NTNU – Trondheim
Norwegian University of
Science and Technology

Conservation and Evolution of Pairwise Gene Co-Expression Correlations in HIV and TB Co-infected Individuals

A Network-based Analysis of Whole-Blood RNA samples

Martin Borud

Chemical Engineering and Biotechnology
Submission date: December 2014
Supervisor: Eivind Almaas, IBT

Norwegian University of Science and Technology
Department of Biotechnology

Abstract

Network analysis of gene co-expression data has been shown to be a strong tool to elucidate biological information from large datasets. In this analysis the goal has been to see if the method developed by Voigt, Nowick & Almaas (2015), based on pairwise gene co-expression correlation, can be used to extract biologically significant information about co-infection of HIV and tuberculosis. Their method is based on the idea that pairwise correlation between genes can be grouped in categories of conserved, divergent and specific correlation, based on their expression in different tissues/samples.

Data has been gathered from the study done by Kaforou et al. (2013), which includes whole-blood RNA samples gathered from HIV positive and HIV negative patients with either active or latent tuberculosis infection. These samples were grouped accordingly and analyzed, generating networks consisting of conserved, divergent and specific correlations. Networks consisted of correlations between either HIV negative or HIV positive samples co-infected with either active or latent TB, resulting in two different networks. These were compared to each other, and the network consisting of HIV positive samples was further analyzed to determine significant nclustering and gene hubs, in the hopes of extracting biological information useful in relation to HIV and TB co-infection.

Clusters of conserved correlation between gene pairs was found to be preserved between the two different networks, consisting of genes that were found to be involved in the general upkeep of cells. Hubs were found to be located in areas of high degree of specific and divergent pairwise gene co-expression correlations, implying involvement in immune response pathways that differ between the case of latent infection of TB and active TB infection in HIV positive patients.

Results from the study performed by Kaforou et al. (2013) were used to determine whether or not the findings from this analysis were consistent with previous findings. TB "Fingerprint" genes are genes whose expression correlated with active TB infection, and could help distinguish between active TB infection from other types of infection, were found to be present to some degree in the network, located in areas of mainly specific and divergent gene pair correlations. This is consistent with what is expected when performing pairwise gene co-expression correlation analysis.

These findings indicate that the method described by Voigt et al. (2015) shows great promise in extraction of biological information from gene co-expression data, but more analysis is required to determine more specific genetic implications of the data considered in this study.

Sammendrag

Nettverksanalyse av gen-koekspresjonsdata har vist seg å være et kraftig verktøy for å hente ut biologisk informasjon fra store datasett. I denne analysen har målet vært å se om metoden utviklet av Voigt, Nowick & Almaas (2015) kan benyttes for å hente ut biologisk interessant informasjon om koinfeksjon av HIV og tuberkulose. Metoden er basert på ideen om at parvis korrelsjon mellom gener kan bli gruppert i henhold til hvorvidt de er konserverte, divergente eller spesifikke korrelajsoner, ut i fra deres ekspresjon i forskjellige vev/prøver.

Data har blitt innhentet fra studien gjennomført av Kaforou et al. (2013), som inkluderer "whole-blood" RNA-prøver samlet fra HIV-positive og HIV-negative pasienter som i tillegg er smittet med enten aktiv eller latent tuberkulose. Disse prøvene ble gruppert og analysert slik at to nettverk ble produsert, bestående av konserverte, divergente og spesifikke korrelasjoner Dette resulterte i to forskjellige nettverk som ble sammenlignet med hverandre, og nettverket av HIV-positive prøver ble videre analysert for å bestemme signifikante "klustere"/klynger og tilstedeværelsen av "hubs". Dette ble gjort i håp om å kunne hente ut biologisk informasjon som kan benyttes i forbindelse med koinfeksjon av HIV og tuberkulose.

Klynger bestående av konserverte korrelasjoner var tilstedeværende i begge nettverk. Disse genene viste seg å være delaktige i generelle genfunksjoner. Hubs ble lokalisert i områder med høy andel spesifikke og divergente korrelasjoner, og det ble funnet indikasjoner påat disse områdene er involvert i immunrespons som er differensiert mellom prøver med latent tuberkulose og prøver med aktiv tuberkulose.

Resultater fra studien gjennomført av Kaforou et al. (2013) ble benyttet for å vise hvorvidt resultatene fra denne analysen stemte overens med tidligere funn. "Fingeravtrykkgener" er gener hvis ekspresjon viser samvarians med aktiv tuberkuloseinfeksjon. Disse kan bidra til åskille prøver med aktiv tuberkulose fra prøver med andre infeksjoner. Omtrent halvparten av fingeravtrykkgenene ble funnet i nettverket bestående av HIV-positive prøver, lokalisert i områder med høy andel spesifikke og divergente korrelasjoner. Dette er konsistent med hva som er forventet.

Resultatene indikerer at metoden beskrevet av Voigt et al. (2015) viser stort potensiale for åhente ut biologisk informasjon fra gen-koekspresjonsdata, men at mer analyse er nødvendig for åbestemme mer spesifikke genetiske implikasjoner i dataene benyttet i denne analysen.

Preface

This thesis is submitted to the Department of Biotechnology of the Norwegian University of Science and Technology (NTNU) in partial fulfillment of the requirements for the Master of Science in Chemical Engineering and Biotechnology. The work was carried out from August to December 2014, under the supervision of Professor Eivind Almaas.

I would like to thank Eivind for his excellent guidance and support, and for helping me understand the quirks of networks, biological systems and computer programming.

I would also like to thank Rune Håkonsen, for his motivational talks, for listening to me rant about scripts not doing what I want them to do, for holding out with me in the last hectic weeks of this work, and last but not least, for delivering a bottle of champagne to me at my work station for when the thesis was finally submitted.

Martin Borud
Trondheim, December 2014

Contents

List of Figures

List of Tables

1. Background

This master thesis is based upon the network analysis method developed by Voigt, Nowick & Almaas (2015) The foundation of this method is that gene sequences that diverge between species, whilst conserved within them, are believed to contribute to phenotypic differences between species (Hudson et al. 1987). It is expected that not only differences in gene sequences, but also in gene expression contribute to such phenotypic differences (Nowick et al. 2009).

This reasoning has been extended to gene co-expression networks, and the aim was to investigate the role of specific genetic interactions by studying correlation between gene pairs and whether this correlation is conserved or differentiates under alternative environmental perturbations. Network analysis has shown great promise in elucidating biological meaning from gene expression sets, protein-protein interaction networks, but correlating network information between species or perturbations is still a work in progress. Examples of environmental perturbations can mean either cell cultures from two (or more) different species or a single cell culture expressed under different regiments.

The premise of this master thesis is an example of the latter; human blood samples with or without HIV infection as well as either active or latent tuberculosis infection. This kind of system has been studied by Kaforou et al. (2013). Their hypothesis was that a unique host blood RNA transcriptional signature could be used to distinguish tuberculosis from other diseases in HIV-infected and -uninfected patients. Co-infection of HIV and TB is a highly significant cause of death, especially in areas of inadequate health systems and poverty. Fast, cheap and reliable detection of infection can help save many lives. Through analysis of patients infected with these diseases, Kaforou et al. (2013) were able to distinguish a "fingerprint" set of expressed RNA specific to patients co-infected with HIV and TB.

Is it possible to detect differences in the gene co-expression networks of patients infected with HIV and/or TB? By introducing the pair-wise gene co-expression correlation analysis of Voigt et al. (2015) to this system, this study is assessing the possibility of detecting gene modules or gene hubs in humans specific to infection of either active or latent tuberculosis and co-infection of HIV. Determining the feasibility of extracting biological sense from networks created using pair-wise correlation, and whether this can be used in unison with other findings regarding co-infection to more easily be able to detect and treat these diseases.

2. Theory

2.1 Network theory

Newman (2003) describes a network as a set of items, known as vertices or *nodes*, that are connected to each other through *edges* (Figure 2.1). The concepts stem from the mathematical area of graph theory. Network theory can be used to describe many structures known to us, such as social networks where nodes represent individuals and edges the connection between them, or the World Wide Web, where nodes represent servers or computers and edges represent pathways of information between them. The edges can take on different kinds of connections between nodes. In our social network the edges between individuals can represent friendliness, animosity or sexual interactions as just some examples. Network theory can be used on a multitude of different data sets capable of network representation, from gene expression (Ruan et al. 2010), to tennis players (Radicchi 2011), to the connection between social interactions and influenza outbreaks (Gardy et al. 2011).

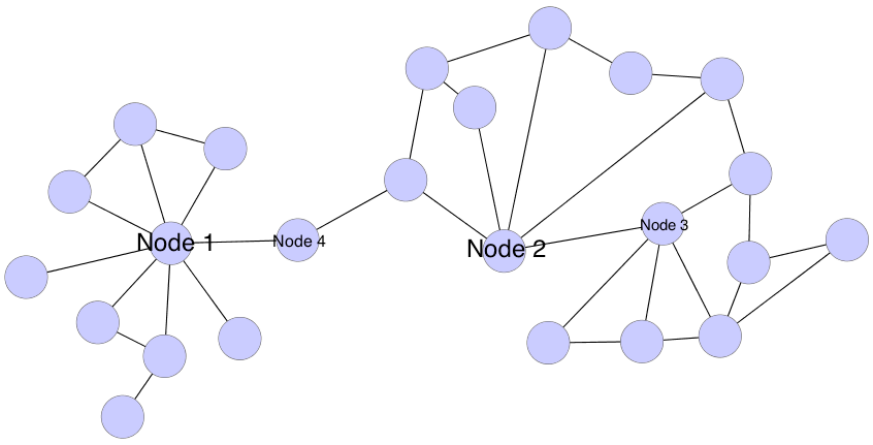


Figure 2.1: Conceptual example of a network. The circles are nodes and the black lines are edges, signifying which nodes are connected to each other.

2.1.1 Adjacency matrix and degree

As defined by Horvath (2011), an *adjacency matrix* $A = (A_{ij})$ can be used to describe the pairwise relationship between a set of n nodes. Each component A_{ij} quantifies the connection strength from node i to node j. Two types of networks give different values for the adjacency matrix. For an *unweighted* network, the value of component A_{ij} equals either 1 or 0. Either the link between the nodes exists or it does not. For a *weighted* network, the value A_{ij} is a real number between 0 and 1. Edges between nodes can also be either *directed* or *undirected*. For an undirected network, there is no difference in going from node i to node j compared to going from node j to i ($A_{ij} = A_{ji}$), while for a directed network, the adjacency matrix may not be symmetrical.

