of P450 proteins can result in drug accumulation, drug-drug interactions and drug toxicity, induction of cytochromes can result in faster drug clearance, interfering with the drug's role and efficiency [168].

Table 9 - Regulation of HIV-Interacting/Iron Binding Proteins: A list of 40 iron binding proteins that are known to directly interact with HIV proteins, the types of interactions and proteins they associate with. The differential expression at the transcript level in HIV, Hepatitis C and Influenza A is shown: downregulated (↓), upregulated (↑) and non-differentially expressed (●). Expression level clusters, according to expression in healthy CD4+ T cells are also indicated: low (L), middle-low (ML), high-low (HL) and high (H).

Gene Symbol	Entrez ID	Cluster	HIV	Hepatitis C	Influenza A	Interaction	HIV Protein
NOX4	50507	L	●	●	↓	activated by	*Gp120*
NOX3	50508	L	●	●	↓	activated by	*Gp120*
PTGS2	5743	L	●	●	●	upregulated by upregulated by	*Gp120* *Tat*
TH	7054	L	↓	↓	↓	downregulated by	*Tat*
NOX5	79400	L	●	●	●	activated by	*Gp120*
NOS3	4846	ML	●	↓	●	inhibited by upregulated by	*Tat* *Gp41*
LTF	4057	ML	●	●	●	inhibits	*Gp120*
HFE	3077	ML	●	●	●	downregulated by	*Nef*
NOX1	27035	ML	●	↓	↓	activated by	*Gp120*
IDO1	3620	ML	↑	↓	●	release induced by	*Gp120*
CYP27B1	1594	ML	↑	↑	↓	activated by	*Matrix*
NOS1	4842	ML	●	●	↓	inhibited by upregulated by	*Tat* *Gp41*
CYBB	1536	ML	↓	↓	●	inhibited by	*Capsid*
PTGS1	5742	ML	●	●	●	upregulated by upregulated by	*Gp120* *Tat*
NOS2	4843	ML	↓	●	●	inhibited by upregulated by upregulated by	*Tat* *Gp120* *Gp41*
ALOX5	240	ML	↓	↓	●	upregulated by	*Gp120*
PPP2CB	5516	ML	↓	↓	↓	inhibits	*Tat*
ERCC2	2068	ML	↓	↓	●	binds	*Tat*
APP	351	MH	↑	↓	●	activated by inhibited by inhibits inhibits upregulated by	*Retropepsin* *Gp41* *Gp120* *Tat* *Tat*
CYCS	54205	MH	●	●	●	released by	*Vpr*

Table 9 (continued)

Gene Symbol	Entrez ID	Cluster	HIV	Influenza A	Hepatitis C	Interaction	HIV Protein
CYP51A1	1595	MH	↓	↓	↓	upregulated by	*Nef*
IKKE	9641	MH	●	●	●	binds phosphorylated by	*Gp120* *Nef*
IKK1	1147	MH	●	●	●	binds phosphorylated by	*Gp120* *Nef*
SDHB	6390	MH	●	●	↓	binds	*Tat*
ABCE1	6059	MH	↑	●	↑	associates with associates with	*Pr55* *Vif*
TFRC	7037	MH	●	●	●	downregulated by downregulated by	*Gp120* *Nef*
HMOX2	3163	MH	↑	●	↓	upregulated by	*Gp120*
IKK2	3551	MH	●	●	●	binds phosphorylated by	*Gp120* *Nef*
GLRX2	51022	MH	●	●	↓	activates	*Retropepsin*
CAT	847	MH	●	●	●	inhibits	*Gp160*
PPP3CA	5530	MH	↓	●	↑	activated by	*Tat*
PPP1CB	5500	MH	●	●	↓	stimulates upregulated by	*Tat* *Gp120*
PPP3CC	5533	MH	●	●	●	activated by	*Tat*
PPP1CA	5499	H	↑	●	↓	downregulated by stimulates	*Gp120* *Tat*
CYC1	1537	H	↓	●	↓	release induced by	*Vpr*
DOCK2	1794	H	●	↑	↑	associates with	*Nef*
PPP2CA	5515	H	●	●	↑	inhibits	*Tat*
PPP3CB	5532	H	●	●	↑	activated by	*Tat*
GLRX5	51218	H	●	↑	●	activates	*Retropepsin*
PPP1CC	5501	H	●	●	●	stimulates	*Tat*

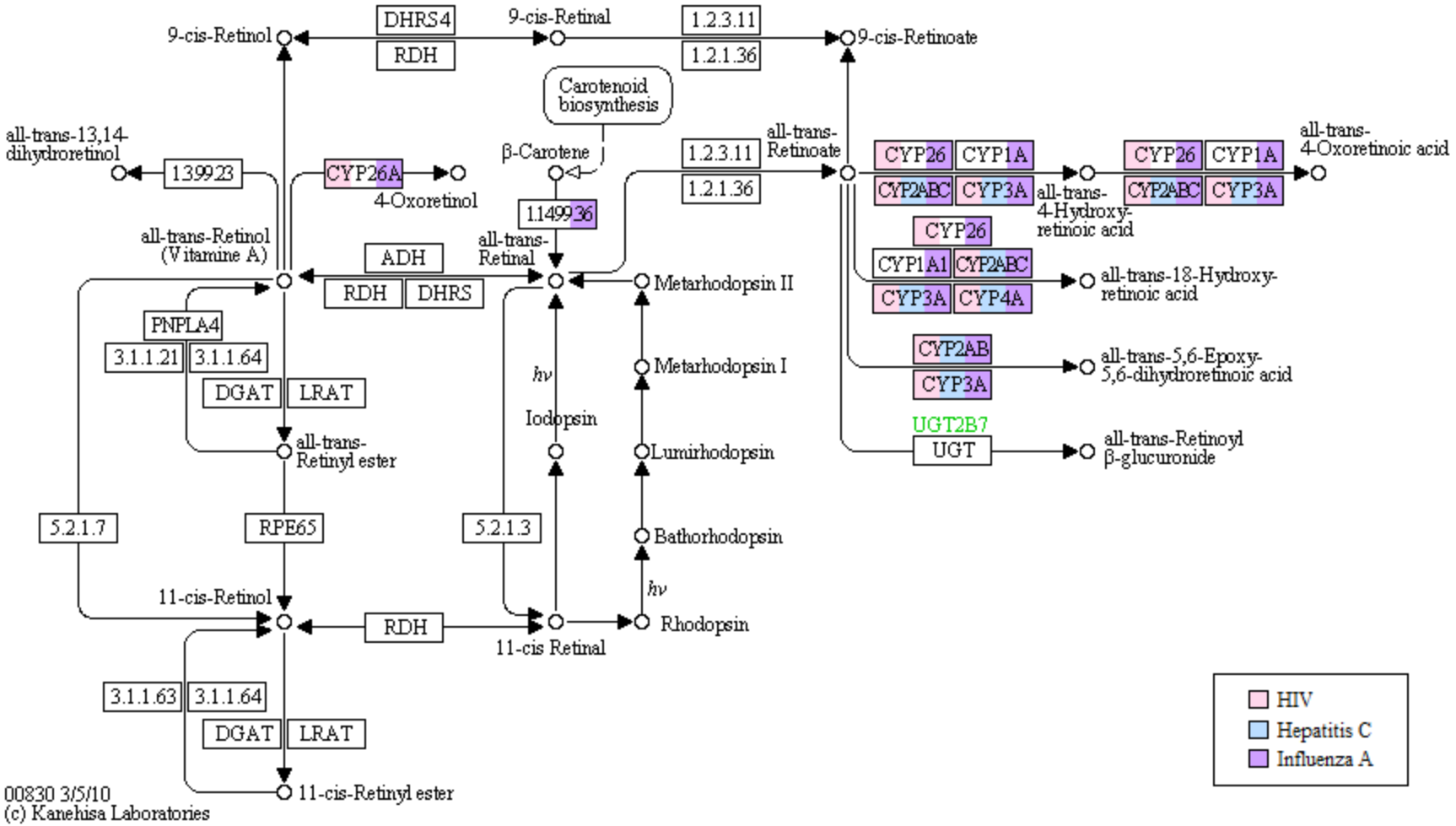

Figure 17 - Retinol Metabolism in Animals: Iron binding genes in KEGG's retinol metabolism pathway that are significantly altered by HIV (pink), Influenza A (blue) and Hepatitis C (purple) infections

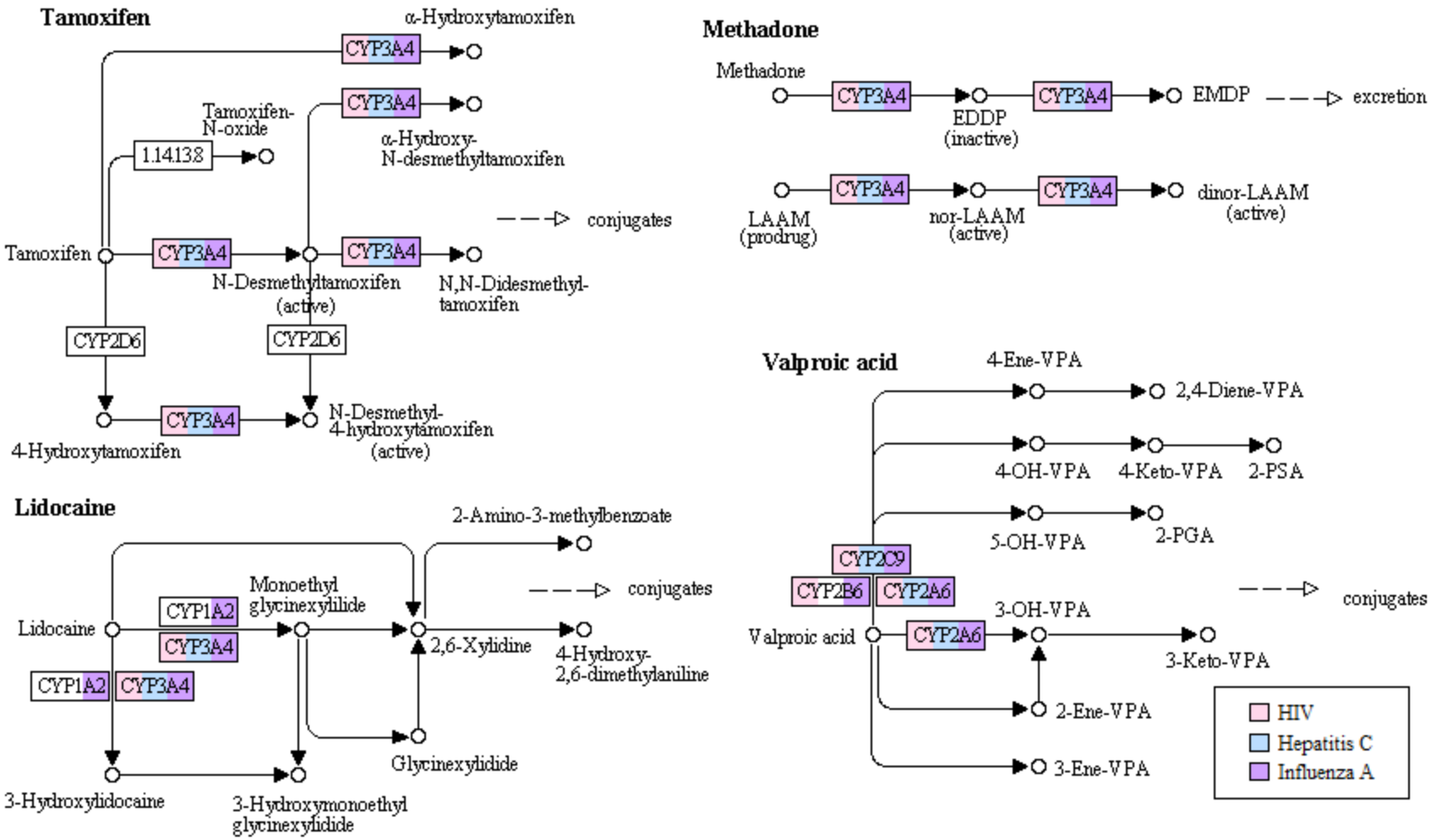

Figure 18 - Drug Metabolism Cytochrome P450: Iron binding genes in KEGG's drug metabolism (Cytochrome P450) that are significantly altered by HIV (pink), Influenza (blue) and Hepatitis C (purple) infection

