Analysis of the Gene Expression Patterns of the Human Immunome

Master's thesis
Heidi Ali
Institute of Medical Technology
University of Tampere
June 2009

Preface

This work was done in the group of Bioinformatics, Institute of Medical Technology, University of Tampere. The Bioinformatics group, lead by Professor Mauno Vihinen, provided me good facilities and a positive climate to carry out this project.

First of all I would like to thank Csaba Ortutay, PhD, for excellent and well-informed guidance. You arranged for me this very interesting, organized and feasible project. My knowledge of Bioinformatics was limited when I started this project and you have patiently shared your wisdom. You explained me facts of immunome, taught programming with R and various other things involved in this project. I really learned a lot from you!

I would like to thank Professor Mauno Vihinen firstly for accepting me as a student of Bioinformatics and as a member of Bioinformatics group and secondly for guiding me in this project. You noticed so much revising in my writing, but I understand it was for my best to improve my thesis and to make me a better writer.

Thank you Salli Kukkurainen, MSc, for proofreading my thesis. Your excellent linguistic skill helped a lot. Thanks for your support while doing this work and writing about it. You have been a good friend and support during this long and hard journey!

Thank you, Jukka Lehtiniemi, for your help in editing the pictures and in making the presentations of the thesis.

Thank you, Hannu Korhonen, for all the help in problems with my computer. You installed Windows again and again. Thanks to Toni Vormisto and Amer Mezian too in solving the computer problems.

Thank you Katri Hituri, MSc, for your support and giving me a good idea for discussion.

Thanks to Sofia, Sanna, Janita, Jouni and Timo for your support, even when you did not understand what I am actually doing.

Thanks to all relatives and friends who were asking me constantly when is your thesis ready. Without this pressure it would probably still be unfinished.

Thank you, Aino, for not tearing this thesis to pieces as you did with my Bachelor's thesis.

In Tampere, June 2009

Heidi Ali

MASTER`S THESIS

Place:	University of Tampere Faculty of Medicine Institute of Medical Technology (IMT) Bioinformatics group
Author:	Ali, Heidi Susan
Title:	Analysis of the Gene Expression Patterns of the Human Immunome
Pages:	54
Supervisors:	Csaba Ortutay, Ph.D. and Professor Mauno Vihinen, Ph.D.
Reviewed by:	Professor Mauno Vihinen and Csaba Ortutay
Date:	June 2009

Abstract

Background and aims: Genes and proteins involved in immune system are called immunome. A network is a system where nodes are connected to each other by edges. In gene network genes are nodes and tissues are edges, and vice versa in tissue networks. Network theory helps to find out the features of the network, such as degree, closeness and community structure of the network. The main aim of this study was to find out the patterns of gene expression in the immunome. Other aims were to find out the correlation between immunome gene expression and protein-protein interactions and the evolutionary age of important genes.

Methods: Immunome gene and tissue networks were created. Gene and tissue communities were mined from the networks with two community analyses. Common gene and tissue clusters were collected from these two community analyses. Degree and closeness values of clustered genes were compared to degree and closeness of immunome protein-protein interactions. The evolutionary age of clustered genes was studied. The correlation between degree in the tissue network and the number of genes in that tissue in the gene network was checked.

Results: The most important result was the discovery of 88 immunome gene clusters holding together 547 genes. The genes in the clusters have similar gene expression patterns. There is no correlation between immunome gene clusters and immunome PPI data. Degree and closeness values are divided evenly to different evolutionary levels. The tissue network yielded 203 immunome cluster tissues, which have similar sets of genes expressed in them. There is a strong correlation between degree of a tissue in tissue network and the number of genes in that tissue in gene network.

Conclusions: Gene and tissue networks were created and common gene and tissue clusters found successfully with help of two community analyses. The main aim was reached by finding 547 immunome clustered genes. Tissue analysis revealed 203 immunome cluster tissues. There was no apparent correlation between clustered genes and immunome PPIs or the evolutionary age of the gene.

PRO GRADU-TUTKIELMA

Paikka:	Tampereen yliopisto Lääketieteellinen tiedekunta Lääketieteellisen teknologian instituutti (IMT) Bioinformatiikka-ryhmä
Tekijä:	Ali, Heidi Susan
Otsikko:	Ihmisen immunomin ilmentymisen analyysi
Sivumäärä:	54
Ohjaajat:	Csaba Ortutay, FT and professori Mauno Vihinen, FT
Tarkastajat:	Professori Mauno Vihinen ja Csaba Ortutay
Aika:	Kesäkuu 2009

Tiivistelmä

Tutkimuksen tausta ja tavoitteet: Immuunijärjestelmän geenejä ja proteiineja kutsutaan immunomiksi. Verkko on systeemi, jossa solmut yhdistyvät toisiinsa kaarilla. Geeniverkossa geenit ovat solmuja ja ilmentymiskudokset kaaria, kudosverkossa päinvastoin. Verkkoteorian avulla voidaan löytää verkon ominaisuuksia, kuten aste, läheisyys ja verkon sisäinen rakenne. Tämän tutkimuksen päätavoite oli löytää immunomin geenien ilmentymisryväksiä. Lisäksi tavoitteena oli tutkia korrelaatiota immunomin geenien ilmentymisen ja proteiini-proteiini-vuorovaikutusten sekä tutkia tärkeiden geenien evolutionaarista ikää.

Tutkimusmenetelmät: Työssä luotiin immunomin geeni- ja kudosverkot sekä etsittiin niiden sisäisiä rakenteita kahdella eri menetelmällä. Yhteiset geeni- ja kudosryväkset, jotka saatiin menetelmien avulla, kerättiin talteen. Ryväsgeenien asteita ja läheisyyksiä verrattiin proteiini-proteiiniverkon asteisiin ja läheisyyksiin. Tutkittiin myös ryväsgeenien evolutionaarista ikää sekä korrelaatiota kudosverkon asteiden ja geeniverkon geenien lukumäärän välillä.

Tutkimustulokset: Tutkimuksen tärkein tulos oli 88 immunomin geeniryvästä, joissa on yhteensä 547 geeniä. Ryväksen geeneillä on samankaltainen ilmentyminen. Immunomin geeniryväksillä ja proteiini-proteiini-vuorovaikutuksilla ei havaittu korrelaatiota. Asteet ja läheisyydet olivat jakautuneet tasaisesti eri evolutionaarisille tasoille. Kudosverkosta saatiin 203 immunomin ryväskudosta, joilla on samankaltainen geenien ilmentyminen. Kudosverkon kudosten asteilla ja geeniverkon geenien lukumäärien välillä oli vahva korrelaatio.

Johtopäätökset: Työssä luotiin geeni- ja kudosverkot, joista löydettiin sisäisiä rakenteita sekä yleisiä geeni- ja kudosryväksiä. Tutkimuksen päätavoite saavutettiin löytämällä 547 immunomin ryväsgeeniä. Kudosverkosta löydettiin 203 immunomin ryväskudosta. Immunomin ryväsgeeneillä ja proteiini-proteiini-vuorovaikutuksilla tai geenien evolutionaarisella iällä ei havaittu korrelaatiota.

Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
CBIL	Center for Bioinformatics Controlled Vocabularies
CD	Clusters of Differentiation
CSV	Comma Separated Value
EST	Expressed Sequence Tag
HIV	Human Immunodeficiency Virus
HPA	Human Protein Atlas
HPRD	Human Protein Reference Database
IRIS	Immunogenetic Related Information Source
MHC	Major Histocompatibility Complex
NK	Natural Killer
ORF	Open Reading Frame
PPI	Protein-protein Interaction
RAG	Recombination Activating Gene
SNP	Single Nucleotide Polymorphism
TCR	T Cell Receptor
TF	Tissue Factor
WAS	Wiskott-Aldrich Syndrome

Table of contents:

1. Introduction

1.1. The human immune system

The immune system is a complex network of cells, tissues and organs that protect organisms against different kinds of foreign molecules and pathogens. It works by identifying and removing various types of viruses, bacteria, parasitic worms etc. Immune system is one of the most complex biological systems, because it has to be able to recognize and attack against so many different types of pathogens which in addition to variation evolve with the time. Immune system is able to sort body's own tissues from pathogenic intruders.

There are two types of immune responses; innate and adaptive, which work together to form active and efficient immune response (Chaplin, 2006). Innate immune responses include antimicrobial peptides, phagocytes and the alternative complement pathway. They are activated instantly after infection and prevent the replication of the infecting pathogen. Innate immune responses do not alter on repeated exposure to an infectious agent, because they are encoded in the germline genes of the host. Adaptive immune responses participate by clonal selection and expansion of lymphocytes. It takes from three to five days for a sufficient number of clones to be produced and to differentiate into effector cells. Adaptive immune response is highly specific for a particular pathogen and it improves each time it faces the same pathogen. Immune responses are produced primarily by leukocytes of which there are several different types.

Immune responses take place in the cells which are organized into tissues and organs, which are called the lymphoid system. The lymphoid system consists of lymphocytes, accessory cells (macrophages and antigen-presenting cells) and in some tissues, epithelial cells. It works in capsulated organs or diffuse lymphoid tissues. The major lymphoid organs and tissues are classified as either primary (central) or secondary (peripheral). Primary lymphoid organs are the major sites of lymphopoiesis (lymphocyte development). Lymphocytes differentiate in primary lymphoid organs from lymphoid stem cells, proliferate, and mature into functional cells. T cells mature in the thymus and

B cells in the fetal liver and adult bone marrow in mammals. Lymphocytes obtain their repertoire of specific antigen receptors to fight against the antigens they meet in their life. The cells are tolerant to autoantigens and recognize only non-self antigens in the periphery. Secondary lymphoid organs include spleen, lymph nodes and mucosa-associated tissues, including the tonsils and Peyer`s patches of the gut. Lymphocytes interact with each other, with accessory cells, and with antigens in secondary lymphoid tissues, which spread the immune response. Immune responses in secondary lymphoid tissues call for phagocytic macrophages, antigen-presenting cells, and mature T and B cells (Male et al., 2006).

The immune system defends the body against foreign pathogens, but the original role of immunity was not to fight infections. This role it has adopted during evolution (Rinkevich, 2004). One of the original roles of immunity has been mate selection in jawed vertebrates. The extreme diversity of the Major histocompatibility complex (MHC) genes is the result of mating preferences. Many studies show that humans tend to choose MHC-dissimilar mates by odor (Havlicek and Roberts, 2008). Interestingly, mate selection does not seem to aim at maximum MHC-dissimilarity, but to optimal dissimilarity. Heterozygosity at the MHC may enhance the immunity in progeny and function to avoid inbreeding.