The nodes that are directly connected to node i, are known as its nearest neighbors in the network, and the number of nearest neighbors of node i is known as connectivy or *degree* (k_i). If we treat the network in Figure 2.1 as unweighted, node 1 has seven nearest neighbors, and $k_1 = 7$, and node 2, which has five nearest neighbors, has degree $k_2 = 5$. More generally, Horvath (2011)defines the degree of node i is as:

$$k_i = \sum_{j \neq i} A_{ij} \tag{2.1}$$

For an unweighted network, k_i equals the number of nearest neighbors, while in a weighted network, k_i is the sum of connection weights between node i and the other nodes. For the remainder of this text, focus will be held on unweighted networks unless otherwise stated.

2.1.2 Hubs and modules

One of the prominent feature of many networks is the occurrence of *hubs* and *modules*. Hubs are defined as nodes with a high degree k_i, meaning that they have connections to many other nodes. These nodes have been shown in several network studies to be important actors in the network structure (Albert et al. 2000, Jeong et al. 2001, Albert & Barabási 2002), though this does not always have to be the case. In Figure 2.1 the three nodes with the highest degree are node 1, node 2 and node 3, with a degree of 7, 5 and 5 respectively.

An example of hub significance is protein-protein interaction networks, where deletion of hub genes from an organism are more likely to be lethal than deletion of non-hub genes (He & Zhang 2006). Alternatively, in a social network, if connections between nodes are based on friendliness, hubs can be seen as the popular individuals. In such a network, the hubs might even be highly connected to each other, forming a clique. In network context highly connected groups of nodes are called modules or clusters.

The network shown in Figure 2.1 we can identify two separate clusters, one including node 1 and the other including node 2 and 3. In gene co-expression networks, gene clusters have been shown to express products linked to the same biological pathways or protein complexes (Segal et al. 2003, Zhang & Horvath 2005). In this context, for the network in Figure 2.1, node 4 could be referred to as a

"bridge" node that link the two clusters together. In gene co-expression it can be seen as a kind of *gatekeeper gene*, responsible for co-regulation of the production of two correlated protein complexes by transmitting information between the them.

There are different methods that can be used to calculate and identify clustering in a network. Horvath (2011) mentions partitioning-around-medoids (PAM) clustering and hierarchal clustering as two methods often used in network applications. For this analysis, the MCODE plugin was used, and the method, which is based on k-coring, is further detailed in Section 3.4. The general approach is to find sets of nodes with robust and strong connections to each other, that stand out from the network in general. One central parameter in regards to clustering is the *clustering coefficient*, which is a density measure of local connections. Dong & Horvath (2007) defines the clustering coefficient as follows:

$$ClusterCoef_i = \frac{\text{number of actual connections between neighbors of i}}{\text{number of possible connections between neighbors of i}}$$

This can be evaluated in the context of an adjacency matrix, resulting in the following expression:

$$ClusterCoef_i = \frac{2\Delta_i}{k_i(k_i-1)} = \frac{1}{k_i(k_i-1)} \sum_{j=,l=1}^{N} a_{ij}a_{jl}a_{li}, \tag{2.2}$$

where Δ_i is the number of triangles node i is a part of. Following this logic, $ClusterCoef_i$ equals 1 if all neighbors of i are connected to each other.

2.1.3 Centrality

The term centrality is used when determining the relative importance of nodes or edges in a network. Degree is one such centrality measure. As mentioned, a node with a high degree could correlate with high significance for the architecture of the network, but there are limitations. Chen et al. (2012) propose that degree as a centrality measure can fail to identify important nodes in some cases.

In Figure 2.1 node 1 has a higher degree than node 2. Still, node 2 may be more influential on the other nodes in the network because of its connectivity beyond its first neighbors. The biological example by Chen et al. (2012) argues that an infection that starts in node 2 will have faster and further-reaching effects than an infection starting in node 1as a result of its nearest neighbors having a higher degree than node 1.

Several centrality measure parameters have been proposed to account for these global effects in a network, including betweenness centrality and closeness centrality, two *geodesic-path-based ranking measures*. Geodesic path is a distance measurement where one "walks" from one node via the least amount of edges necessary to another node. This distance is also known in network theory as the shortest path between nodes.

The betweenness centrality of a node v is defined as the fraction of shortest paths between node pairs that pass through v. Chen et al. (2012) define betweenness centrality $C_B(v)$ for a network with n nodes and m edges as:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{2.3}$$

where σ_{st} is the total number of shortest paths between s and t and $\sigma_{st}(v)$ is the number of shortest paths through v. The rationale behind betweenness centrality is that nodes with a high fraction of shortest paths through it will have significant effect on the network as a whole since information from other nodes will readily pass through it. The "bridge" nodes from Figure 2.1 are examples of nodes with high betweenness centrality.

Closeness centrality $C_C(v)$ takes into account the distance from node v to all other nodes, and focuses on how long it will take to spread information from node v to all other nodes. It is defined as the reciprocal sum of geodesic distances to all other nodes of V:

$$C_C(v) = \frac{1}{\sum_{t \neq v} d_g(v,t)}, \tag{2.4}$$

where $d_g(v,t)$ is the geodesic distance between v and t.

2.2 Gene co-expression networks

The thought behind gene co-expression studies is that genes that are expressed in concert with each other (i.e. one gene's expression displays similar pattern to another gene's expression) is related to phenotypic and functional consequences (Oldham et al. 2006, Nowick et al. 2009). The advantage of studying a network generated from pairwise gene co-expression data is that it is possible to look at the system-level functionality of genes. Co-expression analysis uses correlation data or any other "distance" measure to find similarities between gene expression profiles.

DNA microarray experiments have shown that clusters of genes with correlated expression patterns have protein products that participate in the same pathways (Hughes et al. 2000, Segal et al. 2003). By studying gene co-expression it has been possible to see how changes in gene expression result in phenotypic differences between species. Ebersberger et al. (2002) find that chimpanzees and humans have a high extent of gene sequence homology, and points to differences in the expression of homologous genes as a reason for our different physiologies.

Earlier work has looked at differences in expression between species or tissues, but with difficulties in discerning between functionally significant and insignificant expression changes (Khaitovich et al. 2004). The usage of network analysis on a genome-wide scale have come a long way in elucidating how clusters of genes are conserved or differentiate between species, further establishing their phenotypic effects. Stuart et al. (2003) found that there are both animal-specific components and conserved relationships between newly evolved and ancient modules when comparing the genomes of *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegant* and *Saccharomyces cerevisae*.

Their network analysis also showed that some genes with unknown biological function linked closely to modules of genes with known functions. These genes were experimentally shown to be functionally essential for the metabolic pathways of these modules. This shows that network analysis of gene co-expression networks also opens up the possibility to map the biological significance of genes with previously unknown functions.

2.3 Constructing gene co-expression networks

Construction of gene co-expression networks is conceptually straight-forward. Nodes represent genes, and if two genes are significantly co-expressed across appropriately chosen tissue samples, they are connected. To get to a finalized network is not as straight-forward. It is necessary to make decisions on how to analyze the raw data. There are many statistical aspects to take into account, as pointed out by Zhang & Horvath (2005). What kind of data to analyze, what kind of correlation coefficient to use, and as Zhang and Horvath focus on in their article, what kind of significance weighted networks can have as apposed to unwheigted networks. In this section, the choices done by Voigt, Nowick & Almaas (2015) when establishing their method are detailed.

2.3.1 Determining correlation values

To construct differential gene co-expression networks, the correlation of each gene pair is found using the Spearman's rank correlation (ρ), defined as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.5}$$

Here, for a sample size n, the n raw scores X_i and Y_i are converted to ranks x_i and y_i and $d_i = x_i - y_i$. A value of $+1$ or -1 for ρ signifies a perfect Spearman correlation, where 1 would indicate perfect positive correlation (both parameters increase or decrease in the same direction), while a spearman correlation of -1 indicates that the parameters are inversely correlated (one parameter increases as the other decreases). Values close to 0 signifies little or no correlation between the tested values.

Spearman correlation is only one of several possible methods of defining the relationship between gene pairs. Horvath (2011) discusses the possible other correlation coefficient methods to note; Pearson- and biweight mid-correlation. In general, the difference between Spearman and Pearson correlation is that Spearman measures the rank order of data points, not taking into account exactly where they are on an axis, while Pearson measures how well the two parameters fit a linear relationship. Pearson correlation is more sensitive to outliers (values that deviate from the general trend), while Spearman is overly conservative in many applications. The biweight mid-correlation is more complex than Spearman's rank and Pearson, but combines the advantage of Person-correlation's high power and the Spearman-correlations high robustness.

To analyze changes in correlation, the group defined three possible scenarios for gene pairs present in two tissues/samples:

- Conserved correlations: Strong and similar in both tissues
- Divergent correlations: Strong in both tissues, but dissimilar
- Specific correlations: Strong in one tissue only

The different regions corresponding to the three scenarios are shown in Figure 2.2. Conserved correlations in a gene co-expression data set would be gene pairs whose pairwise expression in an organism is unchanged between environmental perturbations, or highly evolutionary stable gene modules between species.

An example of specific correlations are gene pairs whose role is only evident in one type of tissue in an organism, such as brain-specific genes. When measuring gene expression of these genes in other tissues, one would expect little or no pairwise gene correlation in contrast to expression rate in brain tissue. In general, gene pairs that are strongly correlated in one tissue or species while showing no significant correlation in the other.

Divergent correlations are predicted to be present in gene pairs who have strong absolute correlations in both tissues/perturbations, but these are oppositely correlated (negatively correlated in one and positively correlated in the other).

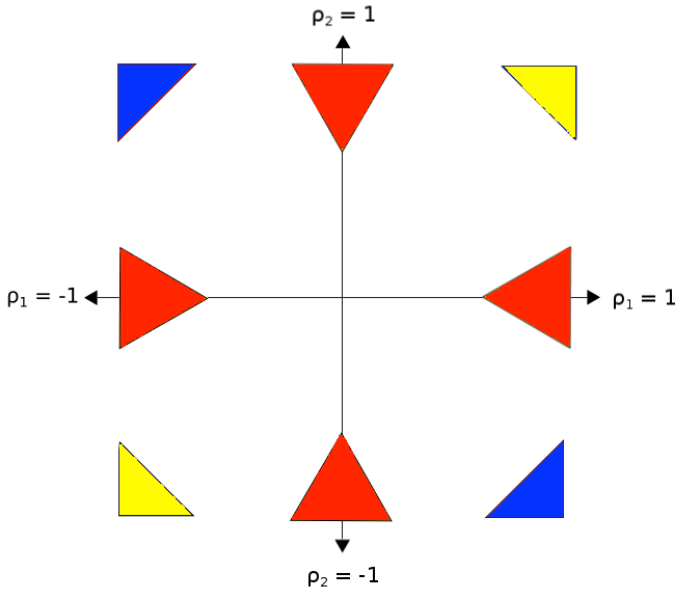


Figure 2.2: Map of possible scenarios of correlation combination. Gene pairs with co-expression correlation values such that they fall within the areas of the red triangles are considered specific to one tissue/sample, while yellow triangles depict strongly conserved pairs, and blue depict strongly divergent pairs.