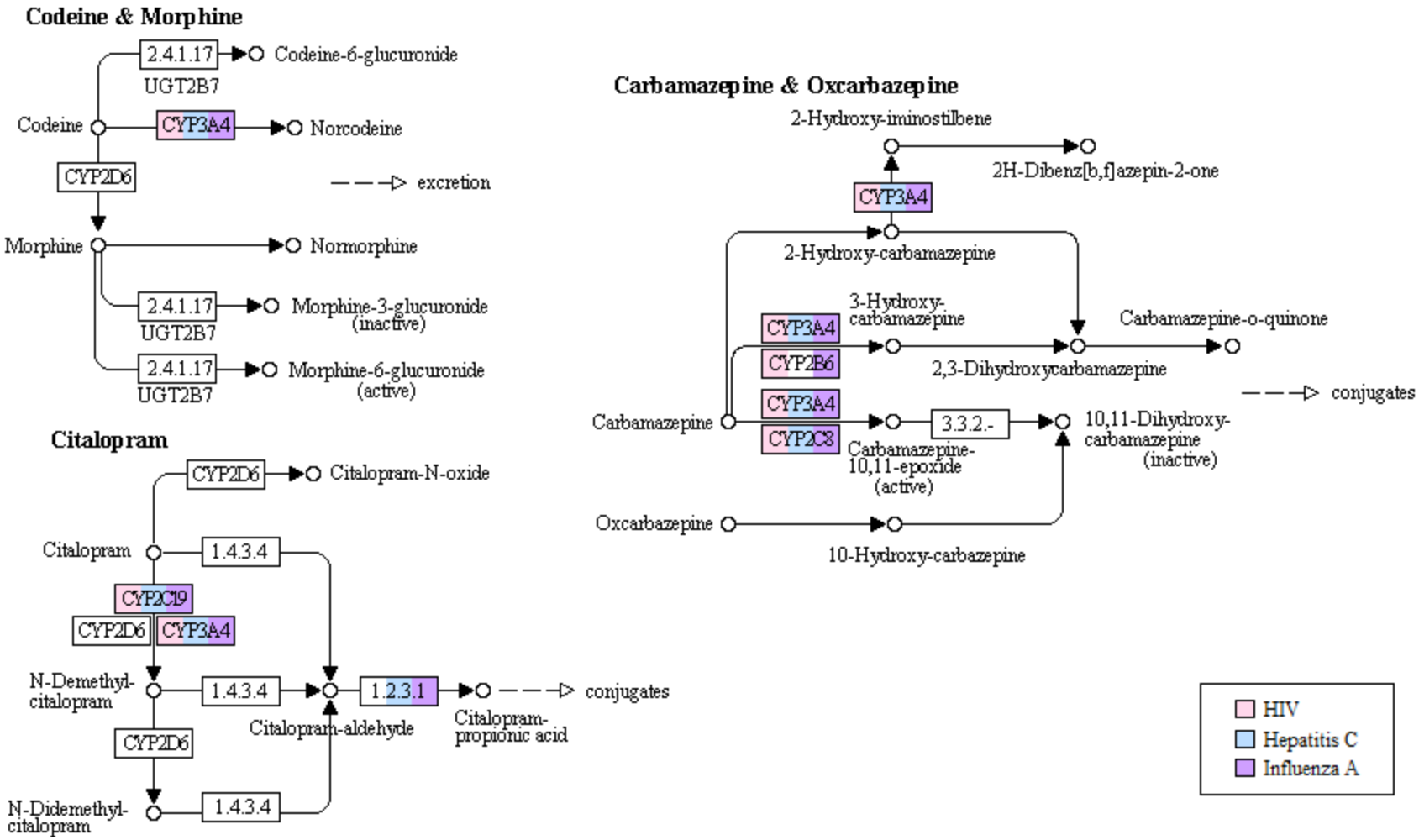

Figure 18 (continued)

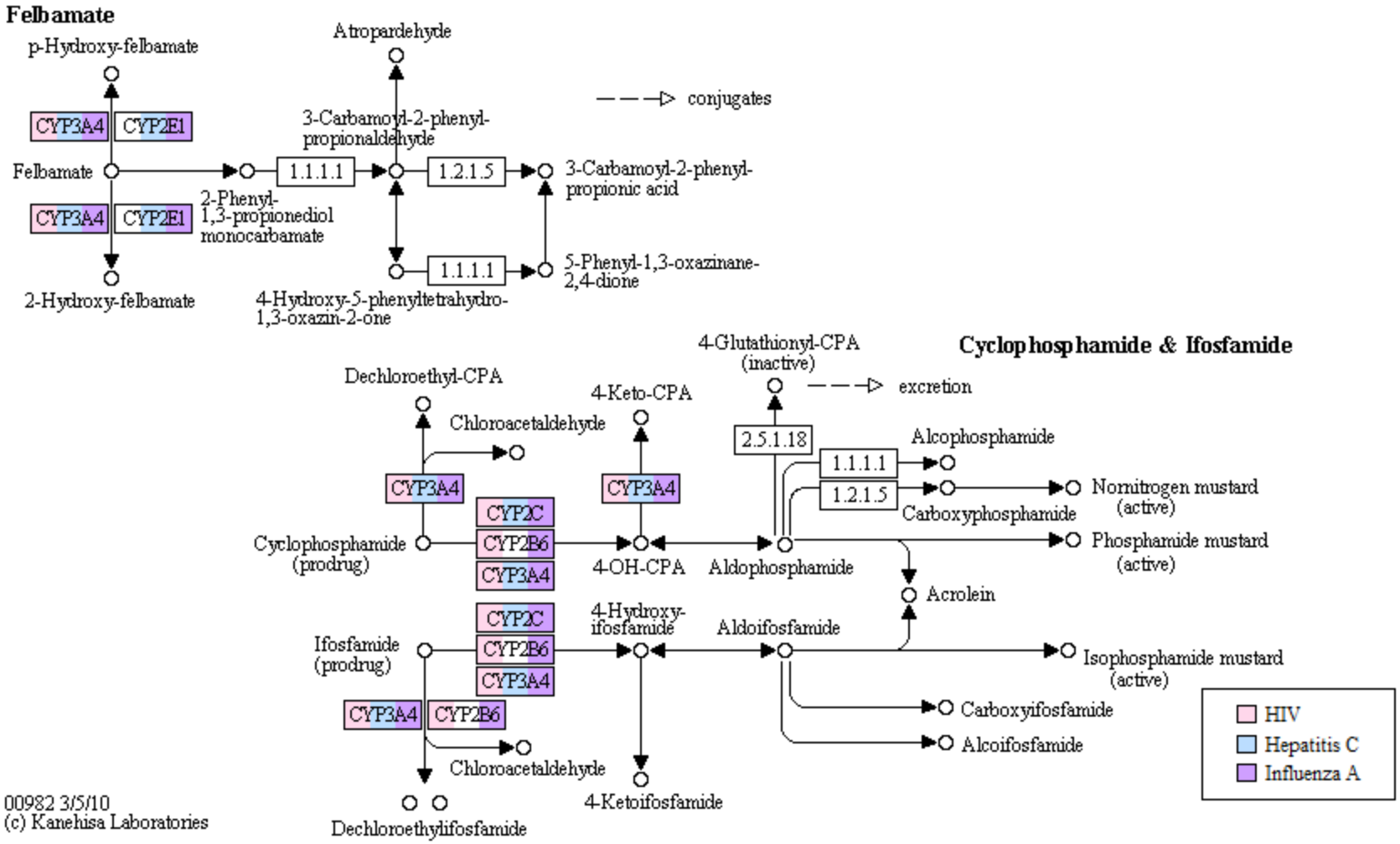

Figure 18 (continued)

More information on the biological processes revealed the enrichment of oxidative stress response. In addition, 12, 11 and 9 of the proteins in the HIV, hepatitis C, and influenza A lists, respectively, were involved in immune system response. This emphasizes the dual role viruses provoke on host cells in terms of metabolic demands and immune response. Nonetheless, the behavior of these genes within the different infections sheds some meaningful insight on the differences between HIV/hepatitis C and influenza A. For example, cytochrome B (*CYBB*) is downregulated in both HIV and hepatitis C, but no changes are observed in influenza. *CYBB* is one of the down-stream proteins in the interferon-gamma (*IFN-γ*) mediated immune response, and its downregulation impairs the oxidative burst response and phagocytosis that are needed to fight infections [169]. *ALOX5* and *ABCE1* are also known to be direct HIV-1 target proteins, and similar changes were perceived in hepatitis C microarray results but not influenza A.

5.5 Discussion

Iron is a vital nutrient for most organisms. It is universally present in the active site of iron binding proteins involved in oxygen transport, energy metabolism and respiratory pathways, DNA synthesis, and metabolite synthesis [143, 145, 170]. However, the reactive properties that typically make iron useful to these proteins also result in free iron being toxic [171]. Due to the high involvement of iron throughout the body's cellular processes, cells commit complex systems to control the availability, reactivity and flux of iron in order to maintain a healthy cellular environment [144].

The initial step for achieving homeostasis is through the regulation of iron absorption from the gut. However the process of transporting iron to usage and storage sites is equally important, in addition to the roles of enterocytes and macrophages [144]. Monocytes and macrophages utilize different pathways to acquire iron. These methods include transferrin-mediated uptake, transmembrane uptake of ferrous and ferric iron, obtaining iron through lactoferrin or ferritin

receptors, as well as through erythrophagocytosis. As a result, the proliferation and differentiation of these cells are not affected by limiting the iron supply through one of these sources [145]. While iron is important and deficiencies result in aberrant cell growth and immune function, iron overload is similarly deleterious [144-145], affecting the proliferation and activation of T-cells, B-cells and natural killer cells [145, 172-173]. One mechanism through which iron loading can affect cells is by inhibiting *IFN-*γ mediated pathways in macrophages, which causes them to lose their ability to kill intracellular pathogens [145]. Moreover, the lack of an iron excretory pathway highlights the importance of homeostatic mechanisms adopted by cells in order to balance out iron needs as opposed to iron overload as well as redox utility as opposed to resulting toxicity.