RAG (Recombination activating gene) transposition caused the structure of immunoglobulin, which lead to the development of the adaptive immune system in the jawed vertebrates soon after their evolutionary differentiation from jawless vertebrates (Agrawal et al., 1998). This change in immunity function may sometimes make it work incorrectly (Rinkevich, 2004). Overactivity of the immune system causes autoimmune diseases, such as Diabetes mellitus type 1 and Addison's disease, where the immune system is attacking against tissues of the body. In immunodeficiency, immune system has decreased or totally absent ability to fight pathogens. Immunodeficiency patients are more vulnerable to the pathogens. Primary immmunodeficiencies usually result from genetic mutations. Acquired immunodeficiencies are results of malnutrition, aging or medications (for example chemotherapy, immunosuppressive drugs). Many diseases attack the immune system. Cancers involved with bone marrow and blood cells (leukemia, lymphoma, multiple myeloma) cause immunodeficiency. HIV (human immunodeficiency virus) attacks the immune system and in the last stage of virus infection causes AIDS (acquired immunodeficiency syndrome).

1.1.1. Systems biology of immune system

Traditional reductionistic biology, practiced by many generations of scientists before us, has focused on identifying individual genes, proteins, and specific functions of cell. Systems biology is instead studying an organism as a network of genes, proteins and biochemical reactions. System biology concentrates on all the components and the interactions between them as one system. These interactions ultimately form the life in cells.

Aderem (Aderem, 2004) urges for revealing immunomic relations by the methods of systems biology. Traditional approaches are reductionistic and do not consider the complexity of the immune system. By using new computational approaches we can understand these complex biological processes.

Many methods of systems biology have been used to study immune system (Louzoun, 2007). One of the most extreme studies is the creation of the computational immune system that behaves analogous to the natural immune system (Forrest and Beauchemin, 2007). It is possible to study with this artificial "immune system" how the immune system works and get new data of the natural immune system. It is much easier to perform experiments on the *in silico* model than on living system. The artificial "immune system" also can be used to solve practical engineering problems such as computer security (Forrest and Beauchemin, 2007).

Cohen ponders the problem of modelling the immune system (Cohen, 2007). Characteristic behaviours of living organisms are emergent properties where the whole is more than the sum of its parts. Information is dynamic: gene activation, the proteome, signalling pathways, enzymatic pathways, replication and death. Biologic systems never rest; everything is on the move. The same applies to the immune system and causes requirements for ideal models.

1.1.2. Immunome

Immunome means the genes and proteins involved in the immune system. The immunome can be defined in various ways.

One earlier attempt to define human immunome is Immunogenetic Related Information Source (IRIS), which has 1562 immune genes (Kelley et al., 2005). Kelley et al. defined the immune gene as a complete gene that produces a functional transcript and demonstrates at least one defense characteristic (Kelley et al., 2005):

- Known or putative function in innate or adaptive immunity
- Participates in the development or maturation of immune system components
- Induced by immunomodulators
- Encodes a protein expressed primarily in immune tissues
- Participates in an immune pathway that results in the expression of defense molecules
- Produces a protein that interacts directly with pathogens or their products

Immunome database (Ortutay et al., 2007A) defines the immunome in a different way and includes 847 genes. These genes and their corresponding proteins were collected from research articles, textbooks and electronic information sources. The focus was on genes and proteins that are directly involved in immunological processes. In addition to clearly defined groups, such as clusters of differentiation (CD) molecules, chemokines, and their receptors, other essential genes were included. The genes that were undoubtedly needed for immunology were included. Immunodeficiency genes were taken from the ImmunoDeficiency Resource (Samarghitean et al., 2007) and IDbases (Piirilä et al., 2006). Proteins that are expressed in nearly all cells were excluded, although their function is needed also in immunity related cells and tissues. Only full-length genes were included; thus, the gene segments of immunoglobulins, B and T cell receptors and MHCs were excluded. In the case of signalling molecules, only those involved in immunity-related cascades were included.

To analyze the emergence of immunological processes, they studied the appearance and accumulation of genes in the evolutionary levels (Ortutay et al., 2007B). Three types of ortholog distributions were identified. These results indicate that most proteins in the human immune system have orthologs only in other mammals. These genes and proteins are mainly involved in adaptive immunity. It seems that these three types of orthologs arise from three different evolutionary routes.

In the first type, 15-30 % of the proteins have orthologs from Eukaryota stage (Ortutay et al., 2007B). Generally after Chordata, a few new genes appear on every level till taxon Mammalia, where almost all of the proteins have orthologs. This is the most general type of distribution: examples include antigen presentation, exogenous antigen and chemokine activity. The original proteins were not related to immune functions, but modified their functions and became involved in defense mechanisms (Ortutay et al., 2007B).

In the second type of distribution, orthologs emerge at a certain level, and then, in one or two levels, almost all of the proteins have orthologs. This type includes T cell activation and T cell receptor complex. These proteins may have resulted from the molecular appearance of the mammals. These proteins mostly arose with the appearance of mammals and represent relatively young defence strategies.

In the third type, the number of proteins with orthologs rises from level to level without a well-defined jump. This type is represented e.g. by integrin complex and blood coagulation. These proteins represent classical processes in which gradually more and more proteins became involved.

Vertebrates, the four largest ortholog groups are complement activation, alternative pathway, integrin complex, integrin-mediated signalling pathway and blood coagulation. In all the analyzed groups, 60 % of the human genes have orthologs. Thus we can assume that these functions already existed when vertebrates appeared (Ortutay et al., 2007A).

Hutton et al. defined the mouse immunome having 360 genes (Hutton et al., 2004). They used 8638 element microarray and probed with mRNA prepared from 65 normal adult and fetal tissues. At the first stage, they selected genes that were more highly expressed in one or more of 6 immunome tissues (lymph nodes from normal and antigen stimulated mice, thymus, activated T cells, spleen, peripheral blood mononuclear cells). At the second stage, they eliminated from 680 genes those with 2-fold or greater expression in brain, spinal cord, heart, kidney, pancreas or stomach, because they do not play role in the immune response. At the third stage, resulting 483 genes were examined by hierarchical cluster analysis. Immune genes were restricted to the ones expressing two-fold or greater in at least one of stimulated or unstimulated lymph nodes, activated T cells, or thymus, in order to exclude proteins of immature erythroid cells and polynuclear leukocytes encoded in spleen and peripheral blood mononuclear cells.

Hutton et al. defined genes that were highly expressed based on their normalized expression (Hutton et al., 2004) being at least 4 times higher in an individual immune tissue relative to their normal expression. They examined highly expressed mouse genes and their human orthologs for the presence of clusters of tissue factor (TF) binding sites, with the additional constraint that at least one of the cis elements present in the cluster was a lymphoid element. The numbers of mouse genes in different tissues were 17 in activated T cells, 7 in thymus and 4 in stimulated lymph node.

1.1.3. Genes of immunome

Current (May 2009) number of immunome genes in Immunome database is 893. In the time the expression data was collected for immunome genes the number was 847. There are so many genes in the immunome that it is not possible to present them all here. Instead just two samples were picked to represent immunome genes: *WAS* and *TNFRSF9*.

WAS (Wiskott-Aldrich syndrome) gene, also known as *THC*, *IMD2* and *WASP*, is located in chromosome X in Xp11.4-p11.21 (Maglott et al., 2005). Mutations in the *WAS* gene results faulty actin polymeration and cause Wiskott-Aldrich syndrome and X-linked thrombocytopenia (XLT) (Imai et al., 2003), which are primary immunodeficiency disorders. Wiskott-Aldrich syndrome is associated with combined immunodeficiency, thrombocytopenia, small platelets, eczema and increased susceptibility to autoimmune disorders and cancers.

Mutation in the different parts of *WAS* cause varying defects. *WAS* mutation may cause defects in NK (natural killer) cells and this may cause the disease (Orange et al., 2002). *WAS* mutation confuses actin polymeration, which in turn disturb TCR endocytosis (McGavin et al., 2001). *WAS* mutation causes impaired formation in the structure of phagocytic cup (Tsuboi et al., 2007).

TNFRSF9 (tumor necrosis factor receptor superfamily, member 9) gene, aliases *ILA*, *4-1BB*, *CD137*, *CDW137* and *MGC2172*, is located in the first chromosome at 1p36 (Maglott et al., 2005). TNFRSF9 protein is a member of the TNF-receptor superfamily. This receptor contributes to the clonal expansion, survival, and development of T cells.

TNFRSF9 regulates the proliferation and survival of CD8^{+} T cells (Laderach et al., 2002, Kim et al., 2008). It has been observed that levels of TNFRSF9 correlate with the rheumatoid arthritis symptoms (Jung et al., 2003). *TNFRSF9* gene might play also a role in interaction among human brain cells (Reali et al., 2003). Neurons, astrocytes and microglia of human brain express *TNFRSF9*.

Genes of the immunome can be divided into nine functional categories (Ortutay and Vihinen, 2007A). Genes can belong to more than one category. Two of the biggest categories are "CD molecules" with 292 genes and "Chemokines and their receptors" with 243 genes. Other remarkable categories by immune system function are "Inflammation" (131 genes), "Adaptive immunity" (103 genes) and "Innate immunity" (100 genes).

1.1.4. Microarray expression data

Microarray is a high-throughput method which works by utilizing the ability of the given probe mRNA to hybridize to the polynucleotide template target from which it originated. It is possible to determine the expression levels of thousands of genes within a cell in a single experiment by measuring the amount of mRNA bound to the array. The hybridization of the target to the probe is usually detected with fluorescence-based detection. Microarrays can be used to measure changes in expression levels or to detect single nucleotide polymorphisms (SNPs). The first microarrays for gene expression profiling were by Schena et al. (Schena et al., 1995). They made differential expression measurements of 45 *Arabidopsis* genes. The first whole genome microarray experiment was done on yeast (Lashkari et al., 1997). Microarrays were used to examine gene expression in yeast grown under a variety of different conditions: heat shock, cold shock, steady-state galactose and glucose.

Churchill divides two-color microarray experiments into three layers (Churchill, 2002). The top layer is biological variation, which is affected by genetic and environmental factors, and also whether samples are pooled or individual. The middle layer is technical variation during extraction, labelling and hybridization. The bottom layer is the measurement error in reading the fluorescent signals, which can be result of dust on the array. According to van Bakel et al., the problem of evaluating microarray technology reliability is that there is no single microarray technology, but it is mixture of many different techniques (van Bakel and Holstege, 2004). Two parameters of data quality are accuracy, which refers to how close a measurement is to a real value, and precision, which describes how often the measurement gives the same result. Usually with microarray experiments, the focus is on precision, as in Churchill's article. van Bakel and Holstege are trying to focus on accuracy in their work by using external RNA controls (van Bakel and Holstege, 2004).