2.3.2 Correlation metrics

To determine whether a gene pair is specific for one tissue/species or divergent or conserved between them, three metrics have been introduced that are non-overlapping and can be generalized to any higher-order comparison:

Conservation (C):

$$C = \frac{|\rho_1 + \rho_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{2.6}$$

Divergence (D):

$$D = \frac{|\rho_1| + |\rho_2| - |\rho_1 + \rho_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{2.7}$$

Specificity (S):

$$S = \frac{\|\rho_1| - |\rho_2\|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{2.8}$$

Here, ρ_1 refers to the correlation value for a gene pair in one tissue or sample and ρ_2 to the correlation value for the same gene pair in another, σ_i is the variation for each of the gene pairs in a tissue/sample. Significant scores are evaluated by introducing a threshold value for each of the metrics, by determining a p-value. Details on determining threshold values for this analysis is given in Section 3.2.

For the three metrics, C, S and D, the variation σ_i is introduced. The method is based upon correlation between two disparate groups, for example two gene sequences taken from two different species. To effectively evaluate the correlation between these two groups, and the significance of their overlap, we need to calculate the variation in each of the sample groups. In the example of gene sequences evaluated between species, a gene sequence with a lot of variation in expression in one species is likely loosely regulated, meaning it is of less significance to the overall system. σ_i is introduced to score correlation values with a high amount of variation in a set of samples correctly. This can be done by subsampling in the different test groups, and calculating its standard deviation of the mean.

2.4 Gene ontology

As previously stated, gene sequences specifying core biological process are shared between many species. This has led to the advent of *gene ontology* studies and identification (Ashburner et al. 2000). A consortium of different agencies and research groups work together to categorize genes and gene products into functional groupings based on their roles. Each gene is given a set of gene ontology (GO) ID-markings, making it possible to check the connection between sets of genes. There are three main categories; molecular functions, cellular components and biological processes. Genes and gene products are also linked to other databases, such as GenBank and Ensembl for easy identification of different gene IDs.

Ashburner et al. (2000) mentions 'cell growth and maintenance' and 'signal transduction' as a couple of the broad biological process terms under which genes

and gene products can be categorized. Molecular functions can incorporate 'enzyme', 'transporter' or 'ligand' among others, while cellular components refers to where in the cell different genes are active, such as 'ribosome' and 'protesome'.

Gene ontology databases, such as the Database for Annotation, Visualization and Integrated Discovery (DAVID) bioinformatics database (Da Wei Huang & Lempicki 2008, Huang et al. 2009), can be used to check whether clusters in co-expression networks are part of a specialized biological process, or if there is a common molecular output from these genes, possibly correlating cluster affiliation and protein complexes or pathways. The DAVID analysis clusters genes that share gene ontologies together and rank them according to an enrichment score based on number of genes from the supplied list which are a part of the given gene ontology.

2.5 HIV and tuberculosis

Tuberculosis (TB) and human immunodeficiency virus/acquired immune deficiency syndrome (HIV/AIDS) ranks as some of the deadliest infectious diseases on a global scale. Estimates from the World Health Organization (WHO) show approximately 9 million new cases of TB in 2013 and 1.5 million deaths, 360 000 of which were HIV positive (WHO Global Tuberculosis Report 2014). In 2010, 14 million individuals were estimated to be dually infected with both HIV and TB (Getahun et al. 2010).

2.5.1 HIV

HIV is an infectious retrovirus responsible for causing AIDS, a condition recognized by progressive failure of the immune system in humans (Weiss 1993). HIV is a systemic infection, meaning that it is found throughout the body of an infected individual. It transmits between humans through body fluids such as blood and semen. The main routes of infection are sexual transmission, parenteral transmission (sharing needles, blood transfusion) and transmission from mother to infant (De Cock et al. 2000). It has been found that HIV susceptibility and transmission increases when co-infected with other sexually transmitted diseases (STDs), such as gonorrhea and chlamydia (Cameron et al. 1989, Mayer & Venkatesh 2011).

HIV infects cells vital to the immune response system in humans such as T helper cells ($CD4^+$ T cells), macrophages and dendritic cells (Cunningham et al. 2010), leading to low levels of $CD4^+$ T cells. When this cell count decline below a critical threshold, cell-mediated immunity is lost, making the body susceptible to opportunistic infections.

One of the trademarks of progressive HIV infection is chronic activation of the immune system. Multiple types of cells involved in immune response show increased turnover and frequency (Brenchley et al. 2006). It is believed that this may result from HIV interfering with the functional organization of the immune system, further allowing viral evolution and spurring on the emergence of AIDS (Grossman et al. 2006).

2.5.2 *Mycobacterium tuberculosis*

TB is caused by infection of the bacteria *Mycobacterium tuberculosis*, and most commonly affect the lungs (Pawlowski et al. 2012). It spreads via air in droplets from infected individuals caused by coughing or sneezing. Symptoms of active TB are coughing with sputum or sometimes blood, chest pains, weakness, weightless, fever and night sweats (WHO Global Tuberculosis Report 2014). Otherwise healthy individuals do not exhibit active TB, as the immune system keeps the infection in check. A TB-infected individual who does not exhibit symptoms is said to have a latent TB infection, or LTBI. TB is held in check by the activation of $CD4^+$ and $CD8^+$ T helper cells, as well as other immune response cell types(Pawlowski et al. 2012). Decline in immune system response may reactivate LTBI to an active state, propagating the disease.

2.5.3 HIV and tuberculosis co-infection

TB and HIV co-infection poses a novel pathogenic scenario , leading to challenges both in regards of diagnosis and therapy (Pawlowski et al. 2012). Management and epidemiology becomes a great deal more complex for both diseases when co-infected in comparison to a mono-infection of either TB or HIV. This severity is related to the fact that both pathogens potentiate one another leading to an accelerated deterioration of the immune system (Pawlowski et al. 2012). LTBI has a 20-fold increase of reactivation to active TB in the presence of HIV infection.

There are treatments available for dealing with HIV/TB co-infection, but there are challenges related to the quality of the health system in many countries where this problem is prevalent. Better and easier ways of early diagnosis could help save the lives of many infected patients (Kaforou et al. 2013).

2.6 Microarray analysis

A microarray is an analytical device based on a solid substrate (glass slide or silicon thin-film cell) containing labeled components from cells, tissues and other biological sources placed in indents or wells on the plate (Schena 2003). These are used to assay large amounts of biological material using high-throughput screening. Types of microarrays include DNA microarrays (cDNA, oligonucleotide microarrays etc.), protein microarrays, peptide microarrays, antibody arrays and others. Prominent manufacturers of microarray technology are Affymetrix, Agilent and Illumina.

In gene expression analysis the sample that is to be analyzed is amplified and hybridized against labeled probes, creating a 2D-array with different probes expressed in different regions. The arrays are scanned by a laser that excites the labeled probes, and registers their expression value as emission levels. This results in values of expression for each of the labeled probes related to each other. The data is usually normalized and have background noise reduced as a part of the procedure (Schena 2003).

2.6.1 Illumina bead array

Illumina bead arrays, as used by Kaforou et al. (2013) in their initial research, is a type of microarray analysis where each array is assembled on an optical imaging fiber bundle consisting of about 50000 individual fibers fused together (Oliphant et al. 2002). These are etched to produce the wells in which the beads that are central to this method is placed. The beads are produced by covalently attaching oligonucleotides to silica beads. The different beads are pooled together and mounted randomly to fibers, and their place is determined post-assembly. The resulting matrix is then processed as usual for microarray analysis. The size of the beads ($3\mu m$ in diameter) makes it possible to increase throughput by increasing matrix size compared to regular microarray chips.

2.6.2 Whole-blood RNA samples

Whole-blood RNA samples contain mRNA collected from peripheral blood samples (Kaforou et al. 2013), meaning minimal preprocessing of the blood before RNA is extracted. Both HIV and TB are systemic infections that affect the entire organism, resulting in activated immune system and stress responses throughout organs/tissues. Whole-blood samples will effectively become a "snapshot" of the state of the infected organism, since protein-producing mRNA strands involved in these reactions will be carried throughout the circulatory system. This makes this kind of sampling a strong candidate to determine the ramifications and effects of a pathogenic infection on a system.

3. Method

3.1 Data collection

The data used in this analysis consists of the findings in the 2013 study by Kaforou et al.: *Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study.* 536 patients from Malawi and South Africa with either active TB infection, latent TB infection (LTBI) or other diseases (OD) were classified as either HIV positive or HIV negative. Whole blood RNA was extracted from blood samples and labeled cRNA analysed on Human HT-12 v.4 expression Beadarrays from Illumina. The datasets were accessed from the Gene Expression Omnibus (GEO) database, accession set GSE37250. The raw data has been adjusted so that background values have been subtracted and quantile normalized prior to being obtained from GEO.

The dataset consisted of microarray data using whole blood RNA samples from 536 patients, with 47323 probes per sample. Samples were grouped as shown in Table 3.1. For the purpose of this analysis, the data from the OD group was discarded, as the focus of the analysis was on changes relating to co-infection of HIV and TB. Probes had been annotated according to the National Center for Biotechnology Information's (NCBI) Reference Sequence collection (RefSeq) for gene identification. Using this as a filter, a perl script was written to select only genes with known protein products for further analysis, annotated in the dataset as NM genes. This was done to reduce the amount of data before processing, but without losing significant and central gene probes. For genes with multiple probes, their mean value was calculated and used in further computation. To group probes in terms of gene target, they were matched with their official gene names, using a perl script and HGNC (HUGO Gene Nomenclature Committee) nomenclature provided from the metadata in the GEO dataset.