Microarray data was retrieved to identify the alterations that the three viral infections considered, namely HIV, hepatitis C and influenza A induce, focusing interest on the differentially expressed iron binding proteins. HIV and hepatitis C are persistent diseases that reside in the host and cannot be cured. While influenza A infections can at times be fatal, the majority of people display uncomplicated, acute febrile respiratory symptoms that last around three to five days or exhibit no symptoms at all [174]. Hence, influenza was chosen as a control by comparing the changes this infection instigates to those brought about by HIV and/or hepatitis C. This allows for a direct investigation of the role of iron in supporting general viral mechanisms. Results revealed the significant changes at the transcriptional level of 101, 122 and 107 iron binding proteins in HIV, hepatitis C and influenza A, respectively. Comparisons among these lists indicated over 50% overlap. However, statistically only the overlap between HIV and hepatitis C was significant. Further investigation of these three lists revealed the inclusion of several cytochrome P450 proteins, which were enriched in several metabolic pathways (Table 8). This indicates the commonalities in the changes these viruses inflict on the cell to exploit the host's machinery.

However, as noted above, some of the main known HIV-1 host protein targets that bind iron displayed similar changes in expression level in hepatitis C but not in influenza infected cells. This implicates that hepatitis C possibly utilizes similar mechanisms to evade the host's immune response. Among those proteins was *CYBB* which is part of the NADPH oxidase enzymatic complex. NADPH oxidase is the main producer of superoxide anion (O_2^-) through the reduction of oxygen. In the cell, superoxide dismutase then acts as an antioxidant by utilizing electrons from copper or zinc for the conversion of superoxide into hydrogen peroxide (H_2O_2). In resting cells, the NADPH oxidase complex is typically dormant. Monocytes and macrophages usually release increased levels of reactive oxygen species (ROS) as a response to certain stimuli. The generation of high levels of ROS, referred to as respiratory burst, plays an important role in the host defense mechanism against pathogens [169, 175]. These reactive species are therefore involved in inflammatory processes, apoptosis, aging and carcinogenesis [176]. Iron is essential for the function of the NADPH oxidase complex with a heme-b acting as the prosthetic redox group in cytochrome b. Iron deficiencies therefore result in reduced enzyme activity [177].

HIV-1 targets NADPH oxidase through various proteins. First, *Gp120* binds to CXC chemokine receptor 4 (*CXCR4*) which in turn activates the NAPDH oxidase complex resulting in increased expression of superoxide radicals and subsequent activation of neutral sphingomyelinase, inducing apoptosis and cell death [178]. On the other hand, *Nef* plays a time-dependent role in this process. In the early stages, *Nef* is responsible for the induction of phosphorylation and cell-membrane translocation of *NCF1* and *NCF2*, hence activating NADPH oxidase, which results in the production of superoxide [169, 175]. Meanwhile, *Gp160* also enhances the respiratory burst and oxidative stress through the production of H_2O_2 [176]. Within 10 hours, *Nef* inhibits NADPH oxidase resulting in a dysregulation in the production of ROS, impairing specific immune functions including the oxidative burst response and phagocytosis. This in turn allows for the development of HIV-1 pathogenesis [169, 175]. In addition, the viral capsid has been shown to

inhibit the *IFN-γ* induced accumulation of the cytochrome B heavy chain mRNA [179], and this inhibition is evident from the observed downregulation of *CYBB* in the microarray analysis.

On the other hand, arachidonate 5-lipoxygenase (*ALOX5*) is a nonheme iron-containing dioxygenase that plays an important role in the biosynthesis of leukotrienes, namely the catalysis of the production of leukotriene LTA4 from arachidonic acid, which can then be converted to LTB4 [180]. Leukotrienes are important inflammatory mediators and LTB4 can then induce the adhesion and activation of leukocytes, *ALOX5* is therefore mainly expressed in the different leukocytes [181]. In addition, *ALOX5* might be capable of inducing cell cytotoxicity by oxidizing cellular membranes [182-183]. Leukotriene synthesis is reduced in the macrophages and peripheral mononuclear cells of HIV patients [184-186], as is supported by similar observation from the microarray analysis results on CD4+ T cells.

Microarray data also revealed the significant elevation in ATP-binding cassette protein (*ABCE1*) levels in CD4+ T cells of HIV-1 infected patients as compared to normal. Typically, *ABCE1* is required for cellular survival, mRNA translation, and ribosome biogenesis. It is the only ATP-binding cassette enzyme that has an amino-terminal iron-sulfur cluster domain, thus necessitating the availability of iron for its functioning [143, 187]. During HIV-1 infection, cellular *ABCE1* interacts with viral *Pr55* (*Gag*) and *Vif* to assist in capsid assembly. While *Vif* is excluded from the mature viral particles, it is essential for viral infectivity. It is therefore a late HIV-1 product, acting in the latter stages of the virus life cycle during viral assembly and/or maturation to enhance the infectivity of the progeny virions [188-189]. *ABCE1* is known to function as an RNAse L inhibitor, suggesting that the viral association with *ABCE1* is possibly to protect the viral RNA from degradation during viral assembly [188]. HIV-1 *Gag* polypeptides are synthesized in the cytoplasm of infected cells and then are trafficked to the plasma membrane. *ABCE1* is then recruited to sites of assembling *Gag* at the membrane and the association

continues throughout capsid formation until the onset of viral maturation and its subsequent release [190].

More literature has been curated for HIV's interactions with host proteins; however, similarities in microarray expression can insinuate comparable mechanisms utilized by hepatitis C. However, the consequences associated with iron overload have been confirmed by much research in both HIV [143, 152-160] and hepatitis C [143, 149, 191-193]. In the latter, not only does this overload correlate with progression of liver disease, fibrosis and carcinoma, it also results in a decreased response to antiviral therapy [149]. Nonetheless, despite all the efforts aimed at understanding the full role of iron in facilitating such viral infections, the mechanism and molecular explanation for the involvement of iron in these viral infection remains to be incomplete [143, 149].

5.6 Conclusion

When HIV and hepatitis C infections hijack the host's machinery, a complex interplay occurs between viral proteins and the host's immune system and iron homeostasis. Infecting viruses have to also posses the ability to enhance cellular metabolism in order to replicate their genome and proteins. As these processes require iron, the virus has to ensure that the iron supply meets its proliferative demands. Microarray analysis of the influence of HIV, hepatitis C and influenza A on iron binding proteins revealed that such proteins are major targets, whether direct or indirect, of viral infections. In addition, while these iron binding proteins could comprise a general theme for viral infections, differences are observed between some of the major genes and proteins affected by persistent and non-persistent infections. Some of these variations, in turn, are crucial for persistent viral survival and their abilities to avert the host's immune system response enabling them to continue to reside within the host.

Chapter 6: Concluding Remarks

Complex diseases and infections are characterized by the multiplicity of genes and pathways that are altered within the patient or host to ensure disease or pathogenic persistence. The advances in microarray technology has provided for a fast assay of the changes in gene regulation accompanied by the introduction of a multitude of perturbations including diseases, pathogens, drugs, gene knockouts and environmental modifications. Such data can provide a quantitative profile of the expression of thousands of genes in a single experiment. Biological significance can then be extended to genes of interest to identify the processes and pathways that succumb to the influence of these perturbations.

DNA gene expression analysis has been widely used in the study of cancer [36-40], but has generated many inconsistencies across studies. Apart from the lab-specific noise in the data that could interfere with the results, these disparities can arise because the number of samples used in studies does not meet the statistical requirements that support the thousands of genes that are assessed within the experiment. However, as more data is deposited into publically accessible databases, researchers can acquire large amounts of data across hundreds of labs to analyze and to add statistical significance and confidence to their results. Consequently, this research focused on an integrative method that could utilize data from different but similar platforms. The research allows for data merger prior to analysis, contradictory to the common meta-analysis methods that combine results after datasets are analyzed separately. Not only does the proposed method have the advantage of reducing experimental noise, it also takes advantage of the changes in sample distributions that have emerged in recent studies. Therefore, this methodology is not restricted to datasets containing both control and test data, but can select for data from any experiment and

utilize the samples that are of interest, provided the experiment has been hybridized on one of the sister Affymetrix platforms; the HG-U133A, HG-U133A2, and HG-U133 Plus 2.0.

The integrated method was based on the previously verified SAM statistical analysis approach as its performance is superior to that of other statistical methods, like the t-test and fold change [75]. In addition, SAM allows for controlling the different parameters including p-values and fold change. Testing the merged SAM approach on five cancer tissues: colon, kidney, liver, lung and pancreas, revealed its ability to capture large amounts of the experimental literature available on cancer that is independent of microarray usage. Moreover, it has surpassed the capabilities of the inverse-variance meta-analysis technique applied to the same data. This is supported by Nadon & Shoemaker [135] who noted that normalizing samples together adds robustness when compared to samples from datasets that have been normalized independently. To understand the significance of the gene lists obtained from the independent tissue analyses, pathways enriched with these genes were identified. The compliment and coagulation cascades and the ECM-receptor interactions were among the common pathways associated with these cancers. Moreover, a combined normal/cancer analysis revealed the important aberrations that occur within the cell cycle, including the differential expression of cyclins A and B and cyclin-dependent kinases (CDK1 and CDK4/6 complex), which are necessary for cell cycle progression. The merged SAM approach application in cancer was expanded to include an additional eight types of tissues to investigate the similarities in gene expression changes that occur across these tissues. The cancer samples included for each tissue were not restricted to a specific grade or type, this allows for the identification of cancer type-independent features. These common genes can prove to be essential drug targets for general cancer therapy and a few are already targeted by existing drugs while others are in the experimental phase.

The scope and utility of this research, however, is not limited to cancer investigations. Therefore, its application was directed to the study of persistent and non-persistent viral infections. In order for an infectious disease to replicate within the host's cells, it must establish a connection with the host. This crosstalk is essential for the different stages of the infection starting with the virus' ability to bind to and fuse into a cell and ending with its escape from the cell to spread to other cells and hosts. However, in between these two points the virus has to employ the cell's machinery through further viral-host protein-protein interactions. These interactions then allow the virus to replicate its genome, but it also triggers other pathways such as immune response and apoptosis, especially for non-persistent infections. Persistent viruses, on the other hand, have developed ways to escape the immune system and ensure their continued existence within the host.

Much of these biological processes and pathways, varying from DNA replication to immune response, involve proteins that bind iron. DNA microarray data was therefore used to identify those iron binding proteins that exhibited altered expression at the transcript level due to viral infections. This is especially important since iron overload occurring in patients with persistent viral infections is believed to correlate with disease progression and decreased survival in hepatitis C and HIV, respectively. Merged microarray results confirmed some of the known effects HIV exerts on the host's proteins it directly interacts with and similar results were observed under hepatitis C infection, suggesting comparable hijacking approaches. However, comparison of results with influenza A, a non-persistent infection chosen as a control, revealed that iron-dependent proteins are the target of viruses in general, as these proteins are mainly involved in metabolic processes. Nonetheless, some of the genes altered exclusively in persistent viruses could be related to viral lengthened survival and escaping the host's immune response.