1.2. The world of networks

Natural and social sciences have been separated from each other having different principles and tools for their studies. This separation has only increased during the last centuries when different fields have fragmented gradually to small isolated islets. The general systems theory (von Bertalanffy, 1950) was one of the first attempts to unify different sciences and find similar properties between different fields. Evolved interdisciplinary network theory developed this idea to practice. The network theory can be applied to many fields of natural and social sciences. In the future, some other properties arising from network theory might bring natural and social sciences closer to each other and even some type of universal systems theory might be found to unify the basis of all the sciences.

A network is a system where nodes or vertices are connected to each other by edges or links. Network models can be classified to three main groups: random (Erdős and Rényi, 1960), small world (Milgram, 1967) and scale-free (Barabási and Albert, 1999). Different networks have the same topological features. Degree (Shaw, 1954) simply shows how many edges a node has to other nodes. Degree distribution P(k), gives the probability that a node has k links. For P(k) the number of nodes with edges is counted and divided by the total number of nodes. Degree centrality shows the effect that a node has on the network. Closeness (Freeman, 1979) centrality of a node shows the centrality of a node based on how close it is to other nodes in the network. Nodes with high closeness have the small total distance to other nodes. The distance between two nodes is the length of the shortest path between them. The closeness centrality for a node is calculated by the inverted sum of distances from other nodes in the network. The topological features of networks, which have been developed in physics, can also be used to generate a model of how the function of a cell is organized (Barabási and Oltvai, 2004).

Random networks have edges distributed randomly, which means quite evenly, through the network, so each node has about the same number of edges. A random network is obtained by setting *n* nodes and adding edges between them at random. The node degrees follow a Poisson distribution, in other words most nodes have roughly the same number of edges. These random networks are purely theoretical models.

Small world means that everything/everybody is connected with each other. Stanley Milgram´s famous "small world" experiment revealed that any person is connected to any other through a short chain of social ties, the average chain length being six people (Milgram, 1967). Most people have heard the phrase “Six degrees of separation”, but actually Milgram himself did not use it, it has become established later. Systems are organized into small world structures, because it is efficient in transforming information, for example infectious diseases spread more easily in small world networks than in regular lattices (Watts and Strogatz, 1998).

Another useful property that shows up from networks is the robustness of scale-free networks, which means that scale-free networks display surprisingly high degree of tolerance against random failures. Although key components regularly malfunction, local failures rarely lead to the loss of the global information-carrying ability of the network. The error tolerance comes at the expense of attack survivability: the diameter of these networks increases rapidly, and they break into many isolated fragments when the most connected nodes are targeted (Albert et al., 2000). Fortuna and Melian (Fortuna and Melian, 2007) showed that scale-free regulatory network allows a larger active network size than random ones by compiled the network of software packages with regulatory interactions (dependences and conflicts) from Debian GNU/Linux operating system. They suggested that this result might have implications for the number of expressed genes at steady state. Small genomes with scale-free regulatory topologies could allow much more expression than large genomes with exponential topologies. This may have implications for the dynamics, robustness and evolution of genomes.

Social ties which form social networks are helping in job hunting, as most of the jobs are found through personal contacts than by the application. The strength of interpersonal ties varies from a person who we meet once a year or less to very close friends and family members. Interestingly the weak ties are stronger in transforming information, because those to whom we are weakly tied are more likely to move in circles different from what we are (Granovetter, 1973).

In scale-free networks, some nodes have only one or few edges, while some have many. These important nodes having many edges to other nodes and thus having high degree are called hubs. Scale-free networks have the probability P(k) that a vertex in the network interacts with k other vertices decays as a power in law, following $P(k) \sim k^{-\gamma}$, where γ is the degree exponent. The value γ determines many properties of the system. The smaller the value γ, the more important the role of the hubs is in the network (Barabási and Albert, 1999). Most of the existing social, technological and biological networks in the world are scale-free networks. Just to give a brief demonstration of these different networks, there are some interesting examples following. Dekker studied the Eurovision song contest as a friendship network, how countries casted their votes to other countries and formed blocks (Dekker, 2007). An example of technological network is the study of transportation system of the subway and buses in Boston and the network they form (Latora and Marchiori, 2002). An example of biological network is the gene-interaction network created to find out genes associated to prostate cancer (Özgűr et al., 2008). They find out that highest degree, eigenvector, closeness and betweenness genes in the gene-interaction network were most likely to be related with the disease. There are also many attempts to capture a part of human protein-protein interaction (PPI) networks in order to model the function of the body. An example of this kind of network is by Ewing at al. using mass spectrometry for finding new PPIs (Ewing et al., 2007). There are also some specialized PPI networks, for example the network of human inherited ataxia-causing proteins (Lim et al., 2006).

1.2.1. Communities in networks

Communities in networks have groups of nodes that are connected to each other with more edges than the rest of the network. Random graphs do not have a community structure. Many real world small world and scale-free networks have community structure. For example the biggest community groups Santa Fe Institute scientist collaboration network has are Structure of RNA, Statistical Physics, Mathematical Ecology and Agent-based Models (Girvan and Newman, 2002). This study revealed that scientists are grouped together by similar research topic or method.

Modularity (Newman, 2004) is a property of a network and a specific proposed division of that network into communities. In high modularity network there are many edges within communities and only a few between them. Modularity is a measure of the quality of a particular division of a network. The modularity value is between 0 and 1. If the number of within-community edges is random, we will get 0. Values approaching 1 indicate networks with the strong community structure (Newman and Girvan, 2004).

Divisive methods are relatively little studied. They start with the network of interest and attempt to find least similar connected pairs of vertices and then remove the edges between them. By doing this repeatedly the network is divided into smaller and smaller components, and the process can be stopped at any stage for taking the components at that stage to be the network communities. Difficulty of the algorithm is the relatively high computational demand (Newman and Girvan, 2004). One of the divisive methods is called edge betweenness community, which tries to find the edges that are most "between" communities. Communities are exposed gradually when these edges are removed one by one. The edge betweenness community algorithm first calculates the betweenness for all the edges in the network and then removes the edge with the highest betweenness. Next it recalculates betweenness for all edges affected by the removal and removes the edge with the highest betweenness. It repeats these calculating and removing steps until no edges remain. The speed of the algorithm is rather slow, which makes it impractical for large networks (Girvan and Newman, 2002).

Fast greedy community analysis has another approach for finding community structures. It is a hierarchical agglomeration algorithm, which works by greedily optimizing modularity. The general idea in optimizing modularity is to repeatedly join together two communities whose amalgamation produces the largest increase in modularity. This method is considerably faster than most previous general algorithms and can be used for very large networks as well (Clauset et al., 2004).

2. Objectives

The main objective of this study was to identify immunome gene groups with similar tissue expression pattern using gene networks with immunome gene expression data and various net analysis tools. Parts of this objective were:

- collecting immunome expression data
- finding gene clusters in the immunome gene network
- doing ontology analysis on the clusters
- seeing if there is any correlation in degree and closeness values between the immunome gene and the protein interaction network. If so, which genes have high values in both networks?
- seeing the correlation in degree and closeness values with the evolutionary age of genes. Are the central genes in this network more ancient?
- finding tissue clusters in the immunome tissue network
- finding out the correlation between degree of these tissues in the immunome tissue network and the number of genes expressed in them in the gene expression data
- finding out the most important tissues (with high degree or/and closeness)

3. Materials and methods

3.1. The stages of this thesis

The ultimate goal of this work was to find out the gene expression pattern of the human immunome.

1. The work was done with the expression data collection from HPA and SOURCE databases for the genes of the immunome. The matching anatomy ontology terms for the tissues of expression were also collected to the table. These first steps were done with Microsoft Excel. The data in these tables was transformed to R (R Development Core Team, 2005), because the work continued with R.
2. The next stage was to generate immunome gene networks with genes as nodes and tissues as edges. Immunome gene networks were created from expression data from the SOURCE database, the HPA database and their unified data. Two community analyses, edge betweenness and fast greedy, were used to find community structures of the network data. These two community analyses were used to collect common gene clusters. Top 20 genes were also collected by their high degree and closeness. Data on the gene cluster data table was compared to the immunome PPI data and the evolutionary age of the immunome genes (Ortutay and Vihinen, 2008).
3. Tissue networks, where tissues were nodes and genes were edges, were generated similarly to gene networks. Community analyses were done the same way as for gene data and gave the tissue clusters. The achieved information of degree of tissues was combined with the original expression data of the number of genes in that tissue to see the correlation. Top 20 tissues were collected by their high degree and closeness.

Figure 1 depicts of the steps of this thesis.