Protein-producing genes were selected on the basis that the main characteristic of this analysis is to find differences in expression that can account for infection of HIV and/or TB. These infections interact in different ways with the immune system response of humans (Section 2.5), resulting in changes in cell production. The hypothesis is that this will reflect on the protein-producing genes in that these include DNA-regulatory proteins (Mitchell & Tjian 1989) and proteins involved in cell proliferation and maintenance (Ensoli et al. 1990, Shaulian & Karin 2001).

Table 3.1: Number of samples collected from dataset GSE37250 of the GEO database. The numbers represent patients with that are either HIV positive and negative, and whether they are co-infected with active tuberculosis, latent tuberculosis or other diseases.

	Active TB	**Latent TB**	**Other diseases**
HIV positive	97	84	92
HIV negative	97	83	83

3.2 Computer analysis

After selection of the protein-producing gene probes and calculation of mean values, the resulting samples consisted of 19396 gene probe values. These were further subsampled into 40 sets of 10000 randomized probes for each of the four categories; HIV positive and active TB, HIV positive and LTBI, HIV negative and active TB and lastly HIV negative and LTBI. This was done to be able compile the large amount of data present in the datasets. Testing showed that needed computer memory for running the code written to calculate correlation values for all 19396 gene probes at the same time was found to be 180 GB of RAM. The available computer equipment had a limit of 130 GB available RAM, making subsampling of the data set necessary. By dividing it into partitions of 10000 data points per sample, the amount of needed memory no longer exceeded 130 GB.

40 subsamplings with 10000 randomized gene expression values per sample were done for each set to be sure that each gene pair would have their pairwise correlation value calculated at least once. The reasoning behind this is straightforward. The probability of one gene occurring in a subsample of 10000 genes: $P(Gene\ occurring) = \frac{10000}{19396} = 0,52$. This implies that there is a one in four chance two genes occur in the same subsample. By doing 40 sets of subsamples, each gene pair should appear approximately ten times. A perl script was written to choose 10000 random gene probes and group them together. Analysis of the resulting subsamples showed that all possible gene pairings were present.

Spearman rank correlation value was calculated to obtain pairwise co-expression scores using Equation 2.5. Variability in co-expression was found by dividing the samples into groups of ten, so that the variance σ could be calculated for each gene pair. This subsampling was done so that no more than one individual sample was shared between subsamples, resulting in enough subsamples to calculate variance while still ensuring independence between subsamples. A C++ script had been made by Voigt et al. (2015) for this purpose. Some changes were done to this program to accommodate for differences in input data.

Four possible combinations of of the sample sets were thought to carry meaningful biological significance, each consisting of two sets of correlation data differing in either HIV status or TB status.

- HIV positive and active TB - HIV negative and active TB
- HIV positive and active TB - HIV positive and LTBI
- HIV negative and LTBI - HIV negative and active TB

- HIV negative and LTBI - HIV positive and LTBI

The reason for this choice was to connect findings in the networks to changes in infection status. By letting one parameter differ between the networks, it would be easier to point to possible explanations, as is common practice in research. One of these sets were focused on, the HIV positive pool of TBa active and LTBI samples. These samples were thought to show differences between active and latent TB infection in HIV patients, and could possibly point to interesting gene modules or hubs that are specific for either of the cases. As pointed out in Section 2.5.3, the novel aspects of HIV and TB-coinfection needs a different approach than monoinfection. The pool of HIV negative TB and LTBI-samples were analyzed to contrast findings in the HIV positive pool.

After pairwise gene correlation had been calculated for each gene pair, a perl script was written and used to select all of the unique gene pairs and their correlation and variation values. Equations 2.6-2.8 were used to evaluate each pairwise correlation metric. Another C++ script supplied by Voigt et al. (2015) was used for this purpose. Some changes had to be made for the data to be correctly handled, and to obtain the wanted metrics.

Using a perl script, a hard cutoff threshold for C, S and D was calculated using a method based on Bonferroni correction. For each of the metrics, a sample of 20000 scores were randomly chosen. The highest metric score in this sampling was select and stored. This was repeated 50000 times, and the mean value of the highest scores was calculated. Gene pairs with metric scores over this threshold. By chosing 20000 random scores, the p-value of this method was $5 * 10^{-5}$. Several iterations using different p-values were done to assess how many significant gene pairings to include, and how stable the resulting cutoff value was.

3.3 Network generation

The resulting sets of gene pairs were imported and further analyzed using the Cytoscape software environment (Cline et al. 2007). The three metrics were merged to create one network including all correlations, as well as individual smaller networks consisting of only conserved, divergent or specific correlations. The main components of these networks were selected and different parameters, such as degree, clustering coefficient, betweenness centrality and closeness centrality, were calculated using the appropriate network analysis plugins readily available in Cytoscape. Correlations between these parameters were calculated using MATLAB (2010)'s native *corrcoeff*-function, based on Pearson correlation (Section 2.3.1). The function takes in a $n \times m$ matrix and evaluates the correlation between each column.

3.4 Network cluster analysis

Cluster analysis was done using the Cytoscape plugin MCODE (Bader & Hogue 2003). Initially created for protein interaction networks, the algorithm is transferrable to

other sets of network data. MCODE analyzes the network and determines clustering in three steps. First all nodes are weighted. This is done by selecting the highest k-core of the node neighborhood and determining each node's local network density. A k-core is a subnetwork of the starting network with minimal degree k. Bader & Hogue (2003) define the parameter *core-clustering coefficient* of a node v to be the density of the highest k-core of the immediate neighborhood of v, including v.

The second step takes the highest weighted node and moves outwards from this. A percentage of the original node weight is determined as the threshold of the clustering algorithm, and nodes above this is included in the cluster and their nearest neighbors are checked against the threshold value. When no more vertices are included, a new seed node that has not been checked is selected and the process repeated. Vertices are not checked more than once in this process, resulting in non-overlapping clusters.

The third step of the MCODE algorithm post-processes the network clusters. They are filtered out if they do not contain at least a network of minimum degree 2 (2-core), and options are available for further post-processing. A "fluff" option is a parameter between 0.0 and 1.0, and all neighbors of nodes in the clusters are checked against this value if they have not already been "seen" by the algorithm. New nodes are included if the neighborhood density is higher than the fluff parameter. Using the fluff option can lead to overlap in clusters. A second option, "haircut" removes singly connected nodes from the clusters, effectively "2-coring" them. A score is given to the resulting clusters, where ranking is based on density of the clusters.

3.5 Gene ontology analysis

The general concept of gene ontology analysis is outlined in Section 2.4. In this analysis, the DAVID bioinformatics database has been used to annotate genes with gene ontology signifiers. The procedure of performing such an analysis is relatively straight-forward. For a set of genes for which function or relation is unknown, the DAVID database web interface (accessible through david.abcc.ncifcrf.gov) can perform gene ontology enrichment analysis. The layout of DAVID is shown in Appendix A.

Here, a list containing gene IDs of different origin can be inserted, or a file containing them uploaded. After the list has been submitted, DAVID will recognize species whose genome match the supplied gene list, and a selection species background can be chosen. Different gene ontology databases can be selected for matching with the gene list, and a functional annotation clustering can be performed, one example of which is given in Appendix A.

Genes are grouped according to annotations, and their function is listed, along with the fraction of supplied genes who fall under that category and the p-value of this selection happening by chance. By evaluating different clusters and the fraction of genes who are a part of the different enrichments, it can be possible to deduce the overall function of the gene cluster/module from which the gene list was extracted.

4. Results

Network representations were created using Cytoscape software package (Da Wei Huang & Lempicki 2008, Huang et al. 2009) and clusters are created using the Cytoscape MCODE plugin (Bader & Hogue 2003). Degree, betweenness centrality, closeness centrality, clustering coefficient and number of conserved, specific and divergent correlations for each node were calculated using Cytoscape's native network analysis plugin.

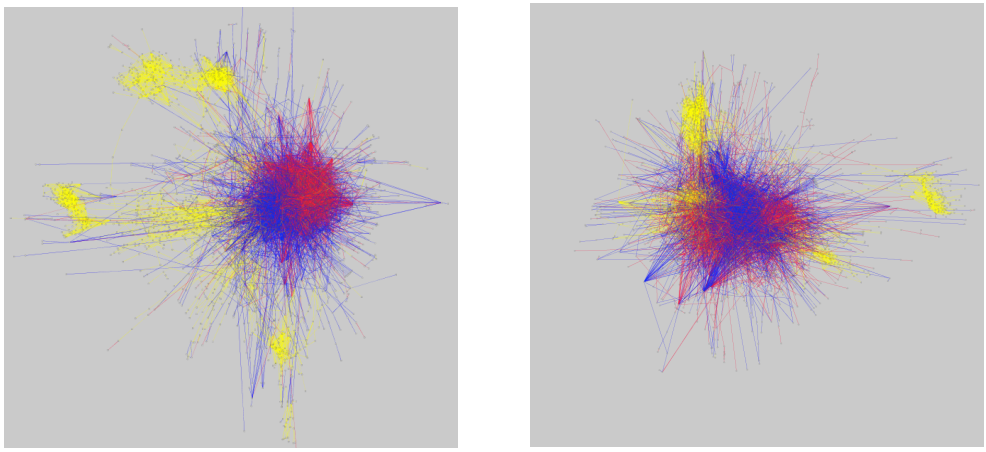
4.1 Networks

Correlation metric datasets for HIV positive and HIV negative data sets for TB active and LTBI were analyzed using Cytoscape. Extraction of the largest connected components yielded networks consisting of 4807 nodes and 14516 edges for the HIV positive dataset and 5169 nodes and 14259 edges for the HIV negative dataset, when using p-value $=5 \times 10^{-5}$ as a correlation metric threshold. Representations of the resulting networks are shown in Figure 4.1, and the number of conserved, divergent and specific correlations are listed in Table 4.1.

Table 4.1: Number of conserved, specific and divergent pairwise gene co-expression correlations for networks consisting of with HIV positive or HIV negative samples co-infected with either active TB or LTBI.

Network	**Correlation**		
	Conserved	**Specific**	**Divergent**
HIV positive (TB active vs LTBI)	4807	5097	4934
HIV negative (TB active vs LTBI)	4455	5156	4648

Two networks were also created using threshold values calculated from p-value $= 7 \times 10^{-5}$ from the HIV negative and HIV positive data pool to analyze a larger set of genes. The resulting networks were merged using Cytoscape's network merging tool. Nodes and edges not present in both networks were filtered out. The resulting network consisted of 6 main components (Figure 4.2), consisting of a total of 431 nodes and 833 edges, all of which were conserved correlations except for divergent correlations between genes CD47 and SFMBT2 and between C4orf32 and DDAH2.