The shortcomings however arise from the disparities in focus on the diseases studied. The majority of the currently available gene expression experiments are dedicated towards the study of cancer as opposed to infectious diseases and other pathologies. For example, thousands of microarray samples exist containing data on a wide range of malignancies, however, only tens of samples exist for HIV infections. Even within cancer itself, a great effort has been devoted to certain malignancies such as breast cancer, where thousands of data points are available for each gene, compared to adrenal, pancreatic and cervical cancers that have far fewer samples. Another limitation is also imposed by the platform restrictions required by the merged SAM. However, Affymetrix has been shown to result in better accuracy than other platforms [194]. Moreover, integrating data from different technologies (Affymetrix/cDNA) is unreliable [133-134].

The molecular pathways affected by cancer also involve noncoding genes; these include the noncoding class of small RNAs referred to as microRNA (miRNA). While these miRNAs are not translated into proteins, they are instead involved in the regulation of mRNA translation [195-196]. Functionally, these miRNAs can reduce the amount of proteins produced by their target transcripts, and thus they are essential for several biological pathways including cell proliferation [195]. Cancer can affect the expression of miRNA through mutations, polymorphisms, chromosomal abnormalities and epigenetic changes. These in turn result in defects in the miRNA biogenesis machinery. Such changes in turn can promote oncogenesis by altering the gene expression of oncogenes and tumor suppressor genes [196]. Future work will add to the obtained cancer knowledge by exploiting changes at the miRNA level associated with the different malignancies. Not only are these abnormal miRNA profiles essential for cancer diagnosis, they can also play a dual role in cancer therapy. Since the miRNA profiles are conserved from the primary tumor to the metastatic cancer, cancer therapy can utilize miRNA as therapeutics and also target them through anti-miRNA therapy [196].

List of References

1. Hatzimanikatis V: Bioinformatics and functional genomics: Challenges and opportunities. *Aiche Journal* 2000, 46:2339-2343.

2. Mathews CK, Van Holde KE, Ahern KG: *Biochemistry.* 3rd edn. San Francisco, Calif.: Benjamin/Cummings; 2000.

3. Bolsover SR, NetLibrary Inc.: Cell biology a short course. In *Book Cell biology a short course* (Editor ed.^eds.), 2nd edition. City: Wiley; 2004.

4. Antisense DNA Oligonucleotide [http://commons.wikimedia.org/wiki/File:Antisense_DNA_oligonucleotide.png]

5. Lee M, Mahato RI: Gene regulation for effective gene therapy. Preface. *Adv Drug Deliv Rev* 2009, 61:487-488.

6. Tajmir-Riahi HA: An overview of protein-DNA and protein-RNA interactions. *Journal of the Iranian Chemical Society* 2006, 3:297-304.

7. Bertram JS: The molecular biology of cancer. *Mol Aspects Med* 2000, 21:167-223.

8. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al: Global variation in copy number in the human genome. *Nature* 2006, 444:444-454.

9. Vogelstein B, Kinzler KW: Cancer genes and the pathways they control. *Nat Med* 2004, 10:789-799.

10. Jaenisch R, Bird A: Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* 2003, 33 Suppl:245-254.

11. Cancer Requires Multiple Mutations [http://commons.wikimedia.org/wiki/File:Cancer_requires_multiple_mutations_from_NIHen.png]

12. Moss SF, Blaser MJ: Mechanisms of disease: Inflammation and the origins of cancer. *Nat Clin Pract Oncol* 2005, 2:90-97; quiz 91 p following 113.

13. Tan WY, Hanin LG: *Handbook of cancer models with applications.* Singapore ; Hackensack, NJ: World Scientific; 2008.

14. Kumar V, Abbas AK, Fausto N, Robbins SL, Cotran RS: *Robbins and Cotran pathologic basis of disease.* 7th edn. Philadelphia: Elsevier Saunders; 2005.

15. Hilleman MR: Strategies and mechanisms for host and pathogen survival in acute and persistent viral infections. *Proc Natl Acad Sci U S A* 2004, 101 Suppl 2:14560-14566.

16. Rambaut A, Posada D, Crandall KA, Holmes EC: The causes and consequences of HIV evolution. *Nat Rev Genet* 2004, 5:52-61.

17. Sharp PM, Bailes E, Robertson DL, Gao F, Hahn BH: Origins and evolution of AIDS viruses. *Biol Bull* 1999, 196:338-342.

18. Sousa AE, Carneiro J, Meier-Schellersheim M, Grossman Z, Victorino RM: CD4 T cell depletion is linked directly to immune activation in the pathogenesis of HIV-1 and HIV-2 but only indirectly to the viral load. *J Immunol* 2002, 169:3400-3406.

19. Reeves JD, Doms RW: Human immunodeficiency virus type 2. *J Gen Virol* 2002, 83:1253-1265.

20. Peterlin BM, Trono D: Hide, shield and strike back: how HIV-infected cells avoid immune eradication. *Nat Rev Immunol* 2003, 3:97-107.

21. Wu L, KewalRamani VN: Dendritic-cell interactions with HIV: infection and viral dissemination. *Nat Rev Immunol* 2006, 6:859-868.

22. Gougeon ML: Apoptosis as an HIV strategy to escape immune attack. *Nat Rev Immunol* 2003, 3:392-404.

23. Memon MI, Memon MA: Hepatitis C: an epidemiological review. *J Viral Hepat* 2002, 9:84-100.

24. Bonkovsky HL, Mehta S: Hepatitis C: a review and update. *J Am Acad Dermatol* 2001, 44:159-182.

25. Tellinghuisen TL, Rice CM: Interaction between hepatitis C virus proteins and host cell factors. *Curr Opin Microbiol* 2002, 5:419-427.

26. McLauchlan J: Properties of the hepatitis C virus core protein: a structural protein that modulates cellular processes. *J Viral Hepat* 2000, 7:2-14.

27. Wolfl M, Rutebemberwa A, Mosbruger T, Mao Q, Li HM, Netski D, Ray SC, Pardoll D, Sidney J, Sette A, et al: Hepatitis C virus immune escape via exploitation of a hole in the T cell repertoire. *J Immunol* 2008, 181:6435-6446.

28. Lauer GM, Walker BD: Hepatitis C virus infection. *N Engl J Med* 2001, 345:41-52.

29. Hay AJ, Gregory V, Douglas AR, Lin YP: The evolution of human influenza viruses. *Philos Trans R Soc Lond B Biol Sci* 2001, 356:1861-1870.

30. Zambon MC: Epidemiology and pathogenesis of influenza. *J Antimicrob Chemother* 1999, 44 Suppl B:3-9.

31. Zambon MC: The pathogenesis of influenza in humans. *Rev Med Virol* 2001, 11:227-241.

32. Mettenleiter TC, Sobrino F: *Animal viruses : molecular biology.* Norfolk, UK: Caister Academic Press; 2008.

33. Inoue E, Wang X, Osawa Y, Okazaki K: Full genomic amplification and subtyping of influenza A virus using a single set of universal primers. *Microbiol Immunol* 2010, 54:129-134.

34. Roxas M, Jurenka J: Colds and influenza: a review of diagnosis and conventional, botanical, and nutritional considerations. *Altern Med Rev* 2007, 12:25-48.

35. Ludwig S, Planz O, Pleschka S, Wolff T: Influenza-virus-induced signaling cascades: targets for antiviral therapy? *Trends Mol Med* 2003, 9:46-52.

36. Mehle A, Doudna JA: A host of factors regulating influenza virus replication. *Viruses* 2010, 2:566-573.

37. Wolff T, O'Neill RE, Palese P: Interaction cloning of NS1-I, a human protein that binds to the nonstructural NS1 proteins of influenza A and B viruses. *J Virol* 1996, 70:5363-5372.

38. Giordano TJ, Kuick R, Else T, Gauger PG, Vinco M, Bauersfeld J, Sanders D, Thomas DG, Doherty G, Hammer G: Molecular classification and prognostication of adrenocortical tumors by transcriptome profiling. *Clin Cancer Res* 2009, 15:668-676.

39. Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, Misra A, Nigro JM, Colman H, Soroceanu L, et al: Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 2006, 9:157-173.

40. Jones J, Otu H, Spentzos D, Kolia S, Inan M, Beecken WD, Fellbaum C, Gu X, Joseph M, Pantuck AJ, et al: Gene signatures of progression and metastasis in renal cell cancer. *Clin Cancer Res* 2005, 11:5730-5739.

41. Bachtiary B, Boutros PC, Pintilie M, Shi W, Bastianutto C, Li JH, Schwock J, Zhang W, Penn LZ, Jurisica I, et al: Gene expression profiling in cervical cancer: an exploration of intratumor heterogeneity. *Clin Cancer Res* 2006, 12:5632-5640.

42. Mas VR, Maluf DG, Archer KJ, Yanek K, Kong X, Kulik L, Freise CE, Olthoff KM, Ghobrial RM, McIver P, Fisher R: Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. *Mol Med* 2009, 15:85-94.

43. Turashvili G, Bouchal J, Baumforth K, Wei W, Dziechciarkova M, Ehrmann J, Klein J, Fridman E, Skarda J, Srovnal J, et al: Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC Cancer* 2007, 7:55.

44. Gyorffy B, Molnar B, Lage H, Szallasi Z, Eklund AC: Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples. *PLoS One* 2009, 4:e5645.

45. Riker AI, Enkemann SA, Fodstad O, Liu S, Ren S, Morris C, Xi Y, Howell P, Metge B, Samant RS, et al: The gene expression profiles of primary and metastatic melanoma

yields a transition point of tumor progression and metastasis. *BMC Med Genomics* 2008, 1:13.

46. Ramasamy A, Mondry A, Holmes CC, Altman DG: Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* 2008, 5:e184.

47. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res* 2002, 62:4427-4433.

48. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A* 2004, 101:9309-9314.

49. Hyrcza MD, Kovacs C, Loutfy M, Halpenny R, Heisler L, Yang S, Wilkins O, Ostrowski M, Der SD: Distinct transcriptional profiles in ex vivo CD4+ and CD8+ T cells are established early in human immunodeficiency virus type 1 infection and are characterized by a chronic interferon response as well as extensive transcriptional changes in CD8+ T cells. *J Virol* 2007, 81:3477-3486.

50. Sedaghat AR, German J, Teslovich TM, Cofrancesco J, Jr., Jie CC, Talbot CC, Jr., Siliciano RF: Chronic CD4+ T-cell activation and depletion in human immunodeficiency virus type 1 infection: type I interferon-mediated disruption of T-cell dynamics. *J Virol* 2008, 82:1870-1883.

51. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, Wittkowski KM, Piqueras B, Banchereau J, Palucka AK, Chaussabel D: Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* 2007, 109:2066-2077.

52. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO, 3rd, Lucas J, Huang Y, Turner R, Gilbert A, Lambkin-Williams R, et al: Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. *Cell Host Microbe* 2009, 6:207-217.

53. Lockhart DJ, Winzeler EA: Genomics, gene expression and DNA arrays. *Nature* 2000, 405:827-836.

54. Mills JC, Roth KA, Cagan RL, Gordon JI: DNA microarrays and beyond: completing the journey from tissue to cell. *Nat Cell Biol* 2001, 3:E175-178.

55. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003, 31:e15.

56. Yang YH, Speed T: Design issues for cDNA microarray experiments. *Nat Rev Genet* 2002, 3:579-588.

57. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: Expression profiling using cDNA microarrays. *Nat Genet* 1999, 21:10-14.