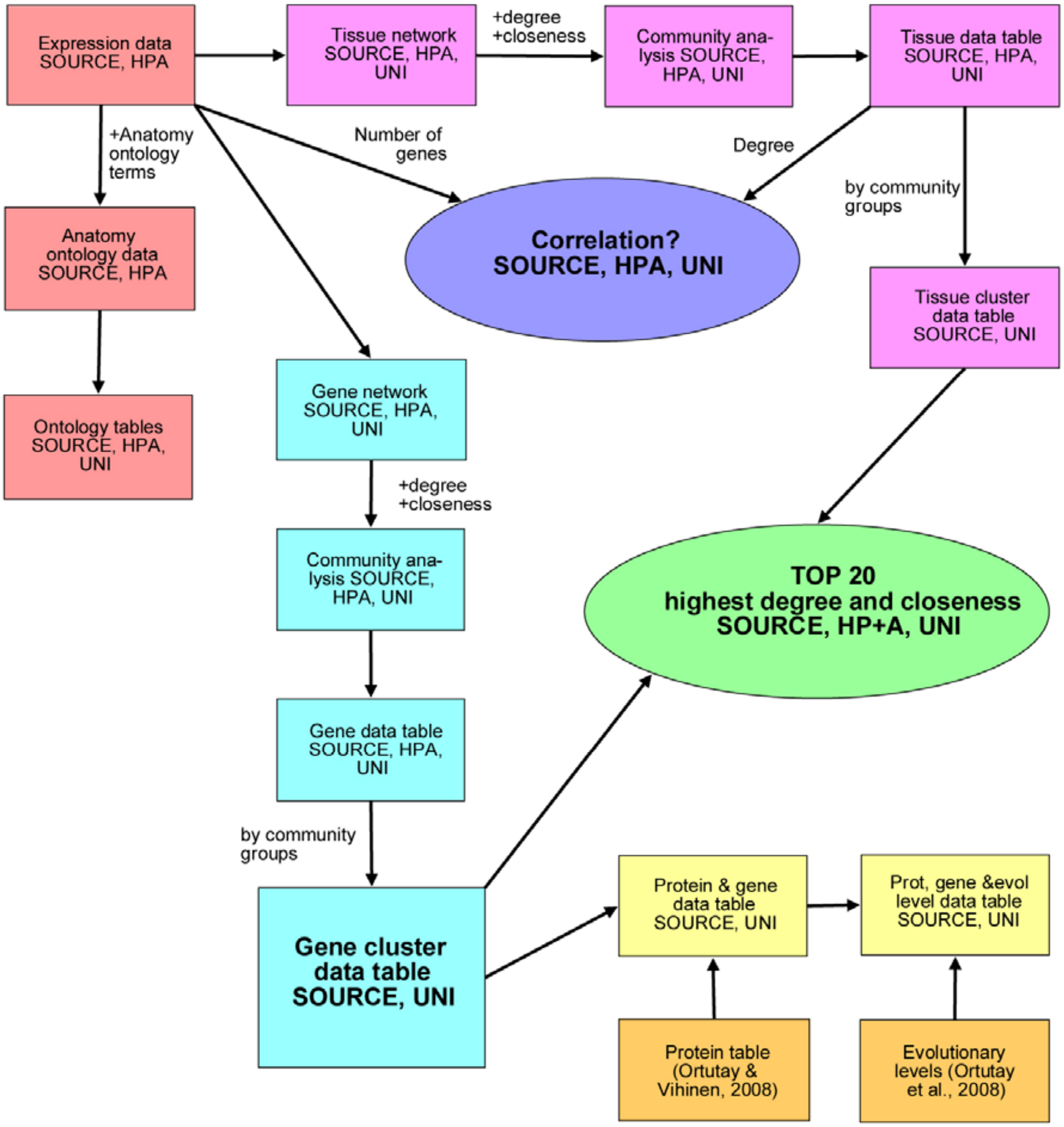
Expression data SOURCE, HPA
Tissue network SOURCE, HPA, UNI
+degree +closeness
Community analysis SOURCE, HPA, UNI
Tissue data table SOURCE, HPA, UNI
+Anatomy ontology terms
Number of genes
Degree
by community groups
Anatomy ontology data SOURCE, HPA
Correlation? SOURCE, HPA, UNI
Tissue cluster data table SOURCE, UNI
Ontology tables SOURCE, HPA, UNI
Gene network SOURCE, HPA, UNI
+degree +closeness
Community analysis SOURCE, HPA, UNI
TOP 20 highest degree and closeness SOURCE, HP+A, UNI
Gene data table SOURCE, HPA, UNI
by community groups
Gene cluster data table SOURCE, UNI
Protein & gene data table SOURCE, UNI
Prot, gene &evol level data table SOURCE, UNI
Protein table (Ortutay & Vihinen, 2008)
Evolutionary levels (Ortutay et al., 2008)

Figure1: Diagram of the study. Salmon color in the up left shows the original expression data collected for immunome genes and the added data of anatomy ontology terms. Light turquoise color shows the gene network analysis: creation of gene networks, community analyses and collecting immunome gene clusters. Light yellow color shows gene cluster data which was added immunome PPI and evolutionary levels data. Pink color shows tissue network analysis. Blue color in the middle checks the correlation between degree of a tissue in tissue network and the number of genes in that tissue in gene network. Bright green color shows the top 20 genes and tissue collected from gene and tissue networks.

3.2. Materials

3.2.1. Expression data for immunome genes

There were 847 genes in human immunome (Ortutay and Vihinen, 2007A) in immunome database. The expression data for these immunome proteins and genes was collected from Human Protein Atlas (HPA) (Table 1) (Uhlen et al., 2005) and SOURCE (Table 2) (Diehn et al., 2003) database, between January and March 2007. HPA database had expression data for only 175 proteins. SOURCE database had expression data for all 847 immunome genes. The expression data from these databases was collected manually, because there was not any ready database with the needed expression data or a script for collecting the data easily.

HPA expression data is based on antibody proteomics. Affinity purified antibodies are used for protein profiling in various tissues and cell types assembled in tissue microarrays. Human Protein Resource (HPR) center, which is located in Stockholm and Uppsala, Sweden, runs this program. The used Atlas version was 2.0, which was updated 30th Oct 2006. At that point Atlas had 1514 antibodies and 1,238,760 images of tissues/cell types.

SOURCE is a unification database which collects data from many databases, including the genetics and molecular biology of genes from the genomes of *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*. Gene expression data of SOURCE is collected from UniGene (Wheeler et al., 2008), Swiss-Prot (The UniProt Consortium, 2009), GeneMap99 (Deloukas et al., 1998), Rhdb (Rodriguez-Tomé and Lijnzaad, 2001) and LocusLink (Wheeler et al., 2003). SOURCE is provided by the Genetics Department, Stanford University.

The PPI data for immunome proteins were collected from Human Protein Reference Database (HPRD) (Peri et al., 2004), which collects information about human proteins. It includes PPIs, post-translational modifications, enzyme-substrate relationships and disease associations. Information in HPRD is collected manually from published literature by expert biologists and by bioinformatics analyses of the protein sequence

Table 1: An example of the immunome gene expression data collected from SOURCE database. On the left column are tissue types in the order of how many immunome genes are expressed in them. Next column presents the number of expressed genes in that tissue. In the following columns are number of genes by gene groups. The last columns give the names of genes that are expressed in the tissue (this table does not show them all, because this is a piece of the original table).

Tissue	Number of genes	CD molecules	Chemokines and receptors	Complement system	Transcription factors	Humoral immunity	Cellular immunity	Phagosytosis	Inflammation	Antigen prosessing and presenting	GENES		
Breast cancer	31	12	12			6	5		3		*CD2*	*IL2RA*	*CD28*
Pooled	31	9	18	1		10	5		11		*CD3D*	*TNFSF5*	*CD48*
leukocyte	28	15	4			4	8		2		*PSTPIP1*	*CD3E*	*CD7*
Lymphocyte	28	13	17	1	2	7	4	1	3		*CD6*	*ITGAX*	*IL2RA*
Spleen	25	5	9	2	1	5	2		4		*CD7*	*CR2*	*FCER2*
myeloid cells, 18 pooled CML cases	22	4	2	2		1	5		7		*MS4A3*	*CEACAM8*	*C5R1*
thymus	20	11	4		2	5	4	2	2	5	*CD1A*	*CD1B*	*CD1C*
Leukopheresis	19	7	4		1	4	3			1	*ITGAL*	*ITGAX*	*MS4A1*
spleen	19	12	6			2	4	1	6	1	*CD5L*	*CD8A*	*ITGB2*

Table 2: An example of the extensive expression data collected from the HPA database. Columns are for tissues and cell types. S in cell type means strong staining and M medium staining. On the left of the table there are gene categories and then numbers of expressed genes in the cell type of the tissue.

Tissue		Adrenal gland				Appendix			
		Cortical cells S	Cortical cells M	Medullar cells S	Medullar cells M	Glandular cells S	Glandular cells M	Lymphoid tissue S	Lymphoid tissue M
Gene category	CD molecules	3	18	0	0	6	12	22	14
	Chemokines and reseptors	4	8	1	0	1	10	3	6
	Complement system	1	1	0	0	1	2	2	3
	Transcription factors	2	2	0	0	1	2	1	5
	Humoral immunity	0	2	0	0	0	1	3	6
	Cellular immunity	1	5	0	0	2	1	5	2
	Phagosytosis	0	1	0	0	0	1	2	1
	Inflammation	2	5	0	0	0	7	1	8
	Antigen	0	1	0	0	1	1	2	2
GENES		*NCAM1*	*CD9*	*IL1RAPL1*	*NP*	*ITGB1*	*CD9*	*CD3E*	*CD2*
		LAMP2	*ITGAM*		*G6PD*	*CEACAM5*	*TNFRSF8*	*CD4*	*ITGAM*
		ABCB1	*CD14*			*CEACAM5*		*CD6*	*CD22*

3.2.2. R and Igraph

R (R Development Core Team, 2005) was used for creating immunome gene and tissue networks and performing the community analysis on them, and finding common immunome gene clusters. R was used for comparing the gene network data to PPI and evolutionary level data. R was used for making figures. The used R algorithms are written as pseudo code. Anyone interested to use the original R codes, can contact the author.

R is a complete system with language, statistical computation and graphics. The R version used was 2.6.0. Additional libraries are available for a variety of specific purposes.

Igraph library version 0.4.4. was used for this work (Csárdi, 2008). Igraph is a tool for graph and network analysis. It makes possible to handle large graphs and to generate random and regular graphs, visualize graphs etc.

3.2.3. CBIL (Center for Bioinformatics Controlled Vocabularies)

CBIL shows the vocabulary of anatomy terms hierarchically for tissues. The vocabulary bases on anatomy terms taken from the Mouse Gene Expression Database at the Jackson Laboratory (Smith et al., 2007). It has been expanded with human anatomy and modified in many areas, especially the haematolymphoid system, based on the 37th edition of Gray's Anatomy (Williams et al., 1996), and the brain, by the contributions of Dr. Jonathan Nissanov of Drexel University. Each anatomy term has been mapped onto the relevant set of Expressed Sequence Tag (EST) libraries in dbEST, a division of GenBank that contains sequence data and other information (Boguski et al., 1993) to increase the reliability of the data. CBIL vocabulary, last updated February 07, 2005, was used on 17th September 2007 in http://www.cbil.upenn.edu/anatomy.php3

3.3. Preliminary work

3.3.1. Anatomy ontology files

The matching anatomy ontology terms for the tissues collected from SOURCE expression data and HPA expression data were picked from CBIL vocabulary to unify unclear names of tissues.

The anatomy ontology term tables were formatted in Microsoft Office Excel. Most specific anatomy category by tissue and cell type was added to data from HPA database. The same was done to data from SOURCE database, although it turned out to be more problematic because SOURCE data had many cancer tissues and tissues from other diseases which did not have a match in anatomy terms. All the disease linked tissues were removed. Tables were changed to Comma Separated Values (CSV) format which is a table file format for storage of data. One line in the CSV file corresponds to a row in the table. Within a line, columns are separated by commas. CSV files are often used for moving tabular data between different computer programs. In this case CSV was used to move data from Microsoft Office Excel to R. The final SOURCE data table had tissue and ontology terms and HPA data table contained tissue, cell type and ontology terms.

3.3.2. Ontology tables

The anatomy ontology files for tissues were used to produce ontology tables for data from SOURCE and HPA databases. Anatomy ontology terms were added to SOURCE and HPA expression data tables. These tables were then merged to get a unified table.

The SOURCE ontology table had tissue, gene and ontology term. The HPA ontology table had tissue, cell type, gene and anatomy ontology term information. The unified data table of SOURCE and HPA data had only gene and ontology.