(a) *(b)*

Figure 4.1: Visual representations of networks created in Cytoscape showing pairwise gene co-expression correlation between TB active and LTBI samples for (a) HIV positive samples and (b) HIV negative samples. Yellow lines signify conserved correlations, blue lines are divergent correlations and red lines are specific correlations. Nodes have been arranged using the organic layout style.

Names of genes located in the different components are listed in supplemental data S1.

4.2 Correlation analysis of network parameters

For each of the four parameters node degree and conserved, divergent and specific correlations per node, betweenness centrality, closeness centrality and clustering coefficient was evaluated for the network created from pairwise gene co-expression correlation network created with HIV positive samples, and are shown in Table 4.2. The result of correlation evaluation between node degree and conserved, divergent and specific correlations are also shown. This was calculated using the MATLAB *corrcoef*-function (Section 3.3).

Table 4.2: Pearson correlation values for selected parameters in network created from correlation between TB active and LTBI samples co-infected with HIV. Correlation values between node degree, as well as conserved, specific and divergent edges for the entire network were calculated against closeness centrality, betweenness centrality and clustering coefficient. Their respective p-values are shown in parentheses (Values of 0 occur as a result of MATLAB's lower bound of representing p-values).

	Betweenness centrality	**Closeness centrality**	**Clustering coefficient**	**Degree**
Degree	0.714 (0)	0.414 (1.39×10^{-198})	0.178 (2.13×10^{-35})	-
Conserved	0.107 (9.0×10^{-14})	0.241 (1.8×10^{-64})	0.521 (0)	0.470 (5.29×10^{-263})
Divergent	0.615 (0)	0.420 (4.2×10^{-205})	-0.147 (1.4×10^{-24})	0.532 (0)
Specific	0.516 (0)	0.535 (0)	-0.105 (2.5×10^{-13})	0.676 (0)

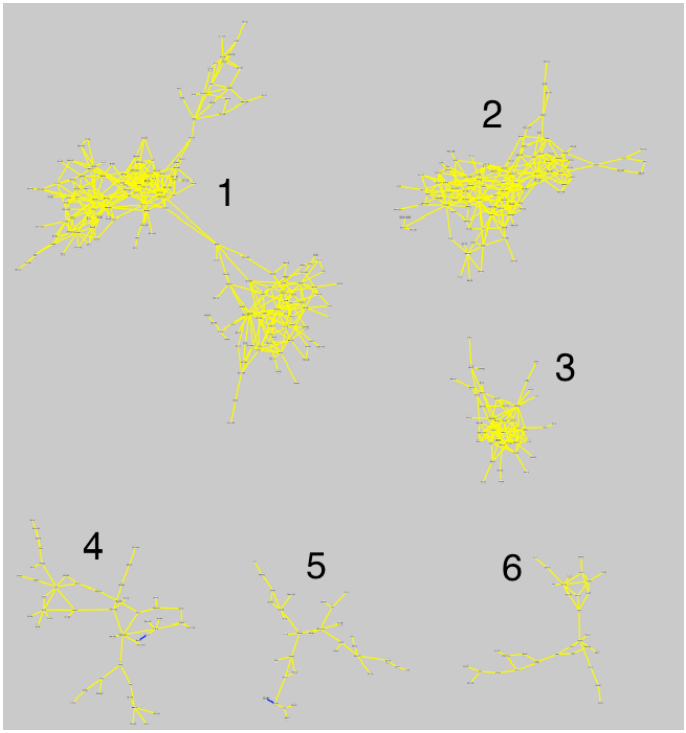


Figure 4.2: Visual representation created in Cytoscape showing genes and their correlations present in both HIV negative and HIV positive data sets. The components are numbered according to size. Yellow edges signify conserved edges, while blue signify divergent edges. Only two pairs of genes exhibit divergent correlations.

4.3 Correlation and centrality analysis of hubs

Using degree k as a measure of hub significance, the ten highest ranking hubs were located for the network created from pairwise gene co-expression correlation between active TB and latent TB infection samples co-infected with HIV. Figure 4.3 shows their placement in the network.

Table 4.3 lists hub degree, betweenness centrality, closeness centrality and clustering coefficient, as well as the fraction of conserved, divergent and specific correlations these hubs are a part of. The total fraction for hub correlations are shown in Figure 4.4.

For each of the top hubs, the average percentage of conserved, specific and divergent correlations among their nearest neighbors was also calculated, as shown in Table 4.3, along with the average degree of each hubs neighborhood and the highest degree among their neighboring nodes. A list of gene names for each hub neighborhood is given in supplemental file S1.

4.4 Clustering

Cluster analysis was done using the Cytoscape plugin MCODE (Bader & Hogue 2003). With degree cutoff 2, and 'haircut' and 'fluff' options on. Node density cutoff

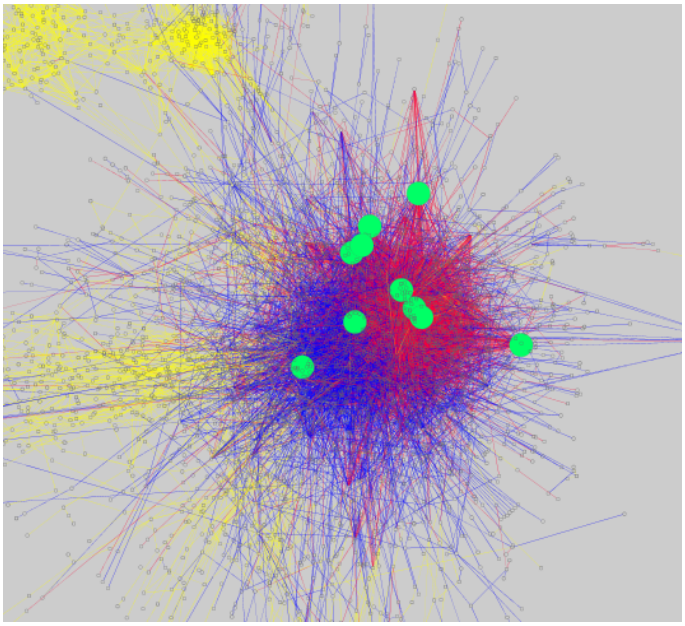


Figure 4.3: Location of hubs shown in closeup of network of pairwise gene co-expression correlation between active TB and LTBI co-infected with HIV. Green circles represent hubs.

was set to 0.1, node score cutoff was set to 0.2, K-core was set to 2, and Max. depth was set to 100. The four clusters with the highest ranking based on the MCODE algorithm had scores of 8.187 through 5.176, and their location in the network is shown in Figure 4.5. Clusters with lower ranking either overlapped with one of higher ranked clusters or consisted of to few genes to be of any significance. Clusters consisted mainly of conserved correlations, with a few divergent correlations in cluster 1 and 4. Cluster closeups are shown in Appendix C, and genes located in each of the clusters are given in supplemental file S1.

When comparing the different modules against the components preserved between HIV positive and HIV negative networks, there was found to be an overlap of approx. 80% of genes present in modules and genes present in the six components.

4.5 Functional analysis of clusters and hubs

4.5.1 Gene ontology analysis of conserved components between HIV positive and HIV negative sample pools

Gene ontology analysis using DAVID for each of the conserved components found when comparing the resulting HIV positive network against HIV negative network (Figure 4.2 resulted in generally low enrichment scores. DAVID evaluated the network to mainly constitute of genes related to cell maintenance and proliferation, finding minor enrichments for transcription activity, organelle structures (ribosomes, mitochondrion, etc.), transport proteins (GTPase and ion channels), a high enrichment of DNA/RNA-binding proteins (i.e. zinc fingers) and membrane proteins involved in electron transport. Component 3 stood out having very high enrichment of ribosomal

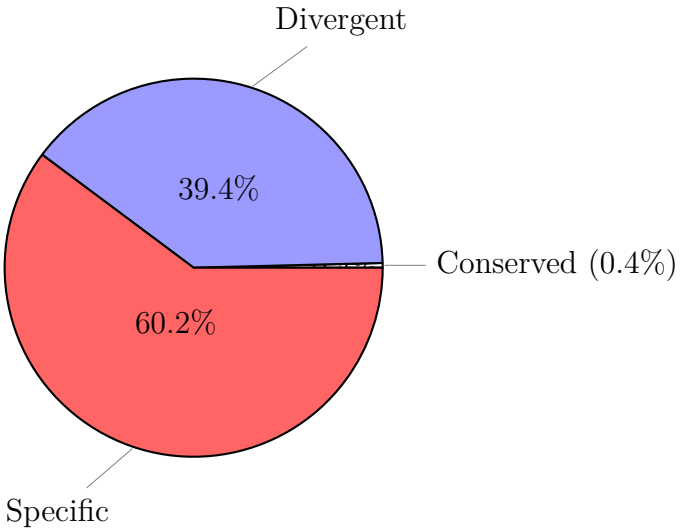


Figure 4.4: Distribution of conserved (0.4%), specific (60.2%) and divergent(39.4%) pairwised co-expression correlations for the top ten hubs in network created from pairwise co-expression correlation between gene pairs in TB active and LTBI samples co-infected with HIV. Threshold determined for edges included in network was calculated using p-value<0.00005*. Hubs ranked based on node degree (k).*

proteins and ribonucleoproteins, with 23 out of 42 genes grouped under production of components for ribonucleotides. A closeup of component 3 with gene names is given in Appendix B.

4.5.2 Gene ontology analysis of clusters in network of HIV positive samples

An analysis of each of the four main clusters in the network consisting of HIV positive samples showed many of the same characteristics as the analysis of the conserved components between HIV positive and HIV negative samples. Low general enrichment, but with some enrichment in production of cell components, cell life cycle, DNA repair and maintenance, metal binding-proteins and cell respiration.

Module 4 was highly enriched with genes involved in ribosome production and activity, the same as component 3 in the preserved network between HIV positive and HIV negative. A comparison of these two modules showed that 38 genes were shared between them (out of 43 genes in component 3 and 51 genes in cluster 4).