58. Celis JE, Kruhoffer M, Gromova I, Frederiksen C, Ostergaard M, Thykjaer T, Gromov P, Yu J, Palsdottir H, Magnusson N, Orntoft TF: Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett* 2000, 480:2-16.

59. Schulze A, Downward J: Navigating gene expression using microarrays--a technology review. *Nat Cell Biol* 2001, 3:E190-195.

60. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA: Maximum likelihood estimation of optimal scaling factors for expression array normalization. *SPIE BIOS* 2001.

61. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003, 19:185-193.

62. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003, 4:249-264.

63. Katz S, Irizarry RA, Lin X, Tripputi M, Porter MW: A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC Bioinformatics* 2006, 7:464.

64. Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli GA, Bicciato S: Novel definition files for human GeneChips based on GeneAnnot. *BMC Bioinformatics* 2007, 8:446.

65. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, et al: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 2005, 33:e175.

66. Barrett JC, Kawasaki ES: Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. *Drug Discov Today* 2003, 8:134-141.

67. Wei C, Li J, Bumgarner RE: Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics* 2004, 5:87.

68. Allison DB, Cui X, Page GP, Sabripour M: Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 2006, 7:55-65.

69. Leung YF, Cavalieri D: Fundamentals of cDNA microarray data analysis. *Trends Genet* 2003, 19:649-659.

70. Lonnstedt I, Speed TP: ReplicatedMicroarray Data. *Stat Sinicia* 2002, 12:31-46.

71. Kerr MK, Martin M, Churchill GA: Analysis of variance for gene expression microarray data. *J Comput Biol* 2000, 7:819-837.

72. Wu TD: Analysing gene expression data from DNA microarrays to identify candidate genes. *J Pathol* 2001, 195:53-65.

73. Long AD, Mangalam HJ, Chan BY, Tolleri L, Hatfield GW, Baldi P: Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework. Analysis of global gene expression in Escherichia coli K12. *J Biol Chem* 2001, 276:19937-19944.

74. Baldi P, Long AD: A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 2001, 17:509-519.

75. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001, 98:5116-5121.

76. Zintzaras E, Ioannidis JP: Meta-analysis for ranked discovery datasets: theoretical framework and empirical demonstration for microarrays. *Comput Biol Chem* 2008, 32:38-46.

77. Lorenz MW, Markus HS, Bots ML, Rosvall M, Sitzer M: Prediction of clinical cardiovascular events with carotid intima-media thickness: a systematic review and meta-analysis. *Circulation* 2007, 115:459-467.

78. Elliott WJ, Meyer PM: Incident diabetes in clinical trials of antihypertensive drugs: a network meta-analysis. *Lancet* 2007, 369:201-207.

79. Sullivan PF, Neale MC, Kendler KS: Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry* 2000, 157:1552-1562.

80. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, Moher D, Becker BJ, Sipe TA, Thacker SB: Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA* 2000, 283:2008-2012.

81. Sutton AJ: *Methods for meta-analysis in medical research.* Chichester ; New York: J. Wiley; 2000.

82. Whitehead A: *Meta-analysis of controlled clinical trials.* Chichester ; New York: John Wiley & Sons; 2002.

83. Choi JK, Yu U, Kim S, Yoo OJ: Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 2003, 19 Suppl 1:i84-90.

84. Choi JK, Choi JY, Kim DG, Choi DW, Kim BY, Lee KH, Yeom YI, Yoo HS, Yoo OJ, Kim S: Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Lett* 2004, 565:93-100.

85. Yang X, Bentink S, Spang R: Detecting common gene expression patterns in multiple cancer outcome entities. *Biomed Microdevices* 2005, 7:247-251.

86. DeConde RP, Hawley S, Falcon S, Clegg N, Knudsen B, Etzioni R: Combining results of microarray experiments: a rank aggregation approach. *Stat Appl Genet Mol Biol* 2006, 5:Article15.

87. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: Coexpression analysis of human genes across many microarray data sets. *Genome Res* 2004, 14:1085-1094.

88. Stuart JM, Segal E, Koller D, Kim SK: A gene-coexpression network for global discovery of conserved genetic modules. *Science* 2003, 302:249-255.

89. Wan X, Pavlidis P: Sharing and reusing gene expression profiling data in neuroscience. *Neuroinformatics* 2007, 5:161-175.

90. Fisher RA: *Statistical methods for research workers.* 5th edn. Edinburgh,: Oliver and Boyd; 1932.

91. Fleiss JL: The statistical basis of meta-analysis. *Stat Methods Med Res* 1993, 2:121-145.

92. Grutzmann R, Boriss H, Ammerpohl O, Luttges J, Kalthoff H, Schackert HK, Kloppel G, Saeger HD, Pilarsky C: Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. *Oncogene* 2005, 24:5079-5088.

93. Cohen J: *Statistical power analysis for the behavioral sciences.* 2nd edn. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

94. Hedges LV, Olkin I: *Statistical methods for meta-analysis.* Orlando: Academic Press; 1985.

95. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, et al: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001, 29:365-371.

96. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* 2007, 35:D760-765.

97. Edgar R, Domrachev M, Lash AE: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002, 30:207-210.

98. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, et al: ArrayExpress--a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* 2003, 31:68-71.

99. Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, 28:27-30.

100. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004, 32:D277-280.

101. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006, 34:D354-357.

102. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000, 25:25-29.

103. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al: The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004, 32:D258-261.

104. Smid M, Dorssers LC, Jenster G: Venn Mapping: clustering of heterologous microarray data based on the number of co-occurring differentially expressed genes. *Bioinformatics* 2003, 19:2065-2071.

105. Pihur V, Datta S: RankAggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* 2009, 10:62.

106. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J: RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 2006, 22:2825-2827.

107. Warnat P, Eils R, Brors B: Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* 2005, 6:265.

108. Hu P GC, Beyene J. Statistical methods for meta-analysis of microarray data: A comparative study. Inf Syst Front 2006; 8: 9-20.

109. Xu L, Tan AC, Winslow RL, Geman D: Merging microarray data from separate breast cancer studies provides a robust prognostic test. *BMC Bioinformatics* 2008, 9:125.

110. Ertel A, Tozeren A: Human and mouse switch-like genes share common transcriptional regulatory mechanisms for bimodality. *BMC Genomics* 2008, 9:628.

111. Ertel A, Tozeren A: Switch-like genes populate cell communication pathways and are enriched for extracellular proteins. *BMC Genomics* 2008, 9:3.

112. Gormley M, Tozeren A: Expression profiles of switch-like genes accurately classify tissue and infectious disease phenotypes in model-based classification. *BMC Bioinformatics* 2008, 9:486.

113. Sanga S, Broom BM, Cristini V, Edgerton ME: Gene expression meta-analysis supports existence of molecular apocrine breast cancer with a role for androgen receptor and implies interactions with ErbB family. *BMC Med Genomics* 2009, 2:59.

114. Gorlov IP, Byun J, Gorlova OY, Aparicio AM, Efstathiou E, Logothetis CJ: Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data. *BMC Med Genomics* 2009, 2:48.

115. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, Desmedt C, Ignatiadis M, Sengstag T, Schutz F, et al: Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008, 10:R65.

116. Xu L, Geman D, Winslow RL: Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics* 2007, 8:275.

117. Dai MH, Wang PL, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, et al: Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res* 2005, 33:-.

118. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing V, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

119. Tibshirani R CG, Hastie T, Narasimhan B. samr: SAM: Significance Analysis of Microarrays. R package version 1.26. http://www-stat.stanford.edu/~tibs/SAM.

120. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009, 4:44-57.

121. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* 2003, 4:P3.

122. Ertel A, Verghese A, Byers SW, Ochs M, Tozeren A: Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Mol Cancer* 2006, 5:55.

123. Hong Y, Ho KS, Eu KW, Cheah PY: A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res* 2007, 13:1107-1114.

124. Su LJ, Chang CW, Wu YC, Chen KC, Lin CJ, Liang SC, Lin CH, Whang-Peng J, Hsu SL, Chen CH, Huang CY: Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics* 2007, 8:140.

125. Galamb O, Spisak S, Sipos F, Toth K, Solymosi N, Wichmann B, Krenacs T, Valcz G, Tulassay Z, Molnar B: Reversal of gene expression changes in the colorectal normal-adenoma pathway by NS398 selective COX2 inhibitor. *Br J Cancer* 2010, 102:765-773.

126. Yap YL, Lam DC, Luc G, Zhang XW, Hernandez D, Gras R, Wang E, Chiu SW, Chung LP, Lam WK, et al: Conserved transcription factor binding sites of cancer markers derived from primary lung adenocarcinoma microarrays. *Nucleic Acids Res* 2005, 33:409-421.

127. Scotto L, Narayan G, Nandula SV, Arias-Pulido H, Subramaniyam S, Schneider A, Kaufmann AM, Wright JD, Pothuri B, Mansukhani M, Murty VV: Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. *Genes Chromosomes Cancer* 2008, 47:755-765.

128. Park MT, Lee SJ: Cell cycle and cancer. *J Biochem Mol Biol* 2003, 36:60-65.

129. Groene J, Mansmann U, Meister R, Staub E, Roepcke S, Heinze M, Klaman I, Brummendorf T, Hermann K, Loddenkemper C, et al: Transcriptional census of 36 microdissected colorectal cancers yields a gene signature to distinguish UICC II and III. *Int J Cancer* 2006, 119:1829-1836.

130. Lin YH, Friederichs J, Black MA, Mages J, Rosenberg R, Guilford PJ, Phillips V, Thompson-Fawcett M, Kasabov N, Toro T, et al: Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer. *Clin Cancer Res* 2007, 13:498-507.

131. Turkheimer FE, Roncaroli F, Hennuy B, Herens C, Nguyen M, Martin D, Evrard A, Bours V, Boniver J, Deprez M: Chromosomal patterns of gene expression from microarray data: methodology, validation and clinical relevance in gliomas. *BMC Bioinformatics* 2006, 7:526.

132. Marty B, Maire V, Gravier E, Rigaill G, Vincent-Salomon A, Kappler M, Lebigot I, Djelti F, Tourdes A, Gestraud P, et al: Frequent PTEN genomic alterations and activated phosphatidylinositol 3-kinase pathway in basal-like breast cancer cells. *Breast Cancer Res* 2008, 10:R101.

133. Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O: Are data from different gene expression microarray platforms comparable? *Genomics* 2004, 83:1164-1168.

134. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 2002, 18:405-412.

135. Nadon R, Shoemaker J: Statistical issues with microarrays: processing and analysis. *Trends Genet* 2002, 18:265-271.

136. Dyrskjot L, Kruhoffer M, Thykjaer T, Marcussen N, Jensen JL, Moller K, Orntoft TF: Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. *Cancer Res* 2004, 64:4040-4048.

137. Corvol JC, Pelletier D, Henry RG, Caillier SJ, Wang J, Pappas D, Casazza S, Okuda DT, Hauser SL, Oksenberg JR, Baranzini SE: Abrogation of T cell quiescence characterizes patients at high risk for multiple sclerosis after the initial neurological event. *Proc Natl Acad Sci U S A* 2008, 105:11839-11844.