Algorithm:

- Open gene data and ontology term tables for SOURCE and HPA
- Put in a table gene expression data
- Add ontology terms to a new column in a table

- Merge SOURCE and HPA data to get unified table

3.3.3. Reformatting gene expression data for R

The collected expression data from HPA and SOURCE databases were formatted with Microsoft Office Excel to simpler form, by removing extra data, to facilitate R analyses. The reformatted HPA data has 3 columns: Tissue, Cell type and Gene. The data shows genes and in which tissues and cell types they are expressed. The reformatted SOURCE data has in the similar way two columns: Tissue and Gene. Both files were saved in CSV format for the following R analyses.

3.4. Gene network analysis

3.4.1. Generating the gene network

In this project the gene network was created purely for statistical analysis. Hence the immunome gene network is artificial and does not exist in the cells or have any common function. The gene network enables finding clusters of genes which are expressed in the same tissues.

The gene expression data was modified to CSV format which R interprets. Immunome gene networks were generated of data from SOURCE database, HPA database and their unified data.

Algorithm:

- Load igraph library
- Open the data file
- Create an empty graph
- Add genes from file as vertices
- Add tissues from file as edges

HPA expression data needed an extra step compared to SOURCE expression data, when tissues and cell types of HPA were combined. This was done because the HPA expression information cannot be recognized only by the tissue nor by the cell types; the same tissue can have many cell types, and also the same cell type can be present in many tissues. In the following analyses, HPA and SOURCE expression data must have the same number of columns for creating their unified data.

3.4.2. Adding degree and closeness to the vertices

Degree (Shaw, 1954) of a node is the number of edges it has in the network. Closeness (Freeman, 1978) centrality measures how many steps are required to access every other node from a given node. The closeness centrality of a node is defined by the inverse of the average length of the shortest paths to/from other nodes in the graph.

Information about degree and closeness of nodes was added for each gene, because they are needed in the further analysis.

Algorithm:

- Simplify the graph by removing nodes´ loops to themselves and multiple edges
- Count degree and closeness for the nodes
- Add the degree and closeness as attributes to the nodes

Immunome gene networks were simplified by removing its edges to itself, which is a sensible, because otherwise they would interfere with the results. Multiple edges between nodes, which appear when two or more genes are expressed in the same tissue, were also removed. It would have been sensible to keep these edges, as they represent the real biological phenomenon, but community analyses did not work with multiple loops, so they had to be removed.

3.4.3. Community analysis

Two different types of community analyses were used to uncover the community structures of the gene networks. Edge betweenness community is a divisive method, which finds the edges that are most "between" communities and their removal one by one reveals the communities. Fast greedy community is a hierarchical agglomeration algorithm, which reveals communities by greedily optimizing modularity. In this study, community analyses were used to reveal the gene clusters which are expressed in the same tissue.

Community to membership methods can be performed varying number of times, which are called steps. Different numbers of steps in community to membership was tested to find out the maximum number of communities. The maximum number was set to loop which counted the ideal modularity (Newman and Girvan, 2004), which is the point where modularity is at its maximum. Modularity presents the division of the network into communities. The modularity value is between 0 and 1. The bigger the value of modularity, the clearer is the community structure of the network. The values of maximum modularities were collected to a table.

Edge betweenness community (Girvan and Newman, 2002) and fast greedy community (Clauset et al., 2004) analyses were performed in the similar manner.

Algorithm:

- Calculate edge betweenness/fast greedy communities
- Create memberships to communities
- Calculate the maximum modularity
- Set the maximum modularity to community to membership
- Set edge betweenness/fast greedy values as vertex attributes

3.4.4. Gene data tables

All the data accumulated from the immunome gene network analyses was stored to the tables. Information about of different node attributes: degree, closeness, edge betweenness community groups and fast greedy community groups was collected to these tables.

3.4.5. Finding common clusters from edge betweenness community groups and fast greedy community groups

The information about edge betweenness community and fast greedy community analyses for gene networks of SOURCE expression data, HPA expression data and their unified data was collected to gene data tables. The next step was to find the common gene clusters of these two community analyses. It was not possible to do gene cluster analysis for HPA data, because community analysis methods were not able to find any community structures from HPA data. Edge betweenness community gave no community groups for HPA data and fast greedy community gave only one big group to which a gene either belonged or not. Finding common clusters was done only for SOURCE and unified data. In the cluster finding algorithm, were first collected all the possible gene pairs from inside each edge betweenness community group. It was than checked if these pairs appeared together in some of the fast greedy groups. The created gene cluster tables included information about common genes and their degree, closeness edge betweenness community groups and fast greedy community groups.

Algorithm:

- Put in a table gene data
- Add column numbers and use them in the following steps instead of genes
- Make pairs of all the combinations of edge betweenness group genes
- Collect pairs of edge betweenness to a table
- Collect the pairs which appear together in some fast greedy group to a table
- Remove duplicates from the table
- Make a new table without gene numbers

3.4.6. Correlation between gene network and PPI network

Previously collected immunome PPI data (Ortutay and Vihinen, 2008) were used to find out if there existed a correlation between immunome gene and PPI network. Immunome PPI data had information about protein vulnerability, closeness and degree.

Immunome proteins used in PPI data derive also from the same Immunome database (Ortutay et al., 2007A) as the genes. PPIs were collected from the Human Protein Reference Database (HPRD) (Peri, 2004). Since only interactions between the immunome proteins were taken into account, no new nodes were added, but proteins without interactions were eliminated from the dataset. The final PPI network had 584 nodes out of the 847 original proteins, forming altogether 1349 interactions (Ortutay et al., 2007A). Interactions which appeared more than once were simplified to single edges.

The vulnerability of the protein network was calculated using the efficiency characteristics of the network. The vulnerability, V_i, of a network associated with the i:th node:

$V_i = (E-E_i)/E$

where E is the global efficiency of the network without the node i and all of its interactions.

Immunome gene data and PPI data tables were combined to see whether there is the correlation between gene and PPI networks. This correlation analysis was done for SOURCE and unified gene cluster data.

Algorithm:

- Open protein and gene data tables
- Merge tables
- Change small closeness values (under $5*10^{-3}$) to NA
- Make a plot of the table

3.4.7. Evolutionary levels

Evolutionary levels (Ortutay et al., 2008) table (Table 3), was created according to the hierarchy in the NCBI taxonomy database (Wheeler et al., 2008). Evolutionary level showed the evolutionary age of the gene. The branches of the taxonomic tree were numbered from *Homo sapiens*, level 0 to Eukaryota, level 9.

Earlier collected data about the evolutionary levels of immunome genes (Ortutay et al., 2008) were combined to the data about immunome gene clusters and PPIs. The immunome evolutionary levels table had gene names and their evolutionary levels.

Table 3: Evolutionary levels.

Evolutionary level	Number
Homo sapiens	0
Mammalia	1
Amniota	2
Tetrapoda	3
Vertebrata	4
Chordata	5
Coelomata	6
Bilateria	7
Fungi/Metazoa	8
Eukaryota	9

3.4.8. List of top genes

Top 20 immunome genes were collected by their degree and closeness to find out if they have some common properties. These top immunome genes were collected from SOURCE network, HPA network, and unified data.

3.5. Tissue network analysis

3.5.1. Generating the tissue network

The tissue network differs from the gene network by having tissues as nodes and genes as edges. The generated immunome tissue networks had tissues as nodes and immunome genes as edges. Immunome tissue networks were created from the expression data from SOURCE database, HPA database and their unified data.

Algorithm:

- Open the data files SOURCE and HPA
- Create an empty graph
- Add tissues from file as vertices
- Add genes from file as edges

3.5.2. Edge betweenness and fast greedy community analyses and finding common clusters from tissue data

Edge betweenness and fast greedy community analyses were performed the same way as the immunome gene network

The immunome tissue data table was created the same way as the immunome gene data table. The tissue data table included data of degree, closeness, edge betweenness community groups and fast greedy community groups.

Finding common immunome tissue groups from edge betweenness and fast greedy community groups was done the same way in pairs (Chapter 3.4.5) as with immunome gene groups.

3.5.3. Correlation between degree of the tissue and the number of genes

The correlation between degree of the tissue in the immunome tissue network and how many immunome genes were expressed in that tissue was checked next. These two values should be near each other, but it is worthwhile to find out if this is true.

The number of genes in the original expression data was counted and then combined to data of degrees of the tissues from immunome tissue networks. The newly created tables had data of tissue, number of genes and degree. This correlation analysis was done for data from SOURCE and HPA database.

3.5.4. List of top tissues

Top 20 immunome tissues were collected, similar to immunome genes, by their degree and closeness. This was done in order to find out if these tissues have common features. Top immunome tissues were collected from SOURCE tissue network, HPA tissue network and their unified data.