4.5.3 Gene ontology analysis of hubs

Gene ontology analysis was performed for each of the ten highest ranking hubs and their neighborhoods. None of the neighborhoods exhibited especially high functionality enrichments, but the general trend for all of the neighborhoods were genes encoding products used in nucleotide binding, energy pathways, apoptosis, transcription, RNA processing and immune response (including leukocyte and T

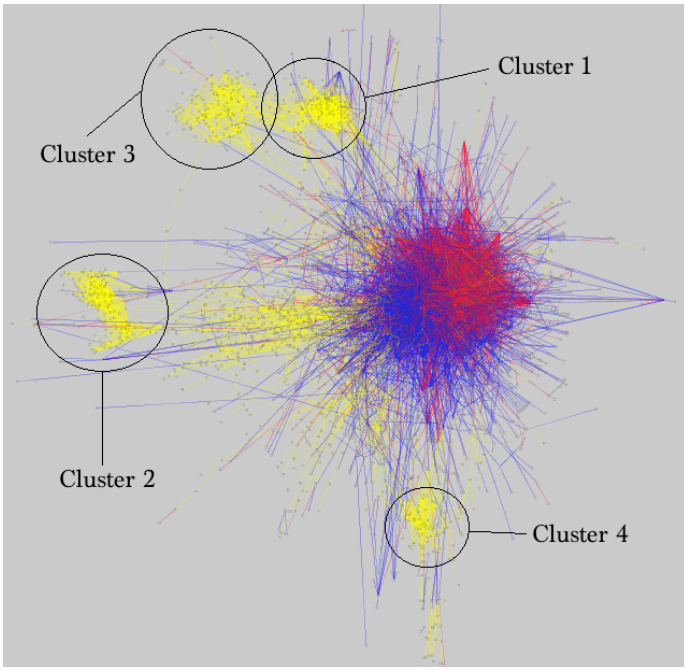


Figure 4.5: Clustering in network created from pairwise co-expression correlation between gene pairs in TB active and LTBI samples co-infected with HIV. Clusters were found using the Cytoscape MCODE plugin, and are ranked according to clustering score given by MCODE.

cell regulation). The occurrence of genes involved in immune response was higher than that of genes present i clusters, though the individual hubs themselves did not represent important molecular functions, nor direct involvement in immune response.

4.6 Detection of TB fingerprint genes

Kaforou et al. (2013) Defined a set of "fingerprint" genes, genes whose transcription profile was unique to samples infected with active TB. Approximately 50% of these genes were detected in the network created from active TB and LTBI samples co-infected with HIV. Fingerprint genes were heavily enriched with either divergent or specific correlations, and their immediate neighborhood were also part of the large area in the network consisting of mainly divergent and specific correlations.

Table 4.3: The ten highest ranking hubs from network generated from pairwise gene co-expression correlation data for samples with active TB compared to samples with latent TB infection. Both sample groups were co-infected with HIV. Hub significance is based on degree k, and the betweenness centrality, closeness centrality and clustering coefficient has been calculated for each of the ten genes, as well as for the neighborhood consisting of each genes nearest neighbors. The fraction of conserved, divergent and specific gene pair correlation have also been calculated for each of the ten hubs, as well as the average fraction in these genes' neighborhood. Fractions over 0.6 are marked with blue. The average degree of the hub neighborhood, as well as the highest degree amongst those neighbors are also specified (gene name in parentheses). Network average for node degree, betweenness centrality, closeness centrality and clustering coefficient are listed as a reference.

Gene	Degree	Betweenness centrality	Closeness centrality	Clustering coefficient	Fraction of Conserved	Fraction of Specific	Fraction of Divergent
OAT	101	0.0342	0.2626	0.0024	0	0.07	0.93
STOM	93	0.0405	0.2631	0.0072	0	0.31	0.69
IDH2	79	0.0163	0.2588	0.0133	0	0.95	0.05
DDX39	73	0.0188	0.2510	0.0156	0	0.88	0.12
WBSCR22	69	0.0142	0.2609	0.0200	0.01	0.99	0
SLC25A5	69	0.0274	0.0255	0.0057	0	0.03	0.97
SARS2	67	0.0282	0.2628	0.0086	0	0.85	0.15
CNNM3	63	0.0248	0.2628	0.0102	0.03	0.44	0.53
MRPL38	61	0.0152	0.2629	0.0126	0	0.92	0.08
ATIC	61	0.0162	0.2602	0.0186	0	0.93	0.7
Network average	6	0.0001	0.1832	0.0635			

Gene	Average degree of neighborhood	Maximum degree in neighborhood	Average fraction in neighborhood of Conserved	Average fraction in neighborhood of Specific	Average fraction in neighborhood of Divergent
OAT	12	69 (SLC25A5)	0.07	0.58	0.35
STOM	12	73 (DDX39)	0.05	0.36	0.58
IDH2	18	61 (ATIC)	0.04	0.14	0.82
DDX39	18	93 (STOM)	0.04	0.29	0.67
WBSCR22	23	73 (DDX39)	0.04	0.23	0.73
SLC25A5	14	101 (OAT)	0.12	0.76	0.12
SARS2	16	73 (RFTN1)	0.08	0.23	0.68
CNNM3	17	73 (DDX39)	0.06	0.48	0.46
MRPL38	20	73 (DDX39)	0.02	0.23	0.75
ATIC	21	79 (IDH2)	0.07	0.20	0.73

5. Discussion

5.1 Evaluation of network layout and parameters

The two main networks created from the HIV positive and HIV negative data sets (Figure 4.1) exhibited similar network traits. In both networks conserved correlations grouped together in clusters that were connected to a larger main group consisting of specific and divergent correlations, suggesting that the main buildup of the pairwise gene co-expression networks are the same, even if the constituents of the networks differ.

Cluster analysis of the HIV positive network agrees with the network layout observed, where clustering is clearly shown to be a characteristic prominent in nodes with a high fraction of conserved correlations (Figure 4.5), while the larger group of divergent and specific correlations exhibit insignificant clustering when analyzed using the Cytoscape MCODE software. When evaluating the correlation between conserved, specific and divergent metrics and clustering coefficient, as listed in Table 4.2, the same picture is drawn, with high correlation between conserved correlations and clustering coefficient, while divergent and specific correlations show slight negative correlation with clustering coefficient.

The average betweenness centrality and clustering coefficient of the network, as seen in Table 4.3, was quite low, suggesting the network layout was dominated by large areas of nodes connected to each other by a few central nodes or hubs.

Calculations showed very strong correlation between node degree and betweenness centrality, reflected by the fact that the top ranking hubs showed high values of betweenness centrality. These observations point to a situation where hubs are responsible for connecting different parts to produce the resulting network. This is mirrored in the correlation to closeness centrality, where divergent and specific correlations rank high. Significant correlation to closeness centrality suggests that genes high in divergent and specific pairwise co-expressions are tightly linked to the rest of the network as a whole. Betweenness centrality was also highly correlated with divergent and specific pairwise co-expressions, which was expected since these are associated with the highest ranking hubs in the system.

5.2 Evaluation of hub analysis

The top ranking hubs were found to be genes with a high fraction of either divergent or specific correlations. Interestingly, when analyzing the same score for their immediate neighbors, there was a markedly different distribution of correlation behaviors, where almost all neighborhoods of "specific" hubs were heavily involved in divergent pairwise co-expression correlations and vice versa. This hub characteristic can imply that the highly connected hubs act as "bridges" between specific and divergent areas in the network, further implicating their importance to the overall functions of the network. As seen in Table 4.3, there was also a tendency for hubs to be connected to other high-ranking hubs, one of which was the DDX39 gene, a prominent participant in many of the other hubs' neighborhoods.

One hub, the CNNM3 gene, stood out as having a neighborhood which was neither heavily specific or divergent, reflecting on its own distribution of correlations, where almost half of its correlations are specific and the other half is divergent. The STOM gene showed a similar, though not as strong, pattern for its nearest neighbors, but was itself markedly more involved in divergent pairwise co-expression correlations.

5.3 Biological interpretations of network analysis

5.3.1 Clustering of conserved correlations

High clustering of conserved correlations may indicate that these are gene modules responsible for the main, vital processes of the cell. Functions and processes that are needed regardless of otherwise environmental perturbations or pathogenic infections. The importance of these conserved correlations is strengthened by the fact that a portion of the conserved clusters where preserved between HIV positive data and HIV negative data.

Gene ontology analysis showed that the clusters where mainly involved in functions that are indeed vital to cell proliferation, such as cell respiration, DNA repair and transcription of RNA. Their location, away from the network center, also points to the fact that their functionality is less affected by alterations other places in the network.

An explanation may be that conserved modules that are vital to main cell functions need to be less sensitive to alterations in the network, thus reducing their interaction with other parts of the network. Another explanation can be that if an alteration should occur in one of these clusters, the rest of the network is less likely to be heavily affected by these changes, leaving the other vital functions unperturbed until the normal state of the gene cluster is restored. This could also explain why the different clusters seem to be located "distant" from each other as well as the main bulk of the network, since this could improve the autonomy of the clusters.

5.3.2 Biological characteristics of divergent and specfic correlations

By studying the main hubs, divergent and specific correlations were found to be highly incorporated with each other in the resulting network. This may imply that there is a complex and interrelated response to changes in infection status of the patients. Co-infection of HIV and TB, as pointed out in Section 2.5.3, has been found to be a more complex situation than monoinfection, caused by TB and HIV's ability to potentiate each other, and the resulting complex network of divergent and specific correlations may reflect this pathogen interaction.

Gene ontology analysis found that genes involved in immune response was more enriched in the specific and divergent correlations than in the conserved, implying that TB and HIV co-infection leads to a new regime of immune response compared to HIV infection alone. Detection of fingerprint genes from the Kaforou et al. (2013) article in specific and divergent correlations further bolster this implication. It is expected that these genes should not appear as conserved genes, since their function is what identified them as specific to active TB infection in the initial study.

The enrichment of immune response in specific and divergent correlations could also occur as a result of HIV's role in disturbing the functionality of the immune system, as some of the genes were responsible for negative regulation of T cells and lymphocytes. This would imply that the occurrence of active TB in these patients is a result of earlier deactivation of the immune system by HIV, and that latent TB could have converted to active state. It is difficult to say something conclusive about the role of these pairwise gene correlations, since it has not been evaluated whether the co-expression profiles are specific to either patients with active TB or latent TB infection, and how divergent correlations are expressed in each of the cases. Evaluation of the individual expression profiles for each probe might clarify this.

6. Concluding remarks

The network analysis method developed by Voigt et al. (2015) shows great promise in extracting biological information from gene expression data sets. The resulting gene co-expression correlation networks found in this analysis show that the method is capable of correctly identifying the different characteristics of pairwise gene correlation. Preservation of conserved correlations between systems (HIV positive samples and HIV negative samples) and the detection of fingerprint genes in specific/divergent areas both imply this.