138. Falt S, Merup M, Gahrton G, Lambert B, Wennborg A: Identification of progression markers in B-CLL by gene expression profiling. *Exp Hematol* 2005, 33:883-893.

139. Gormley M, Dampier W, Ertel A, Karacali B, Tozeren A: Prediction potential of candidate biomarker sets identified and validated on gene expression data from multiple datasets. *BMC Bioinformatics* 2007, 8:415.

140. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008, 36:D901-906.

141. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J: DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* 2006, 34:D668-672.

142. Bisognin A, Coppe A, Ferrari F, Risso D, Romualdi C, Bicciato S, Bortoluzzi S: A-MADMAN: annotation-based microarray data meta-analysis tool. *BMC Bioinformatics* 2009, 10:201.

143. Drakesmith H, Prentice A: Viral infection and iron metabolism. *Nat Rev Microbiol* 2008, 6:541-552.

144. Doherty CP: Host-pathogen interactions: the role of iron. *J Nutr* 2007, 137:1341-1344.

145. Weiss G: Iron and immunity: a double-edged sword. *Eur J Clin Invest* 2002, 32 Suppl 1:70-78.

146. Weiss G, Wachter H, Fuchs D: Linkage of cell-mediated immunity to iron metabolism. *Immunol Today* 1995, 16:495-500.

147. Seligman PA, Kovar J, Gelfand EW: Lymphocyte proliferation is controlled by both iron availability and regulation of iron uptake pathways. *Pathobiology* 1992, 60:19-26.

148. Savarino A, Pescarmona GP, Boelaert JR: Iron metabolism and HIV infection: reciprocal interactions with potentially harmful consequences? *Cell Biochem Funct* 1999, 17:279-287.

149. Franchini M, Targher G, Capra F, Montagnana M, Lippi G: The effect of iron depletion on chronic hepatitis C virus infection. *Hepatol Int* 2008, 2:335-340.

150. Boelaert JR, Weinberg GA, Weinberg ED: Altered iron metabolism in HIV infection: mechanisms, possible consequences, and proposals for management. *Infect Agents Dis* 1996, 5:36-46.

151. de Monye C, Karcher DS, Boelaert JR, Gordeuk VR: Bone marrow macrophage iron grade and survival of HIV-seropositive patients. *AIDS* 1999, 13:375-380.

152. Gordeuk VR, Delanghe JR, Langlois MR, Boelaert JR: Iron status and the outcome of HIV infection: an overview. *J Clin Virol* 2001, 20:111-115.

153. Clark TD, Semba RD: Iron supplementation during human immunodeficiency virus infection: a double-edged sword? *Med Hypotheses* 2001, 57:476-479.

154. Friis H, Gomo E, Nyazema N, Ndhlovu P, Krarup H, Madsen PH, Michaelsen KF: Iron, haptoglobin phenotype, and HIV-1 viral load: a cross-sectional study among pregnant Zimbabwean women. *J Acquir Immune Defic Syndr* 2003, 33:74-81.

155. Costagliola DG, de Montalembert M, Lefrere JJ, Briand C, Rebulla P, Baruchel S, Dessi C, Fondu P, Karagiorga M, Perrimond H, et al.: Dose of desferrioxamine and evolution of HIV-1 infection in thalassaemic patients. *Br J Haematol* 1994, 87:849-852.

156. Delanghe JR, Langlois MR, Boelaert JR, Van Acker J, Van Wanzeele F, van der Groen G, Hemmer R, Verhofstede C, De Buyzere M, De Bacquer D, et al: Haptoglobin polymorphism, iron metabolism and mortality in HIV infection. *AIDS* 1998, 12:1027-1032.

157. Olsen A, Mwaniki D, Krarup H, Friis H: Low-dose iron supplementation does not increase HIV-1 load. *J Acquir Immune Defic Syndr* 2004, 36:637-638.

158. Salmon-Ceron D, Fontbonne A, Saba J, May T, Raffi F, Chidiac C, Patey O, Aboulker JP, Schwartz D, Vilde JL: Lower survival in AIDS patients receiving dapsone compared with aerosolized pentamidine for secondary prophylaxis of Pneumocystis carinii pneumonia. Study Group. *J Infect Dis* 1995, 172:656-664.

159. Jacobus DP: Randomization to iron supplementation of patients with advanced human immunodeficiency virus disease--an inadvertent but controlled study with results important for patient care. *J Infect Dis* 1996, 173:1044-1045.

160. Weinberg GA: Iron overload as a mechanism for the lowered survival in AIDS patients receiving dapsone-iron protoxalate for secondary prophylaxis of Pneumocystis carinii pneumonia. *J Infect Dis* 1996, 174:241-242.

161. Pinney JW, Dickerson JE, Fu W, Sanders-Beer BE, Ptak RG, Robertson DL: HIV-host interactions: a map of viral perturbation of the host system. *AIDS* 2009, 23:549-554.

162. Ptak RG, Fu W, Sanders-Beer BE, Dickerson JE, Pinney JW, Robertson DL, Rozanov MN, Katz KS, Maglott DR, Pruitt KD, Dieffenbach CW: Cataloguing the HIV type 1 human protein interaction network. *AIDS Res Hum Retroviruses* 2008, 24:1497-1502.

163. Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG: Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res* 2009, 37:D417-422.

164. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E: Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 2009, 10:136.

165. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 2010, 38:D142-148.

166. Hartigan A, Wong M: Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics* 1979, 28:100-108.

167. Ameny MA, Raila J, Walzel E, Schweigert FJ: Effect of iron and/or vitamin A re-supplementation on vitamin A and iron status of rats after a dietary deficiency of both components. *J Trace Elem Med Biol* 2002, 16:175-178.

168. Guengerich FP: Cytochrome p450 and chemical toxicology. *Chem Res Toxicol* 2008, 21:70-83.

169. Olivetta E, Mallozzi C, Ruggieri V, Pietraforte D, Federico M, Sanchez M: HIV-1 Nef induces p47(phox) phosphorylation leading to a rapid superoxide anion release from the U937 human monoblastic cell line. *J Cell Biochem* 2009, 106:812-822.

170. Masse E, Arguin M: Ironing out the problem: new mechanisms of iron homeostasis. *Trends Biochem Sci* 2005, 30:462-468.

171. Isom HC, McDevitt EI, Moon MS: Elevated hepatic iron: a confounding factor in chronic hepatitis C. *Biochim Biophys Acta* 2009, 1790:650-662.

172. De Sousa M: T lymphocytes and iron overload: novel correlations of possible significance to the biology of the immunological system. *Mem Inst Oswaldo Cruz* 1992, 87 Suppl 5:23-29.

173. Brekelmans P, van Soest P, Leenen PJ, van Ewijk W: Inhibition of proliferation and differentiation during early T cell development by anti-transferrin receptor antibody. *Eur J Immunol* 1994, 24:2896-2902.

174. Harper SA, Bradley JS, Englund JA, File TM, Gravenstein S, Hayden FG, McGeer AJ, Neuzil KM, Pavia AT, Tapper ML, et al: Seasonal influenza in adults and children--diagnosis, treatment, chemoprophylaxis, and institutional outbreak management: clinical practice guidelines of the Infectious Diseases Society of America. *Clin Infect Dis* 2009, 48:1003-1032.

175. Olivetta E, Pietraforte D, Schiavoni I, Minetti M, Federico M, Sanchez M: HIV-1 Nef regulates the release of superoxide anions from human macrophages. *Biochem J* 2005, 390:591-602.

176. Lachgar A, Sojic N, Arbault S, Bruce D, Sarasin A, Amatore C, Bizzini B, Zagury D, Vuillaume M: Amplification of the inflammatory cellular redox state by human immunodeficiency virus type 1-immunosuppressive tat and gp160 proteins. *J Virol* 1999, 73:1447-1452.

177. Kurtoglu E, Ugur A, Baltaci AK, Mogolkoc R, Undar L: Activity of neutrophil NADPH oxidase in iron-deficient anemia. *Biol Trace Elem Res* 2003, 96:109-115.

178. Jana A, Pahan K: Human immunodeficiency virus type 1 gp120 induces apoptosis in human primary neurons through redox-regulated activation of neutral sphingomyelinase. *J Neurosci* 2004, 24:9531-9540.

179. Nong Y, Kandil O, Tobin EH, Rose RM, Remold HG: The HIV core protein p24 inhibits interferon-gamma-induced increase of HLA-DR and cytochrome b heavy chain mRNA levels in the human monocyte-like cell line THP1. *Cell Immunol* 1991, 132:10-16.

180. Feisst C, Pergola C, Rakonjac M, Rossi A, Koeberle A, Dodt G, Hoffmann M, Hoernig C, Fischer L, Steinhilber D, et al: Hyperforin is a novel type of 5-lipoxygenase inhibitor with high efficacy in vivo. *Cell Mol Life Sci* 2009, 66:2759-2771.

181. Radmark O, Werz O, Steinhilber D, Samuelsson B: 5-Lipoxygenase: regulation of expression and enzyme activity. *Trends Biochem Sci* 2007, 32:332-341.

182. Maccarrone M, Navarra M, Catani V, Corasaniti MT, Bagetta G, Finazzi-Agro A: Cholesterol-dependent modulation of the toxicity of HIV-1 coat protein gp120 in human neuroblastoma cells. *J Neurochem* 2002, 82:1444-1452.

183. Maccarrone M, Navarra M, Corasaniti MT, Nistico G, Finazzi Agro A: Cytotoxic effect of HIV-1 coat glycoprotein gp120 on human neuroblastoma CHP100 cells involves activation of the arachidonate cascade. *Biochem J* 1998, 333 (Pt 1):45-49.

184. Coffey MJ, Phare SM, Cinti S, Peters-Golden M, Kazanjian PH: Granulocyte-macrophage colony-stimulating factor upregulates reduced 5-lipoxygenase metabolism in peripheral blood monocytes and neutrophils in acquired immunodeficiency syndrome. *Blood* 1999, 94:3897-3905.

185. Coffey MJ, Phare SM, Kazanjian PH, Peters-Golden M: 5-Lipoxygenase metabolism in alveolar macrophages from subjects infected with the human immunodeficiency virus. *J Immunol* 1996, 157:393-399.

186. Coffey MJ, Phare SM, George S, Peters-Golden M, Kazanjian PH: Granulocyte colony-stimulating factor administration to HIV-infected subjects augments reduced leukotriene synthesis and anticryptococcal activity in neutrophils. *J Clin Invest* 1998, 102:663-670.

187. Rodnina MV: Protein synthesis meets ABC ATPases: new roles for Rli1/ABCE1. *EMBO Rep* 2010, 11:143-144.

188. Lake JA, Carr J, Feng F, Mundy L, Burrell C, Li P: The role of Vif during HIV-1 infection: interaction with novel host cellular factors. *J Clin Virol* 2003, 26:143-152.

189. Cullen BR: HIV-1 auxiliary proteins: making connections in a dying cell. *Cell* 1998, 93:685-692.

190. Dooher JE, Schneider BL, Reed JC, Lingappa JR: Host ABCE1 is at plasma membrane HIV assembly sites and its dissociation from Gag is linked to subsequent events of virus production. *Traffic* 2007, 8:195-211.