4. Results

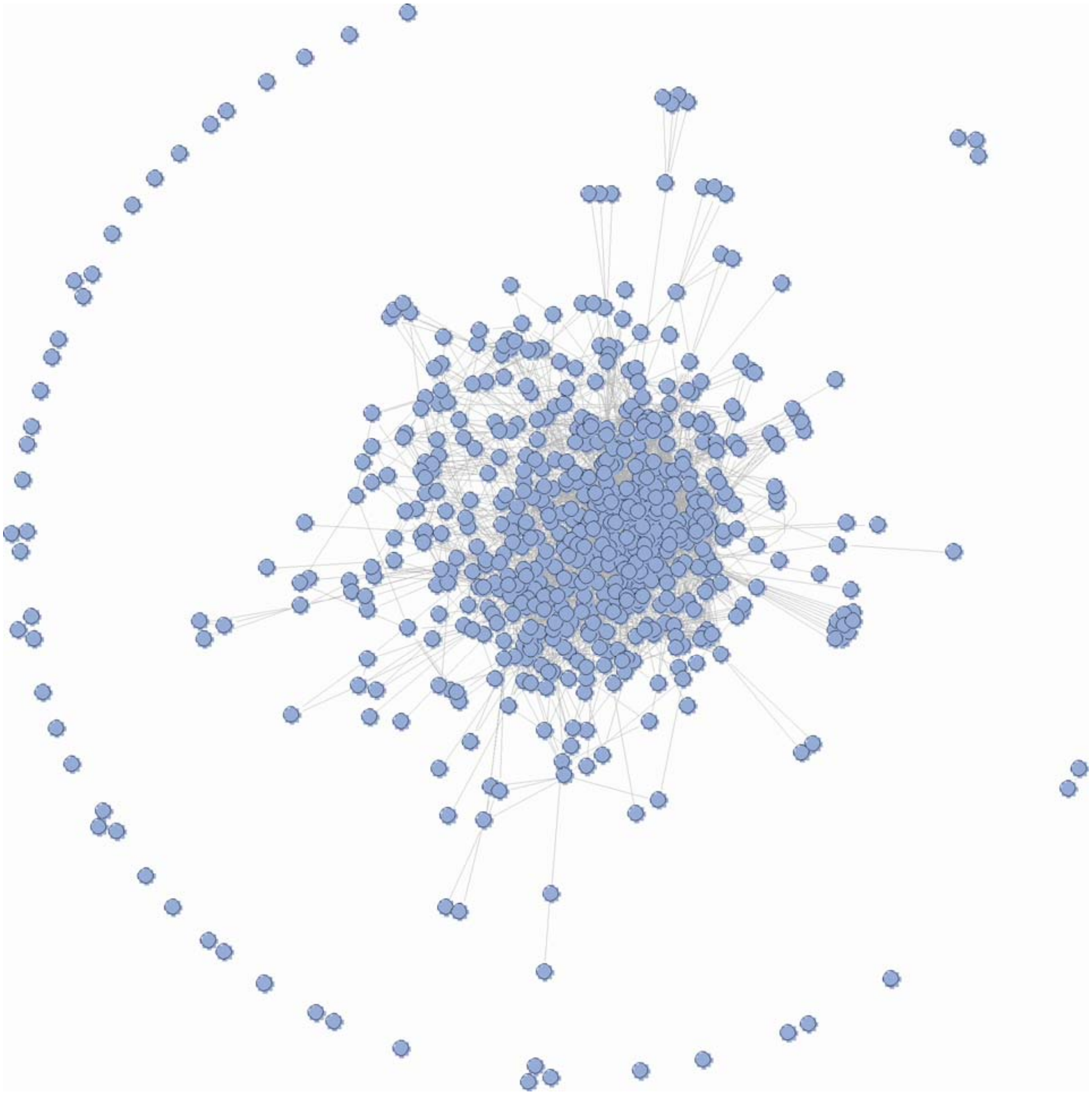

Figure 2: Immunome gene network generated from expression data collected from SOURCE database. Immunome genes are nodes (spots) and tissues are the connecting edges (links). The edges have different lengths just for pleasing the eye and revealing the community structure of the network. A large cluster of immunome genes is shown in the middle of the picture. The circle around the cluster is formed by individual immunome genes which are not expressed in any tissues (by SOURCE expression data) and thus do not have edges.

4.1. Gene network analysis

Immunome gene network analysis was done to find out common immunome gene clusters that have similar gene expression. First immunome gene networks were created. Next two community analyses were done on the gene networks and common clusters were found from their community groups. These collected immunome gene clusters were the main results of this study, because their characteristics can be studied further. The immunome gene cluster results were compared to immunome the PPI results and the information about evolutionary age of the genes to see if there is correlation. Gene network analysis yielded degree and closeness values for genes and enabled collecting top immunome genes by their degree and closeness.

4.1.1. Gene networks

Immunome gene networks have immunome genes as nodes and tissues that they are expressed as edges (Figure 2). Gene networks were done using data from SOURCE database, HPA database and their unified expression data. Gene networks enabled the finding of common gene expression patterns out of large expression data.

4.1.2. Values of maximum modularities

Maximum modularities, expressing how networks are divided to communities, were collected as the byproduct of edge betweenness and fast greedy community analyses. The higher modularity values with immunome gene network data from SOURCE stand for higher division to communities than with gene network data from HPA (Table 4).

Table 4: Maximum modularities in edge betweenness and fast greedy communities of gene network data from SOURCE database, HPA database and unified data.

Community analysis	SOURCE	HPA	UNIFIED
Edge betweenness community	0.50	0.0011	0.25
Fast greedy community	0.47	0.044	0.28

4.1.3. Gene data

Degree, closeness, edge betweenness and fast greedy community groups for the immunome genes was revealed from the collected immunome data. The edge betweenness and fast greedy community analyses revealed the community structures of gene networks from SOURCE expression data and unified data, and the lack of community structure in the gene network from HPA expression data. Figure 3 present the division of groups inside these two community analyses.

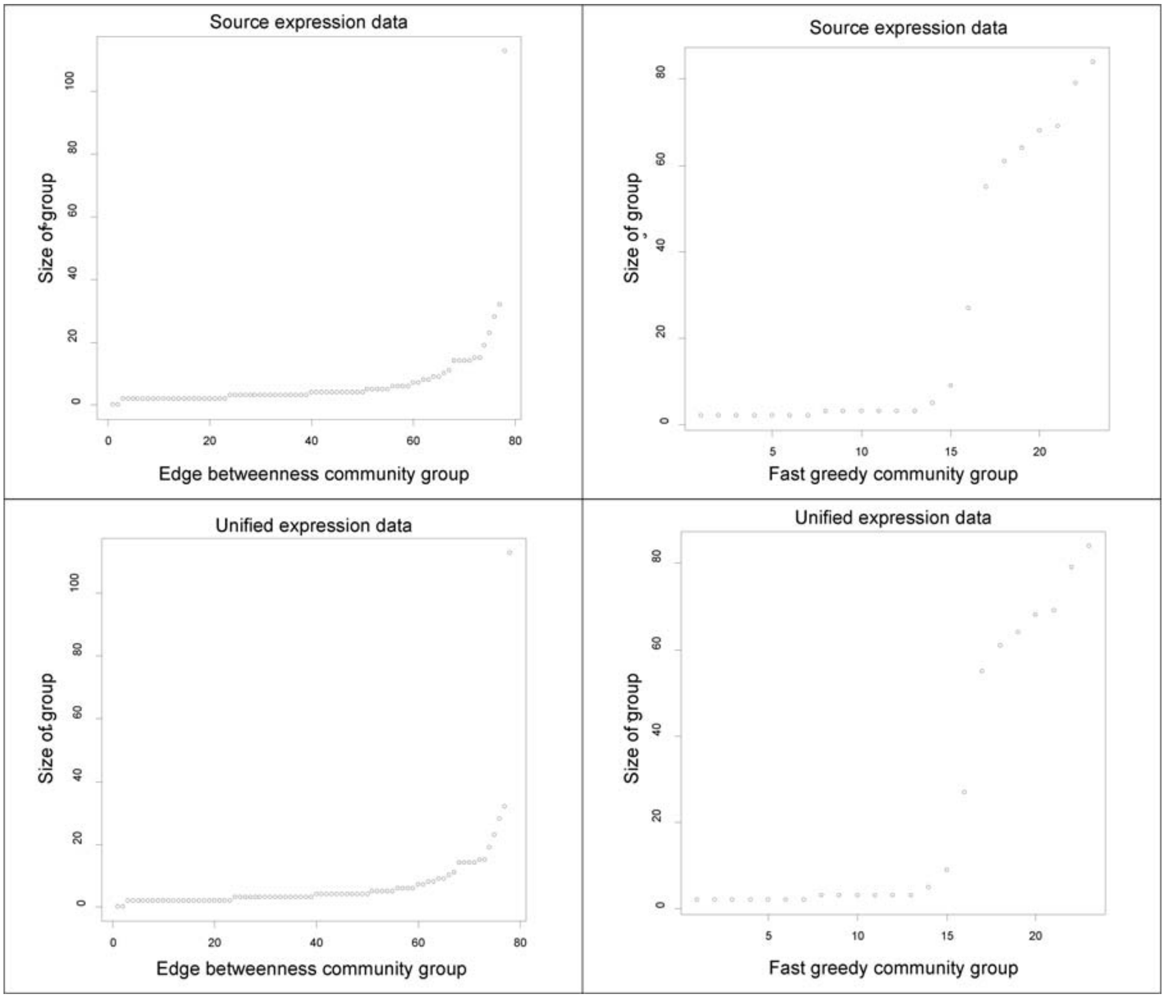

Figure 3: Division of immunome genes in immunome gene network to edge betweenness and fast greedy community groups. Each spot represent a community group in gene network.

4.1.4. Common gene clusters from edge betweenness groups and fast greedy groups

Common immunome gene clusters were found by searching for gene pairs that appear together in the same community group in both edge betweenness and fast greedy community analyses. This method revealed 547 clustered genes in data from SOURCE database and 566 clustered genes in unified data, thus only about 300 genes were eliminated this way. The immunome clustered genes were divided into 88 gene clusters (Table 5) with sizes varying from 2 to 32 genes. Genes belonging to the same cluster have similar gene expression patterns which can be looked at more detailed. For example, cluster number 8 (Table 5) has 15 genes, which are expressed mostly in leukopheresis and lymph node.

Table 5: 88 immunome gene clusters from the gene network created from SOURCE expression data.

Cluster	Immunome clustered genes
1	C5R1 CAMP CD300E CEACAM8 CLEC12A CLEC4A CLEC4D CLEC5A CTSG CXCL3 DEFA1 DEFA4 EPX FCAR FCER1A FCRL5 IFIT1L IGSF2 IL18RAP IRAK2 LILRA3 LTFLY86 LYZ MPO MS4A3 PFC PGLYRP1 PIK3CG PRG2 RHAG TRAF6
2	CARD15 CCL2 CCL20 CCL4 CCL4L2 CCL5 CCL7 CCR4 CCRL2 CD160 CD3D CD48 CD69 CD83 CLEC7A CSF3R FCER1G IFNG IL17F IL2 IL22 IL4 IL8 KLRB1 KLRC1 KLRD1 LCP2 TNFSF5
3	ACE C2 C4BPA C5 C6 C8A C8B C8G CCRN4L CD209L CFHR1 CFHR2 CFHR5 CRP CXCL2 HAMP IL13 IL5RA MBL2 RFXAP RNASE7 TNFSF11TNFSF4
4	BST1 C1QTNF5 CD3Z CD7 CD72 CMTM3 CXCL6 CXCR3 FCER2 IKBKG IL11RA IL12A IL15RA KLRK1 LCK LTB4R PLA2G7 PTPRCAP SOCS1 TBX21 TNFRSF13B TRADD WAS
5	CCR3 CCR6 CD244 CD300LB CD300LF CD33 CD3E CD5 CD84 CMRF-35H GZMB GZMK ICAM2 IL2RB LAIR1 LILRB1 LY9 NCR1 NCR3 PDCD1 SIGLEC5 ZAP70
6	BATF CCL22 CCL3 CD28 CD6 CD86 IL10 IL2RA KIR3DL1 KLRC2 PSTPIP1 SLAMF1 SOCS3 STAT4 TNFRSF4 TNFRSF9 TNFSF8 XCL1
7	BLR1 CCL11 CCL3L1 CCL3L3 CCR7 CISH CXCL11 GNLY IL7R IRAK4 ITGAX JAK3 LAX1 PPBP RAG2 TNFRSF8
8	ADAM8 CCR5 CD164L2 CR1 CRADD CSF1R CSF2RB GP5 HLA-DOA IL28RA ITGA2B ITGAL MHC2TA MS4A1 TLR9
9	CASP10 CCL15 CD38 CLEC10A CMTM8 DEFB1 FGFR2 GZMA IL18R1 IL1RAPL2 MASP1 MME NFATC1 SELE TNFSF6
10	CCL19 CCL25 CD1A CD1B CD1C CD1E CD2 CD209 CD8B1 IL21R ITGA4 PRSS16 RAG1 TCF7
11	CHUK IL17D IL1R1 ITGB3 JAK2 KDR PDGFRA PDGFRB PLXNC1 SH2D1A TCF8 TLR3 TNFSF10 TNFSF7