Pairwise gene co-expression correlation analysis has been shown to effectively separate correlations that are significant and possibly exclusive to HIV negative samples from those in HIV positive samples, while at the same time preserving the conserved correlations shared by both systems. At the same time this method can help elucidate the functional connection between conserved modules and genetic relations that are specific to different environmental perturbations, in this case pathogenic infection status.

Further work is needed to specify the roles of network structures that have been detected in this analysis. By evaluating the expression values of the different probes in latent and active TB infection, it could be possible to select co-expression correlations that are specific to each case, and possible detect functional gene correlations that can be used to improve detection of latent TB or even support other research in treatment of HIV and TB co-infection.

It could be beneficial to perform a new analysis on the same data using even stricter thresholds, to limit the amount of included genes further. This could possibly give an even clearer picture of the essential clusters and the role of the specific and divergent correlations. Using a different method for selecting genes to analyze could also be of interest, to evaluate different aspects of the infection response, for example by choosing a set of genes whose immune response already has been defined, or by analyzing genes involved in signal transmission, since these genes also would be of great importance in immune response.

Introducing weighted correlation metric thresholding to the analysis is also a possible path to improve the understanding of the connection between network and biological function. Weighted correlations could help determine which pairwise correlations are of more significance without losing too much of the underlying information in the network.

Lastly, expanding the analysis to include all of the possible pairings of sample groups provided by Kaforou et al. (2013) could help determining whether the general pattern of these network structures incorporate all of the different infection scenarios, and if the scope of this method can be expanded. This would further strengthen the method as a tool for examining biological systems and interactions.

7. Bibliography

Albert, R. & Barabási, A.-L. (2002), 'Statistical mechanics of complex networks', *Reviews of modern physics* **74**(1), 47.

Albert, R., Jeong, H. & Barabási, A.-L. (2000), 'Error and attack tolerance of complex networks', *Nature* **406**(6794), 378–382.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000), 'Gene ontology: tool for the unification of biology', *Nature genetics* **25**(1), 25–29.

Bader, G. D. & Hogue, C. W. (2003), 'An automated method for finding molecular complexes in large protein interaction networks', *BMC bioinformatics* **4**(1), 2.

Brenchley, J. M., Price, D. A., Schacker, T. W., Asher, T. E., Silvestri, G., Rao, S., Kazzaz, Z., Bornstein, E., Lambotte, O., Altmann, D., Blazar, B. R., Rodriguez, B., Teixeira-Johnson, L., Landay, A., Martin, J. N., Hecht, F. M., Picker, L. J., Lederman, M. M., Deeks, S. G. & Douek, D. C. (2006), 'Microbial translocation is a cause of systemic immune activation in chronic hiv infection', *Nat Med* **12**(12), 1365–1371.

Cameron, D. W., Simonsen, J. N., D'Costa, L. J., Ronald, A. R., Maitha, G. M., Gakinya, M. N., Cheang, M., Ndinya-Achola, J. O., Piot, P. & Brunham, R. C. (1989), 'Female to male transmission of human immunodeficiency virus type 1: risk factors for seroconversion in men', *Lancet* **2**(8660), 403–7.

Chen, D., Lü, L., Shang, M.-S., Zhang, Y.-C. & Zhou, T. (2012), 'Identifying influential nodes in complex networks', *Physica a: Statistical mechanics and its applications* **391**(4), 1777–1787.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B. et al. (2007), 'Integration of biological networks and gene expression data using cytoscape', *Nature protocols* **2**(10), 2366–2382.

Cunningham, A. L., Donaghy, H., Harman, A. N., Kim, M. & Turville, S. G. (2010), 'Manipulation of dendritic cell function by viruses', *Curr Opin Microbiol* **13**(4), 524–529.

Da Wei Huang, B. T. S. & Lempicki, R. A. (2008), 'Systematic and integrative analysis of large gene lists using david bioinformatics resources', *Nature protocols* **4**(1), 44–57.

De Cock, K. M., Fowler, M. G., Mercier, E., de Vincenzi, I., Saba, J., Hoff, E., Alnwick, D. J., Rogers, M. & Shaffer, N. (2000), 'Prevention of mother-to-child hiv transmission in resource-poor countries: translating research into policy and practice', *JAMA* **283**(9), 1175–82.

Dong, J. & Horvath, S. (2007), 'Understanding network concepts in modules', *BMC systems biology* **1**(1), 24.

Ebersberger, I., Metzler, D., Schwarz, C. & Pääbo, S. (2002), 'Genomewide comparison of dna sequences between humans and chimpanzees', *The American Journal of Human Genetics* **70**(6), 1490–1497.

Ensoli, B., Barillari, G., Salahuddin, S. Z., Gallo, R. C. & Wong-Staal, F. (1990), 'Tat protein of hiv-1 stimulates growth of cells derived from kaposi's sarcoma lesions of aids patients', *Nature* **345**(6270), 84–86.

Gardy, J. L., Johnston, J. C., Sui, S. J. H., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R. et al. (2011), 'Whole-genome sequencing and social-network analysis of a tuberculosis outbreak', *New England Journal of Medicine* **364**(8), 730–739.

Getahun, H., Gunneberg, C., Granich, R. & Nunn, P. (2010), 'Hiv infection-associated tuberculosis: the epidemiology and the response', *Clin Infect Dis* **50**(3), S201–207.

Grossman, Z., Meier-Schellersheim, M., Paul, W. E. & Picker, L. J. (2006), 'Pathogenesis of hiv infection: what the virus spares is as important as what it destroys', *Nature medicine* **12**(3), 289–295.

He, X. & Zhang, J. (2006), 'Why do hubs tend to be essential in protein networks?', *PLoS Genet* **2**(6), e88.

Horvath, S. (2011), *Weighted Network Analysis: Applications in Genomics and Systems Biology*, Springer Science and Business Media.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009), 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic acids research* **37**(1), 1–13.

Hudson, R. R., Kreitman, M. & Aguadé, M. (1987), 'A test of neutral molecular evolution based on nucleotide data', *Genetics* **116**(1), 153–159.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. & Friend, S. H. (2000),

'Functional discovery via a compendium of expression profiles', *Cell* **102**(1), 109–126.

Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. (2001), 'Lethality and centrality in protein networks', *Nature* **411**(6833), 41–42.

Kaforou, M., Wright, V. J., Oni, T., French, N., Anderson, S. T., Bangani, N., Banwell, C. M., Brent, A. J., Crampin, A. C., Dockrell, H. M., Eley, B., Heyderman, R. S., Hibberd, M. L., Kern, F., Langford, P. R., Ling, L., Mendelson, M., Ottenhoff, T. H., Zgambo, F., Wilkinson, R. J., Coin, L. J. & Levin, M. (2013), 'Detection of tuberculosis in hiv-infected and -uninfected african adults using whole blood rna expression signatures: a case-control study', *PLoS Med* **10**(10), e1001538.

Khaitovich, P., Weiss, G., Lachmann, M., Hellmann, I., Enard, W., Muetzel, B., Wirkner, U., Ansorge, W. & Pääbo, S. (2004), 'A neutral model of transcriptome evolution', *PLoS biology* **2**(5), e132.

MATLAB (2010), *version 7.10.0 (R2010a)*, The MathWorks Inc., Natick, Massachusettes, United States.

Mayer, K. H. & Venkatesh, K. K. (2011), 'Interactions of hiv, other sexually transmitted diseases, and genital tract inflammation facilitating local pathogen transmission and acquisition', *Am J Reprod Immunol* **65**(3), 308–16.

Mitchell, P. J. & Tjian, R. (1989), 'Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins', *Science* **245**(4916), 371–378.

Newman, M. E. (2003), 'The structure and function of complex networks', *SIAM review* **45**(2), 167–256.

Nowick, K., Gernat, T., Almaas, E. & Stubbs, L. (2009), 'Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain', *Proceedings of the National Academy of Sciences* **106**(52), 22358–22363.

Oldham, M. C., Horvath, S. & Geschwind, D. H. (2006), 'Conservation and evolution of gene coexpression networks in human and chimpanzee brains', *Proceedings of the National Academy of Sciences* **103**(47), 17973–17978.

Oliphant, A., Barker, D. L., Stuelpnagel, J. R. & Chee, M. S. (2002), 'Beadarray technology: enabling an accurate, cost-effective approach to high-throughput genotyping', *Biotechniques* **32**(6), 56–58.

Pawlowski, A., Jansson, M., Sköld, M., Rottenberg, M. E. & Källenius, G. (2012), 'Tuberculosis and hiv co-infection', *PLoS Pathog* **8**(2), e1002464.

Radicchi, F. (2011), 'Who is the best player ever? a complex network analysis of the history of professional tennis', *PloS one* **6**(2), e17249.

Ruan, J., Dean, A. K. & Zhang, W. (2010), 'A general co-expression network-based approach to gene expression analysis: comparison and applications', *BMC systems biology* **4**(1), 8.

Schena, M. (2003), *Microarray Analysis*, 1st edition edn, John Wiley and Sons,Inc.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. & Friedman, N. (2003), 'Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data', *Nat Genet* **34**(2), 166–176.

Shaulian, E. & Karin, M. (2001), 'Ap-1 in cell proliferation and survival.', *Oncogene* **20**(19), 2390–2400.

Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. (2003), 'A gene-coexpression network for global discovery of conserved genetic modules', *Science* **302**(5643), 249–255.

Voigt, A., Nowick, K. & Almaas, E. (2015), 'A composite network of conserved and tissue specific gene interactions', *in progress* .

Weiss, R. A. (1993), 'How does hiv cause aids?', *Science* **260**(5112), 1273–1279.

Zhang, B. & Horvath, S. (2005), 'A general framework for weighted gene co-expression network analysis', *Statistical applications in genetics and molecular biology* **4**(1).

Appendices

A. DAVID gene ontology utility

Layout and output from the DAVID gene ontology functional annotation clustering utility.