191. Kageyama F, Kobayashi Y, Kawasaki T, Toyokuni S, Uchida K, Nakamura H: Successful interferon therapy reverses enhanced hepatic iron accumulation and lipid peroxidation in chronic hepatitis C. *Am J Gastroenterol* 2000, 95:1041-1050.

192. Cho H, Lee HC, Jang SK, Kim YK: Iron increases translation initiation directed by internal ribosome entry site of hepatitis C virus. *Virus Genes* 2008, 37:154-160.

193. Angelucci E, Muretto P, Nicolucci A, Baronciani D, Erer B, Gaziev J, Ripalti M, Sodani P, Tomassoni S, Visani G, Lucarelli G: Effects of iron overload and hepatitis C virus positivity in determining progression of liver fibrosis in thalassemia following bone marrow transplantation. *Blood* 2002, 100:17-21.

194. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, et al: Multiple-laboratory comparison of microarray platforms. *Nat Methods* 2005, 2:345-350.

195. Calin GA, Croce CM: MicroRNA-cancer connection: the beginning of a new tale. *Cancer Res* 2006, 66:7390-7394.

196. Visone R, Croce CM: MiRNAs and cancer. *Am J Pathol* 2009, 174:1131-1138.

Appendix A: General cancer SAM genes

Table A1 - Annotation of General Cancer SAM Genes: List of genes that appeared to be differentially expressed in at least 70% of the iterations when selecting 10 random samples from each tissue for general normal to cancer tissue comparisons

Entrez ID	Frequency (%)	Gene Symbol	Gene Name
125	100	*ADH1B*	alcohol dehydrogenase 1B (class I), beta polypeptide
699	100	*BUB1*	BUB1 budding uninhibited by benzimidazoles 1 homolog (yeast)
701	100	*BUB1B*	BUB1 budding uninhibited by benzimidazoles 1 homolog beta (yeast)
705	100	*BYSL*	bystin-like
983	100	*CDC2*	cell division cycle 2, G1 to S and G2 to M
990	100	*CDC6*	cell division cycle 6 homolog (S. cerevisiae)
991	100	*CDC20*	cell division cycle 20 homolog (S. cerevisiae)
1019	100	*CDK4*	cyclin-dependent kinase 4
1058	100	*CENPA*	centromere protein A
1063	100	*CENPF*	centromere protein F, 350/400ka (mitosin)
1111	100	*CHEK1*	CHK1 checkpoint homolog (S. pombe)
1164	100	*CKS2*	CDC28 protein kinase regulatory subunit 2
1282	100	*COL4A1*	collagen, type IV, alpha 1
1300	100	*COL10A1*	collagen, type X, alpha 1(Schmid metaphyseal chondrodysplasia)
1736	100	*DKC1*	dyskeratosis congenita 1, dyskerin
1786	100	*DNMT1*	DNA (cytosine-5-)-methyltransferase 1
2146	100	*EZH2*	enhancer of zeste homolog 2 (Drosophila)
2237	100	*FEN1*	flap structure-specific endonuclease 1
2305	100	*FOXM1*	forkhead box M1
2537	100	*IFI6*	interferon, alpha-inducible protein 6
3161	100	*HMMR*	hyaluronan-mediated motility receptor (RHAMM)
4171	100	*MCM2*	minichromosome maintenance complex component 2
4172	100	*MCM3*	minichromosome maintenance complex component 3
4174	100	*MCM5*	minichromosome maintenance complex component 5
4176	100	*MCM7*	minichromosome maintenance complex component 7
4192	100	*MDK*	midkine (neurite growth-promoting factor 2)
4306	100	*NR3C2*	nuclear receptor subfamily 3, group C, member 2
4495	100	*MT1G*	metallothionein 1G
4499	100	*MT1M*	metallothionein 1M
4830	100	*NME1*	non-metastatic cells 1, protein (NM23A) expressed in
5050	100	*PAFAH1B3*	platelet-activating factor acetylhydrolase, isoform Ib, gamma subunit 29kDa
5111	100	*PCNA*	proliferating cell nuclear antigen
5984	100	*RFC4*	replication factor C (activator 1) 4, 37kDa
6241	100	*RRM2*	ribonucleotide reductase M2 polypeptide
6491	100	*STIL*	SCL/TAL1 interrupting locus
6696	100	*SPP1*	secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1)
6790	100	*AURKA*	aurora kinase A
7123	100	*CLEC3B*	C-type lectin domain family 3, member B
7153	100	*TOP2A*	topoisomerase (DNA) II alpha 170kDa
7203	100	*CCT3*	chaperonin containing TCP1, subunit 3 (gamma)
7272	100	*TTK*	TTK protein kinase
8317	100	*CDC7*	cell division cycle 7 homolog (S. cerevisiae)
8318	100	*CDC45L*	CDC45 cell division cycle 45-like (S. cerevisiae)
8480	100	*RAE1*	RAE1 RNA export 1 homolog (S. pombe)

Table A1 (continued)

Entrez ID	Frequency (%)	Gene Symbol	Gene Name
8607	100	*RUVBL1*	RuvB-like 1 (E. coli)
8914	100	*TIMELESS*	timeless homolog (Drosophila)
8985	100	*PLOD3*	procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3
9055	100	*PRC1*	protein regulator of cytokinesis 1
9133	100	*CCNB2*	cyclin B2
9212	100	*AURKB*	aurora kinase B
9232	100	*PTTG1*	pituitary tumor-transforming 1
9319	100	*TRIP13*	thyroid hormone receptor interactor 13
9636	100	*ISG15*	ISG15 ubiquitin-like modifier
9735	100	*KNTC1*	kinetochore associated 1
9768	100	*KIAA0101*	KIAA0101
9787	100	*DLG7*	discs, large homolog 7 (Drosophila)
9833	100	*MELK*	maternal embryonic leucine zipper kinase
9837	100	*GINS1*	GINS complex subunit 1 (Psf1 homolog)
9918	100	*NCAPD2*	non-SMC condensin I complex, subunit D2
9928	100	*KIF14*	kinesin family member 14
10051	100	*SMC4*	structural maintenance of chromosomes 4
10112	100	*KIF20A*	kinesin family member 20A
10351	100	*ABCA8*	ATP-binding cassette, sub-family A (ABC1), member 8
10460	100	*TACC3*	transforming, acidic coiled-coil containing protein 3
10535	100	*RNASEH2A*	ribonuclease H2, subunit A
10635	100	*RAD51AP1*	RAD51 associated protein 1
10643	100	*IGF2BP3*	insulin-like growth factor 2 mRNA binding protein 3
11004	100	*KIF2C*	kinesin family member 2C
11065	100	*UBE2C*	ubiquitin-conjugating enzyme E2C
11130	100	*ZWINT*	ZW10 interactor
22827	100	*PUF60*	poly-U binding splicing factor 60KDa
22974	100	*TPX2*	TPX2, microtubule-associated, homolog (Xenopus laevis)
23636	100	*NUP62*	nucleoporin 62kDa
24137	100	*KIF4A*	kinesin family member 4A
25788	100	*RAD54B*	RAD54 homolog B (S. cerevisiae)
25928	100	*SOSTDC1*	sclerostin domain containing 1
26586	100	*CKAP2*	cytoskeleton associated protein 2
29127	100	*RACGAP1*	Rac GTPase activating protein 1
51203	100	*NUSAP1*	nucleolar and spindle associated protein 1
51659	100	*GINS2*	GINS complex subunit 2 (Psf2 homolog)
54478	100	*FAM64A*	family with sequence similarity 64, member A
54892	100	*NCAPG2*	non-SMC condensin II complex, subunit G2
55165	100	*CEP55*	centrosomal protein 55kDa
55257	100	*C20orf20*	chromosome 20 open reading frame 20
55732	100	*C1orf112*	chromosome 1 open reading frame 112
55872	100	*PBK*	PDZ binding kinase
56992	100	*KIF15*	kinesin family member 15
57405	100	*SPC25*	SPC25, NDC80 kinetochore complex component, homolog (S. cerevisiae)
64151	100	*NCAPG*	non-SMC condensin I complex, subunit G
79005	100	*SCNM1*	sodium channel modifier 1
79019	100	*CENPM*	centromere protein M
79581	100	*GPR172A*	G protein-coupled receptor 172A
79762	100	*C1orf115*	chromosome 1 open reading frame 115
79801	100	*SHCBP1*	SHC SH2-domain binding protein 1
84823	100	*LMNB2*	lamin B2
142	99	*PARP1*	poly (ADP-ribose) polymerase family, member 1
332	99	*BIRC5*	baculoviral IAP repeat-containing 5 (survivin)
443	99	*ASPA*	aspartoacylase (Canavan disease)
471	99	*ATIC*	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase
890	99	*CCNA2*	cyclin A2

Table A1 (continued)

Entrez ID	Frequency (%)	Gene Symbol	Gene Name
790	99	*CAD*	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase
871	99	*SERPINH1*	serpin peptidase inhibitor, clade H (heat shock protein 47), member 1, (collagen binding protein 1)
994	99	*CDC25B*	cell division cycle 25 homolog B (S. pombe)
1062	99	*CENPE*	centromere protein E, 312kDa
2949	99	*GSTM5*	glutathione S-transferase M5
4288	99	*MKI67*	antigen identified by monoclonal antibody Ki-67
4886	99	*NPY1R*	neuropeptide Y receptor Y1
9493	99	*KIF23*	kinesin family member 23
9569	99	*GTF2IRD1*	GTF2I repeat domain containing 1
9631	99	*NUP155*	nucleoporin 155kDa
10894	99	*LYVE1*	lymphatic vessel endothelial hyaluronan receptor 1
11339	99	*OIP5*	Opa interacting protein 5
25840	99	*METTL7A*	methyltransferase like 7A
29107	99	*NXT1*	NTF2-like export factor 1
79888	99	*AYTL2*	acyltransferase like 2
84981	99	*MGC14376*	hypothetical protein MGC14376
898	98	*CCNE1*	cyclin E1
1503	98	*CTPS*	CTP synthase
3627	98	*CXCL10*	chemokine (C-X-C motif) ligand 10
4128	98	*MAOA*	monoamine oxidase A
4318	98	*MMP9*	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)
4940	98	*OAS3*	2'-5'-oligoadenylate synthetase 3, 100kDa
6659	98	*SOX4*	SRY (sex determining region Y)-box 4
6772	98	*STAT1*	signal transducer and activator of transcription 1, 91kDa
9123	98	*SLC16A3*	solute carrier family 16, member 3 (monocarboxylic acid transporter 4)
10212	98	*DDX39*	DEAD (Asp-Glu-Ala-Asp) box polypeptide 39
23492	98	*CBX7*	chromobox homolog 7
55355	98	*URLC9*	up-regulated in lung cancer 9
1434	97	*CSE1L*	CSE1 chromosome segregation 1-like (yeast)
5708	97	*PSMD2*	proteasome (prosome, macropain) 26S subunit, non-ATPase, 2
22880	97	*MORC2*	MORC family CW-type zinc finger 2
23594	97	*ORC6L*	origin recognition complex, subunit 6 like (yeast)
51373	97	*MRPS17*	mitochondrial ribosomal protein S17
1776	96	*DNASE1L3*	deoxyribonuclease I-like 3
4320	96	*MMP11*	matrix metallopeptidase 11 (stromelysin 3)
8662	96	*EIF3B*	eukaryotic translation initiation factor 3, subunit B
10095	96	*ARPC1B*	actin related protein 2/3 complex, subunit 1B, 41kDa
10403	96	*NDC80*	NDC80 homolog, kinetochore complex component (S. cerevisiae)
10797	96	*MTHFD2*	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase
51660	96	*BRP44L*	brain protein 44-like
57122	96	*NUP107*	nucleoporin 107kDa
80308	96	*FLAD1*	FAD1 flavin adenine dinucleotide synthetase homolog (S. cerevisiae)
1469	95	*CST1*	cystatin SN
2335	95	*FN1*	fibronectin 1
3624	95	*INHBA*	inhibin, beta A
5433	95	*POLR2D*	polymerase (RNA) II (DNA directed) polypeptide D
23213	95	*SULF1*	sulfatase 1
79075	95	*DCC1*	defective in sister chromatid cohesion homolog 1 (S. cerevisiae)
103	94	*ADAR*	adenosine deaminase, RNA-specific