12	C1QBP CD248 CD34 FCGR3A FCGR3B GUSB HLA-C IL11 IL12RB1 LIFR MARCO NBS1 SELP WASF3
13	C1QG CLEC4E CMRF35 CSF2 CXCL5 ICOS ICSBP1 IL1F9 IL23A IL6 PLAUR PTPNS1 TNF TRAF1
14	AIRE CASP2 DEFB119 DEFB123 GP1BA HRH4 IL12B IL17 IL21 IL26 IL3 IL5 IL9 XCR1
15	CCL21 CD53 CD79B CD80 CTLA4 CXCL13 FOXP3 HLA-DOB HLA-DQB2 HLA-DRA IL24 NCF1 TNFRSF13C TNFRSF17
16	CASP8 CCR2 CMKLR1 CXCR6 CYSLTR1 GYPE IL18BP LAG3 RIPK1 SDC1 SLAMF6 TFRC TLR5
17	ABCB1 CCL18 CCR1 CD109 CD59 GZMM PLAA PSME2 TNFSF13B
18	CD200R2 DEFB106A FCAMR IL27 IL28A IL29 LILRA5 NCR2 PPIA
19	CCL14 CD302 CD47 FADD IL17RB KIT NP TNFRSF18
20	IGLL1 IL1A IL1RAP ITGB4 PILRA PSIP1 PSMB8 RFXANK
21	CCL24 CD97 IL8RA ITGB2 LILRA6 LILRB3 LTA SELL
22	BANK1 CXCL10 CXCL9 IGJ IL7 IRF1 POU2AF1 THBD
23	ANPEP CEACAM5 IL10RB NOS2A PSMF1 SDFR1 TRAF5
24	CD226 CD3G IL1R2 LY64 PTPN22 PTPRC PTPRJ
25	ERGIC2 IL17RE IL1RL2 IL31RA TLR1 YWHAZ
26	CASP1 CD58 HLA-DRB4 IL9R LTB PAFAH2
27	CCL16 CRLF1 CXCL12 KEL MASP2 TIRAP
28	C1QTNF3 IL18 IL1RN ISGF3G PSMB6 S100A8
29	CEACAM3 IL22RA2 PSG1 SEMA7A SIGLEC6
30	AICDA CD79A CXCR4 NCF4 SPN
31	CD9 HLA-DQB1 IFI27 MAPK14 MYLK
32	IL12RB2 SIVA STAT2 TNFRSF10B TNFRSF1B
33	CD320 IFITM1 PTDSR TLR2 TNFRSF6
34	HRH2 MBP SOCS6 SOCS7
35	CSF2RA IGSF8 IL17RD PROCR
36	ALK GYPA GYPB RHD
37	CCL1 CCL27 IL17C MPL
38	APS CEACAM1 CX3CL1 CYBA
39	IRAK1BP1 ITGA2 TRAF3 WASF1
40	CHL1 COLEC12 EBF FY
41	CLEC4C GP9 ICAM1 IL1RL1
42	HLA-DMB IL10RA RFX1 TCN2
43	CCBP2 CD274 IL17B LAIR2

44	C1QTNF7 CD200R1 LPO PROM1
45	CCL23 MS4A5 PGLYRP2
46	IL6R NFATC2 PLK3
47	IGF1R IL20RA TNFRSF14
48	CD24 MIF TNFRSF1A
49	G6PD LIG4 NCAM1
50	ENC1 FOXK2 ITGAE
51	INSR MSR1 SMARCAL1
52	C1QA CYLN2 SIGIRR
53	DDR1 FGFR3 IL4R
54	C1QL3 EBF2 IL1RAPL1
55	ITFG1 PSMB9 PVRL1
56	C1QL4 CLCF1 TNFRSF12A
57	CCRL1 IRF2 LAMP3
58	CDW92 CMTM6 MX1
59	CCL28 CRLF3 ITGAV
60	CD19 DEFA6 FLT3
61	CMKOR1 SDF2L1 TRAF3IP1
62	C3AR1 IF SLAMF7
63	ENG LU TNFRSF10C
64	PDCD1LG2 TNFRSF11A
65	C4BPB CMTM2
66	CSF1 MR1
67	IL23R RHCE
68	CLU IL13RA2
69	CD22 IL8RB
70	CD4 CD74
71	CD33L3 CSF3
72	BF NDUFS3
73	IL1F7 PDGFB
74	A2ML1 IL1F6
75	EPO LYG2
76	ANP32B MRC1

77	LAMP2 TM4SF2
78	CXCL14 DCLRE1C
79	HLA-E NFATC4
80	C1RL ISG20
81	L1CAM SN
82	CD3EAP PECAM1
83	HLA-F IL27RA
84	CXCL1 MST1R
85	EBI2 PPP3R1
86	CASP7 PPP3CA
87	PLA2R1 SCARB2
88	ITGB1 TNFSF14

4.1.5. Correlation between gene network and PPI network

Correlation between immunome gene cluster data and PPI data was checked by using network features degree, closeness and vulnerability. This analysis was done for clustered genes from the gene network of SOURCE database expression data (Figure 4) and the gene network of unified data (Figure 5).

Results show that there is a very weak correlation between vulnerability, closeness and degree inside the immunome PPI data (Figure 4, Figure 5). There is a high correlation between degree and closeness values in immunome clustered genes from the gene network of SOURCE database expression data (Figure 4), as in immunome clustered genes from gene network of HPA database expression data (Figure 5). When comparing vulnerability, closeness or degree of immunome PPI data to degree or closeness values of immunome clustered genes from the gene network based on SOURCE data (Figure 4), one cannot see the clear correlation. The same is true in the immunome clustered genes from gene network of HPA data (Figure 5). In conclusion, there is no significant correlation between the immunome gene network and the PPI network.

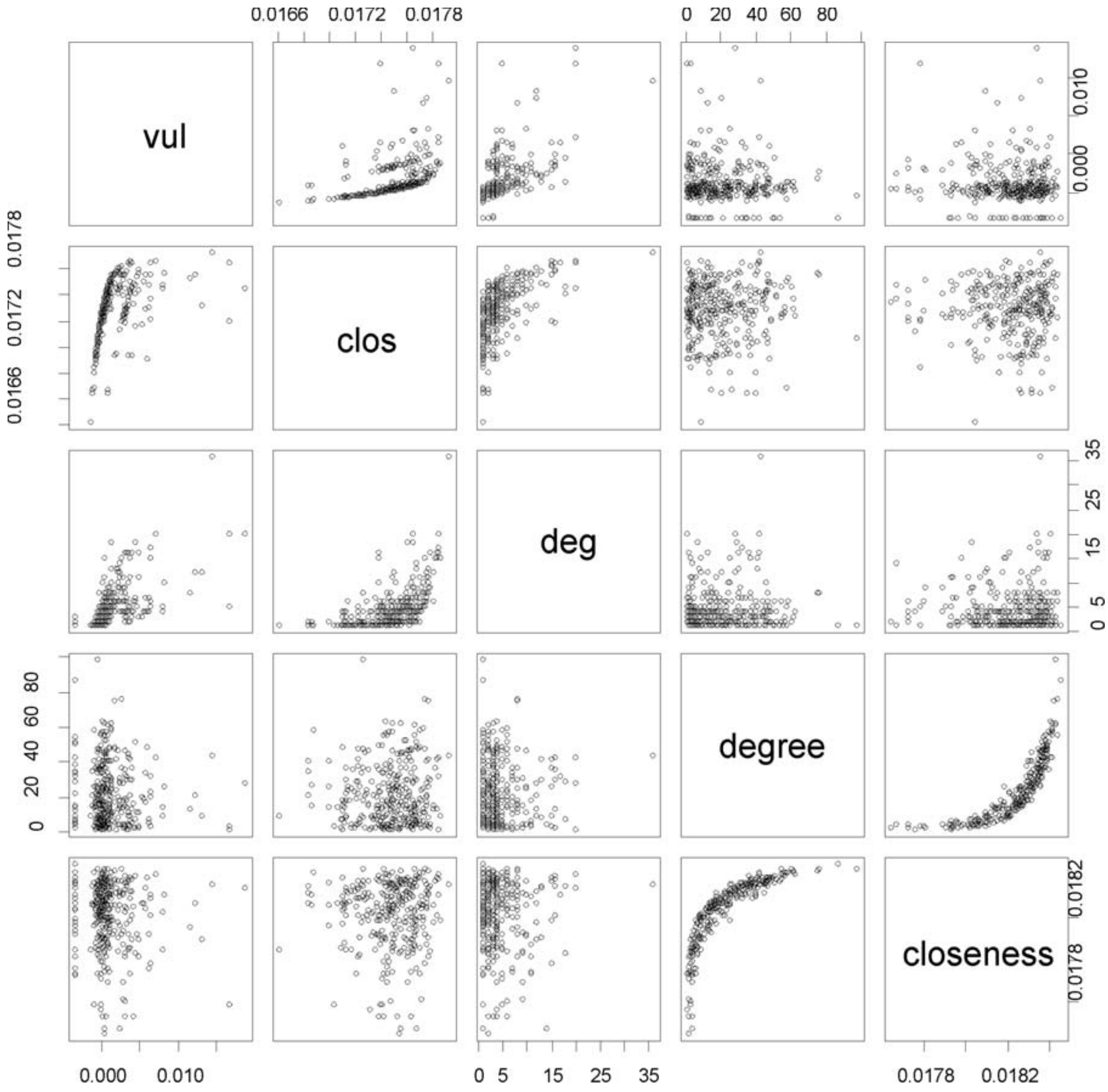

Figure 4: Correlation between immunome clustered genes from the gene network of SOURCE database expression data and PPI data. Three upper values vul (vulnerability), clos (closeness) and deg (degree) are from immunome PPI data. The degree and closeness are from immunome gene cluster data of data from SOURCE database.