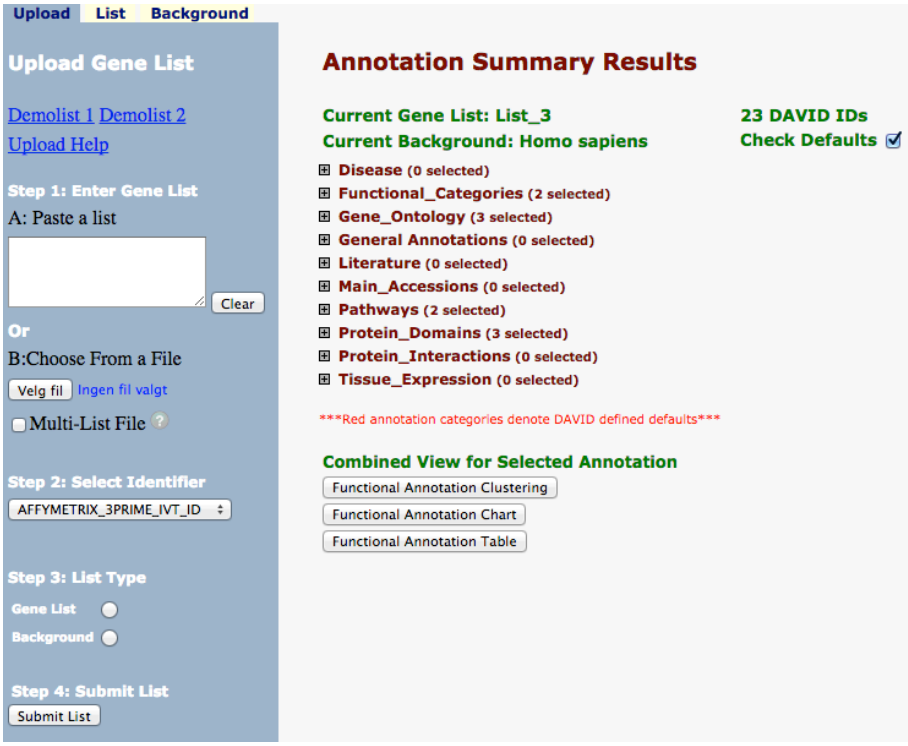


Figure A.1: The main layout of the DAVID gene ontology utility web interface.

	Annotation Cluster 1	Enrichment Score: 9.17	G		Count	P_Value	Benjamini
☐	SP_PIR_KEYWORDS	ribonucleoprotein	RT		13	3.8E-17	1.7E-15
☐	KEGG_PATHWAY	Ribosome	RT		12	5.5E-16	3.9E-15
☐	GOTERM_CC_FAT	ribosome	RT		12	3.4E-15	1.8E-13
☐	SP_PIR_KEYWORDS	ribosomal protein	RT		11	3.6E-15	8.2E-14
☐	GOTERM_BP_FAT	translational elongation	RT		10	4.3E-15	6.7E-13
☐	GOTERM_MF_FAT	structural constituent of ribosome	RT		11	8.4E-15	4.6E-13
☐	GOTERM_CC_FAT	ribonucleoprotein complex	RT		14	3.8E-14	9.7E-13
☐	GOTERM_BP_FAT	translation	RT		12	1.0E-13	7.8E-12
☐	SP_PIR_KEYWORDS	ribosome	RT		8	1.4E-12	2.1E-11
☐	GOTERM_CC_FAT	ribosomal subunit	RT		9	9.2E-12	1.6E-10
☐	GOTERM_CC_FAT	cytosolic ribosome	RT		8	2.3E-11	2.9E-10
☐	SP_PIR_KEYWORDS	protein biosynthesis	RT		8	1.1E-9	1.3E-8
☐	GOTERM_CC_FAT	cytosolic part	RT		8	2.0E-9	2.0E-8
☐	GOTERM_MF_FAT	structural molecule activity	RT		11	4.5E-9	1.2E-7
☐	GOTERM_CC_FAT	cytosol	RT		12	1.0E-6	8.7E-6
☐	GOTERM_CC_FAT	large ribosomal subunit	RT		5	3.1E-6	2.3E-5
☐	GOTERM_CC_FAT	cytosolic large ribosomal subunit	RT		4	2.7E-5	1.7E-4
☐	GOTERM_CC_FAT	cytosolic small ribosomal subunit	RT		4	3.1E-5	1.8E-4
☐	GOTERM_CC_FAT	non-membrane-bounded organelle	RT		13	1.2E-4	5.4E-4
☐	GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		13	1.2E-4	5.4E-4
☐	GOTERM_CC_FAT	small ribosomal subunit	RT		4	1.2E-4	5.2E-4
☐	GOTERM_MF_FAT	RNA binding	RT		7	4.1E-4	7.5E-3
☐	SP_PIR_KEYWORDS	phosphoprotein	RT		9	6.3E-1	9.7E-1

Figure A.2: Example of DAVID functional annotation clustering output. One of the clusters are shown.

B. Component closeup

Figure B.1 shows a closeup of component 3, including gene names, from the network consisting of nodes and correlations present in both HIV positive and HIV negative pairwise gene co-expression correlation networks.

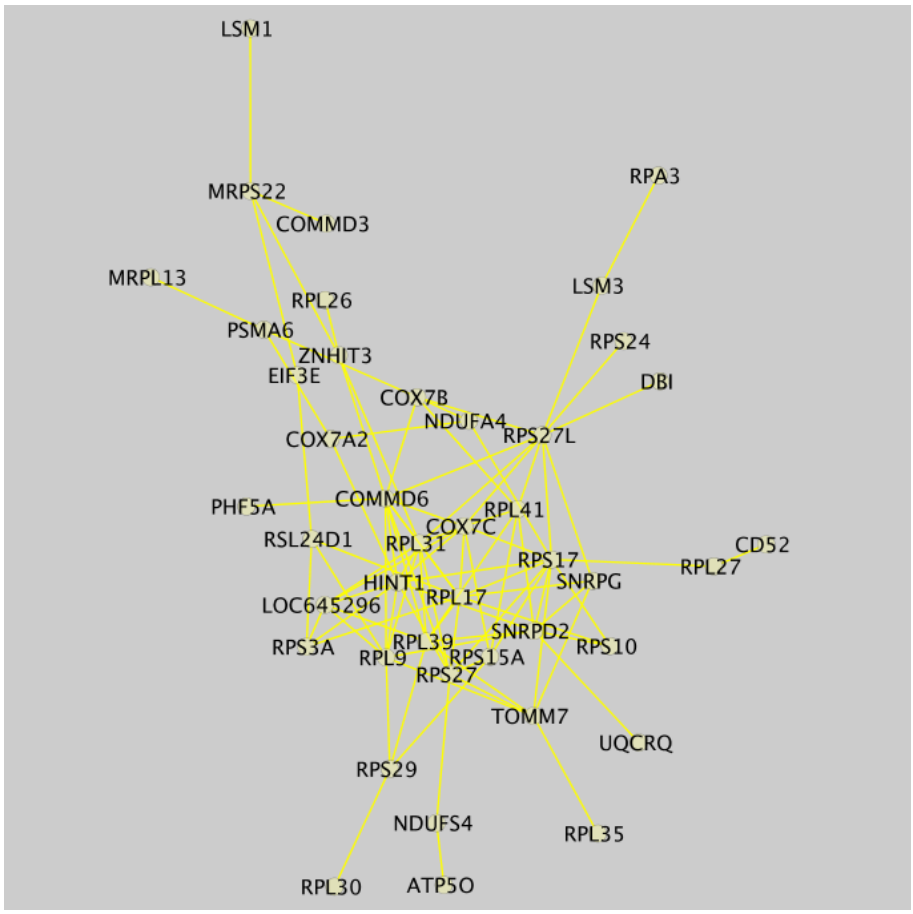
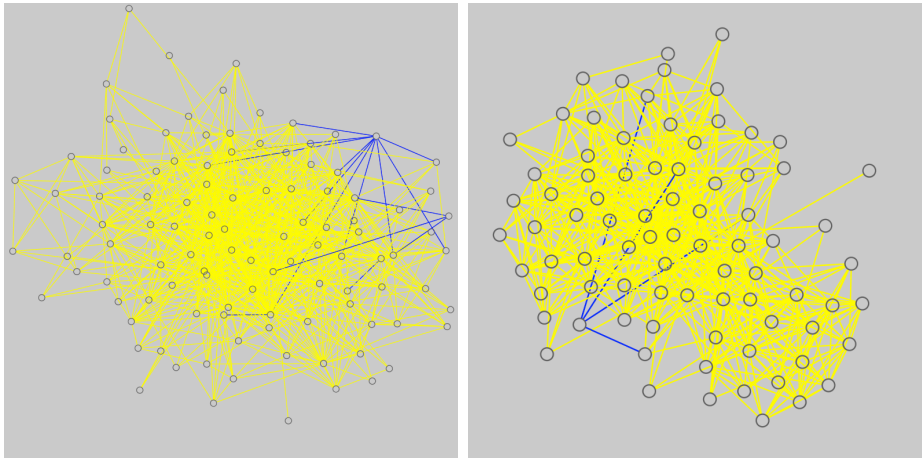


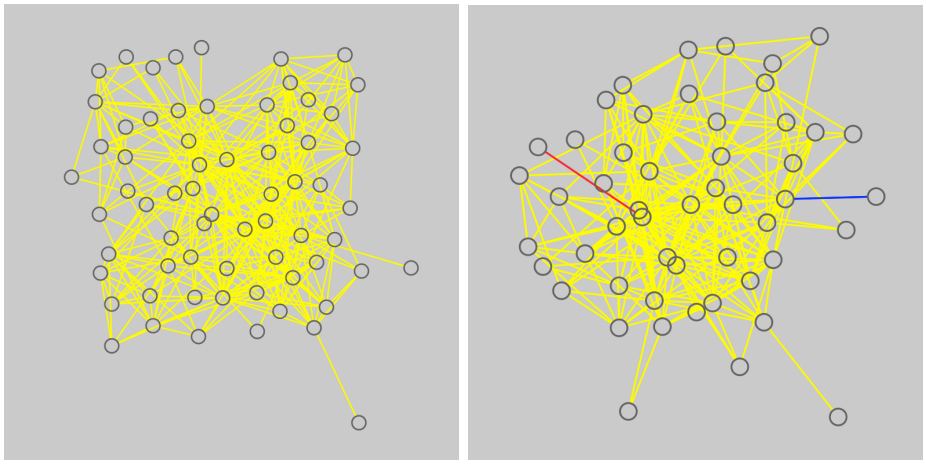
Figure B.1: Correlation network for component detected in preserved gene correlations between HIV positive and HIV negative samples. This component had the highest gene ontology enrichment score of the six components when analyzed in DAVID. All correlations are conserved in regards to metric score.

C. Cluster closeups

Figure C.1 shows closeups of clusters found using the Cytoscape MCODE plugin on data from correlation analysis of samples with active TB versus latent TB infection, co-infected with HIV.



(a) Cluster 1 *(b) Cluster 2*



(c) Cluster 3 *(d) Cluster 4*

Figure C.1: Closeup of clusters from correlation analysis of samples with active TB versus latent TB infection, co-infected with HIV. Yellow edges signify conserved correlations, blue divergent and red are specific correlations.