Table A1 (continued)

Entrez ID	Frequency (%)	Gene Symbol	Gene Name
2191	94	*FAP*	fibroblast activation protein, alpha
5436	94	*POLR2G*	polymerase (RNA) II (DNA directed) polypeptide G
7004	94	*TEAD4*	TEA domain family member 4
9688	94	*NUP93*	nucleoporin 93kDa
54512	94	*EXOSC4*	exosome component 4
81930	94	*KIF18A*	kinesin family member 18A
3832	93	*KIF11*	kinesin family member 11
5202	93	*PFDN2*	prefoldin subunit 2
6472	93	*SHMT2*	serine hydroxymethyltransferase 2 (mitochondrial)
6944	93	*VPS72*	vacuolar protein sorting 72 homolog (S. cerevisiae)
7371	93	*UCK2*	uridine-cytidine kinase 2
9314	93	*KLF4*	Kruppel-like factor 4 (gut)
10606	93	*PAICS*	phosphoribosylaminoimidazole carboxylase, phosphoribosylaminoimidazole succinocarboxamide synthetase
54517	93	*PUS7*	pseudouridylate synthase 7 homolog (S. cerevisiae)
79980	93	*DSN1*	DSN1, MIND kinetochore complex component, homolog (S. cerevisiae)
4237	92	*MFAP2*	microfibrillar-associated protein 2
6626	92	*SNRPA*	small nuclear ribonucleoprotein polypeptide A
29901	92	*SAC3D1*	SAC3 domain containing 1
762	91	*CA4*	carbonic anhydrase IV
5348	91	*FXYD1*	FXYD domain containing ion transport regulator 1 (phospholemman)
23306	91	*TMEM194*	transmembrane protein 194
54981	91	*C9orf95*	chromosome 9 open reading frame 95
1890	90	*ECGF1*	endothelial cell growth factor 1 (platelet-derived)
4017	90	*LOXL2*	lysyl oxidase-like 2
4173	90	*MCM4*	minichromosome maintenance complex component 4
4521	90	*NUDT1*	nudix (nucleoside diphosphate linked moiety X)-type motif 1
4751	90	*NEK2*	NIMA (never in mitosis gene a)-related kinase 2
7704	90	*ZBTB16*	zinc finger and BTB domain containing 16
9238	90	*TBRG4*	transforming growth factor beta regulator 4
7102	89	*TSPAN7*	tetraspanin 7
7965	89	*JTV1*	JTV1 gene
23397	89	*NCAPH*	non-SMC condensin I complex, subunit H
5690	88	*PSMB2*	proteasome (prosome, macropain) subunit, beta type, 2
11335	88	*CBX3*	chromobox homolog 3 (HP1 gamma homolog, Drosophila)
4794	87	*NFKBIE*	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, epsilon
10248	87	*POP7*	processing of precursor 7, ribonuclease P/MRP subunit (S. cerevisiae)
10376	87	*TUBA1B*	tubulin, alpha 1b
51512	87	*GTSE1*	G-2 and S-phase expressed 1
84722	87	*PSRC1*	proline/serine-rich coiled-coil 1
5993	86	*RFX5*	regulatory factor X, 5 (influences HLA class II expression)
10974	86	*C10orf116*	chromosome 10 open reading frame 116
54908	86	*CCDC99*	coiled-coil domain containing 99
55226	86	*NAT10*	N-acetyltransferase 10
2104	85	*ESRRG*	estrogen-related receptor gamma
2535	85	*FZD2*	frizzled homolog 2 (Drosophila)
4001	85	*LMNB1*	lamin B1
5138	85	*PDE2A*	phosphodiesterase 2A, cGMP-stimulated
6358	85	*CCL14*	chemokine (C-C motif) ligand 14
6628	85	*SNRPB*	small nuclear ribonucleoprotein polypeptides B and B1
7045	85	*TGFBI*	transforming growth factor, beta-induced, 68kDa
7049	85	*TGFBR3*	transforming growth factor, beta receptor III

Table A1 (continued)

Entrez ID	Frequency (%)	Gene Symbol	Gene Name
25903	85	*OLFML2B*	olfactomedin-like 2B
79866	85	*C13orf34*	chromosome 13 open reading frame 34
5427	84	*POLE2*	polymerase (DNA directed), epsilon 2 (p59 subunit)
5531	84	*PPP4C*	protein phosphatase 4 (formerly X), catalytic subunit
5725	84	*PTBP1*	polypyrimidine tract binding protein 1
6185	84	*RPN2*	ribophorin II
8228	84	*PNPLA4*	patatin-like phospholipase domain containing 4
55131	84	*RBM28*	RNA binding motif protein 28
1408	83	*CRY2*	cryptochrome 2 (photolyase-like)
1462	83	*VCAN*	versican
3925	83	*STMN1*	stathmin 1/oncoprotein 18
5033	83	*P4HA1*	procollagen-proline, 2-oxoglutarate 4-dioxygenase (proline 4-hydroxylase), alpha polypeptide I
9134	83	*CCNE2*	cyclin E2
9413	83	*C9orf61*	chromosome 9 open reading frame 61
10926	83	*DBF4*	DBF4 homolog (S. cerevisiae)
51092	83	*SIDT2*	SID1 transmembrane family, member 2
55143	83	*CDCA8*	cell division cycle associated 8
687	82	*KLF9*	Kruppel-like factor 9
79134	82	*TMEM185B*	transmembrane protein 185B
3978	81	*LIG1*	ligase I, DNA, ATP-dependent
64754	81	*SMYD3*	SET and MYND domain containing 3
7852	80	*CXCR4*	chemokine (C-X-C motif) receptor 4
80157	80	*FLJ21511*	hypothetical protein FLJ21511
3014	79	*H2AFX*	H2A histone family, member X
55038	79	*CDCA4*	cell division cycle associated 4
217	78	*ALDH2*	aldehyde dehydrogenase 2 family (mitochondrial)
1675	78	*CFD*	complement factor D (adipsin)
4494	78	*MT1F*	metallothionein 1F
7329	78	*UBE2I*	ubiquitin-conjugating enzyme E2I (UBC9 homolog, yeast)
10537	78	*UBD*	ubiquitin D
25896	78	*INTS7*	integrator complex subunit 7
1875	77	*E2F5*	E2F transcription factor 5, p130-binding
3576	77	*IL8*	interleukin 8
5412	77	*UBL3*	ubiquitin-like 3
7076	77	*TIMP1*	TIMP metallopeptidase inhibitor 1
55723	77	*ASF1B*	ASF1 anti-silencing function 1 homolog B (S. cerevisiae)
64943	77	*NT5DC2*	5'-nucleotidase domain containing 2
1017	76	*CDK2*	cyclin-dependent kinase 2
2819	76	*GPD1*	glycerol-3-phosphate dehydrogenase 1 (soluble)
3248	76	*HPGD*	hydroxyprostaglandin dehydrogenase 15-(NAD)
9452	76	*ITM2A*	integral membrane protein 2A
288	75	*ANK3*	ankyrin 3, node of Ranvier (ankyrin G)
813	75	*CALU*	calumenin
2189	75	*FANCG*	Fanconi anemia, complementation group G
5328	75	*PLAU*	plasminogen activator, urokinase
5351	75	*PLOD1*	procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 1
5982	75	*RFC2*	replication factor C (activator 1) 2, 40kDa
6338	75	*SCNN1B*	sodium channel, nonvoltage-gated 1, beta (Liddle syndrome)
7433	75	*VIPR1*	vasoactive intestinal peptide receptor 1
9079	75	*LDB2*	LIM domain binding 2
865	74	*CBFB*	core-binding factor, beta subunit
2960	74	*GTF2E1*	general transcription factor IIE, polypeptide 1, alpha 56kDa
4501	74	*MT1X*	metallothionein 1X
6895	74	*TARBP2*	TAR (HIV-1) RNA binding protein 2
27286	74	*SRPX2*	sushi-repeat-containing protein, X-linked 2
56920	74	*SEMA3G*	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3G

Table A1 (continued)

Entrez ID	Frequency (%)	Gene Symbol	Gene Name
6414	73	*SEPP1*	selenoprotein P, plasma, 1
2690	72	*GHR*	growth hormone receptor
3068	72	*HDGF*	hepatoma-derived growth factor (high-mobility group protein 1-like)
5987	72	*TRIM27*	tripartite motif-containing 27
9603	72	*NFE2L3*	nuclear factor (erythroid-derived 2)-like 3
29980	72	*DONSON*	downstream neighbor of SON
79833	72	*GEMIN6*	gem (nuclear organelle) associated protein 6
7754	71	*ZNF204*	zinc finger protein 204
633	70	*BGN*	biglycan
3131	70	*HLF*	hepatic leukemia factor
5437	70	*POLR2H*	polymerase (RNA) II (DNA directed) polypeptide H
9168	70	*TMSB10*	thymosin, beta 10
10964	70	*IFI44L*	interferon-induced protein 44-like
63924	70	*CIDEC*	cell death-inducing DFFA-like effector c
93594	70	*WDR67*	WD repeat domain 67

Vita

Noor Dawany

Education
Drexel University, Philadelphia, PA
PhD in Biomedical Engineering, June 2010

Spalding University, Louisville, KY
BSc in Biology and Mathematics, November 2004

Research
Computational
Bioinformatics Research Assistant, 2007-2010
Center for Integrated Bioinformatics
Drexel University, Philadelphia, PA

Laboratory
Research Assistant, 2004-2005
Spalding University, Louisville, KY

Publications
Dawany N, Tozeren A. Asymmetric integration of microarray data outperforms meta-analysis approach in multiple cancer types. *BMC Bioinformatics* (in review).

Dawany N, Dampier W, Tozeren A. Large-scale integration of microarray data reveals genes and pathways common to multiple cancer types. *International Journal of Cancer* (in review).

Sarmady M*, Dawany N*, Tozeren A. Connectivity map for viral crosstalk with host iron binding proteins (in preparation).
* contributed equally to the article

Teaching Experience
Teaching Assistant, 2008-2010
Drexel University, Philadelphia, PA

- Genome Information Engineering
- Quantitative Systems Biology
- Engineering Principles Living Systems (I&II)
- Foundations of Biomedical Engineering