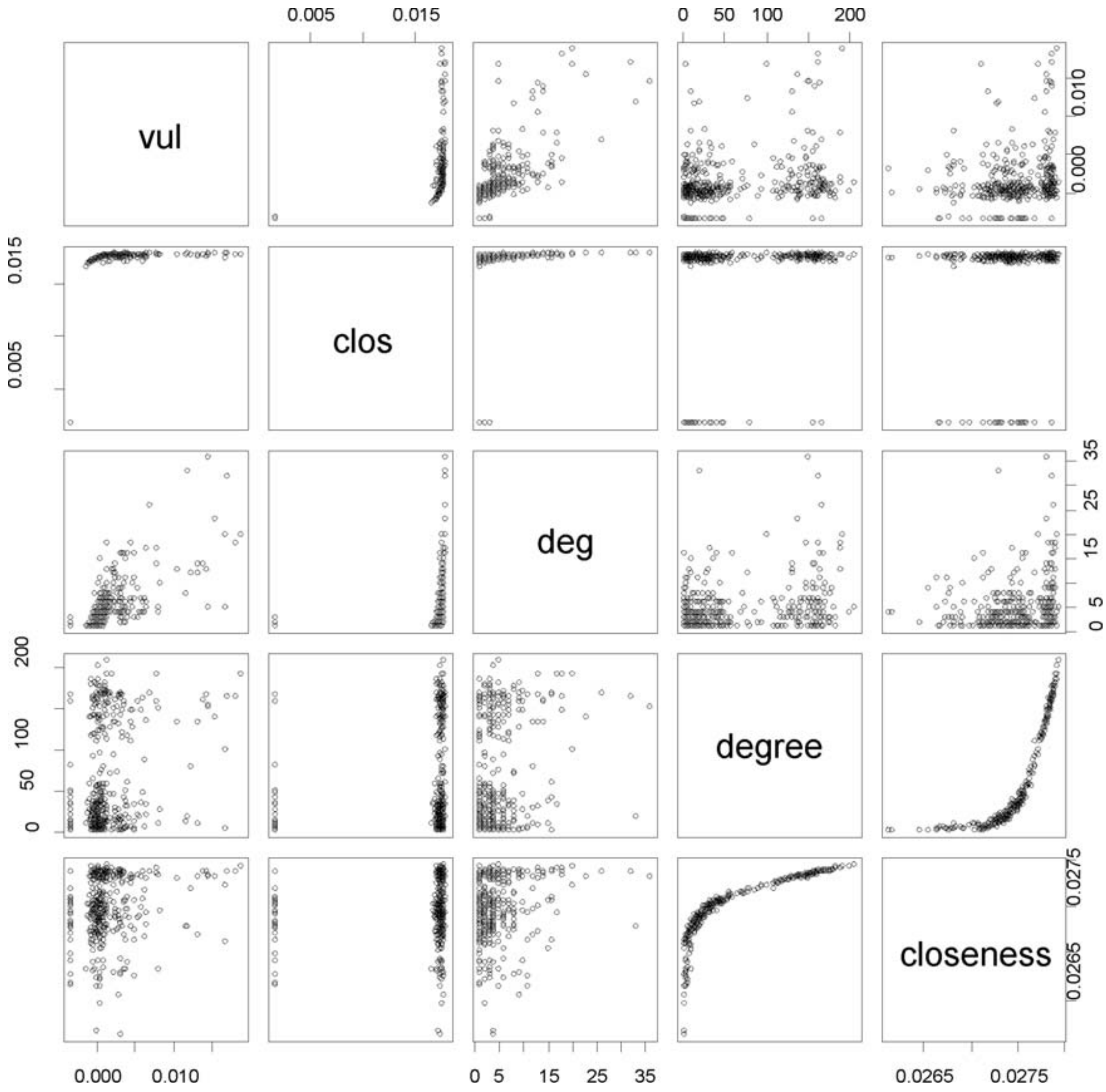

Figure 5: Correlation between immunome clustered genes from the gene network of unified expression data and PPI data. Three upper values vul (vulnerability), clos (closeness) and deg (degree) are from immunome PPI data. The degree and closeness are from immunome clustered genes from gene network of unified expression data.

4.1.6. Correlation of degree and closeness values with the evolutionary age of the genes

Information about the evolutionary age of immunome genes based on their emergence was added to gene cluster and PPI data. The evolutionary age of a gene can be expressed by evolutionary levels, with numbers ranging from 1 (*Homo sapiens)* to 9 (Eukaryota)

(Table 3). Evolutionary levels were used to find out if degree or closeness values from immunome gene network vary with the evolutionary age of the genes.

Degree and closeness values for clustered genes from SOURCE expression data are on top panels and unified data on bottom panels in Figure 6. The idea of the correlation analysis was to see if there is a trend between degree or closeness and evolution level.

The results (Figure 6) indicate that the location and importance of a gene in the gene network are independent of the evolutionary age of the gene.

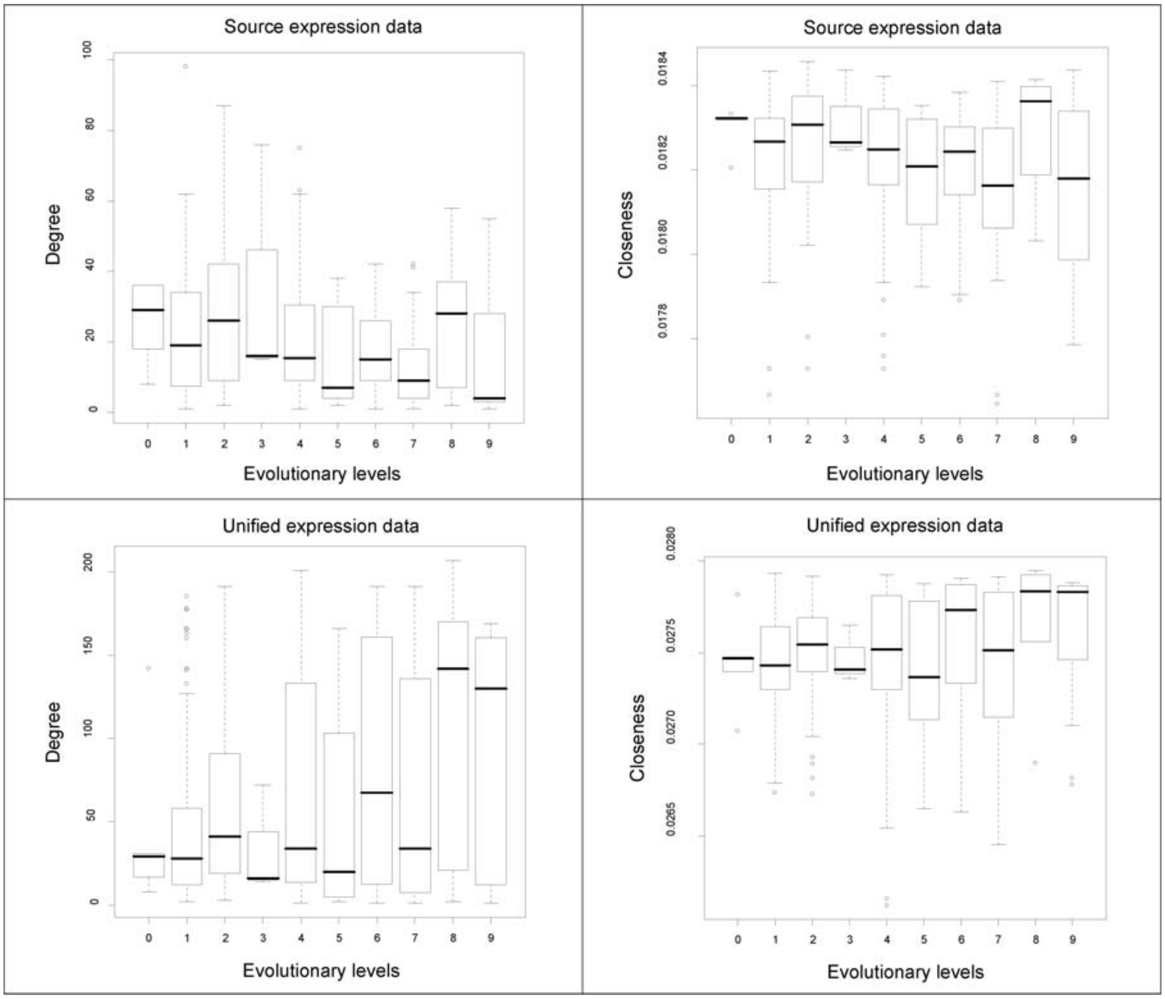

Figure 6: Evolutionary levels (Table 5) by degree and closeness express the evolutionary age of immunome genes.

4.1.7. Genes with the highest degree or closeness

Genes with the highest degree and closeness in the gene network were collected to Table 6. The genes with the high degree and closeness values are the most central in the gene network, indicating of having the most related genes with similar gene expression patterns. For example *TNFRSF9* is in highest degree in SOURCE expression data and in high degree and closeness in unified expression data. *WAS* is in highest degree and closeness in both SOURCE and unified expression data. There is a high correlation between degree and closeness (Figures 4 and 5), and thus many same genes are shared in the different columns of the table 6.

Table 6: Immunome genes with highest degree and closeness.
The table has data of genes with highest degree and closeness in gene networks of SOURCE expression data and unified data. The two genes with gray color are present in many of these highest groups.

SOURCE data		UNIFIED data	
degree	closeness	degree	closeness
EPX	*IL18RAP*	*CASP10*	*CASP10*
IL10	*XCL1*	*CASP2*	*CASP2*
TNFRSF9	*BLR1*	*CCR2*	*CCR3*
XCL1	*IL7R*	*CCR7*	*CD2AP*
BLR1	*CLEC7A*	*CD2AP*	*CHUK*
IL7R	*IL8*	*CD86*	*CXCR3*
CLEC7A	*KLRB1*	*CHUK*	*FOXP3*
KLRB1	*NCR1*	*CXCR3*	*G6PD*
LY9	*PPBP*	*FOXP3*	*IL12A*
NCR1	*KIR3DL1*	*IL12A*	*IL23A*
PPBP	*CCR5*	*IL23A*	*ITGB1*
KIR3DL1	*ICOS*	*ITGB1*	*PTDSR*
SLAMF6	*SLAMF6*	*NP*	*TNFRSF9*
LAX1	*LAX1*	*PTDSR*	*TRAF1*
GZMK	*GZMK*	*TNFRSF9*	*TRAF6*
SIGLEC5	*SIGLEC5*	*TRAF1*	*WAS*
CD28	*CD28*	*TRAF6*	*ZAP70*
IL2RA	*IL2RA*	*WAS*	
CD7	*CD7*	*ZAP70*	
WAS	*WAS*		

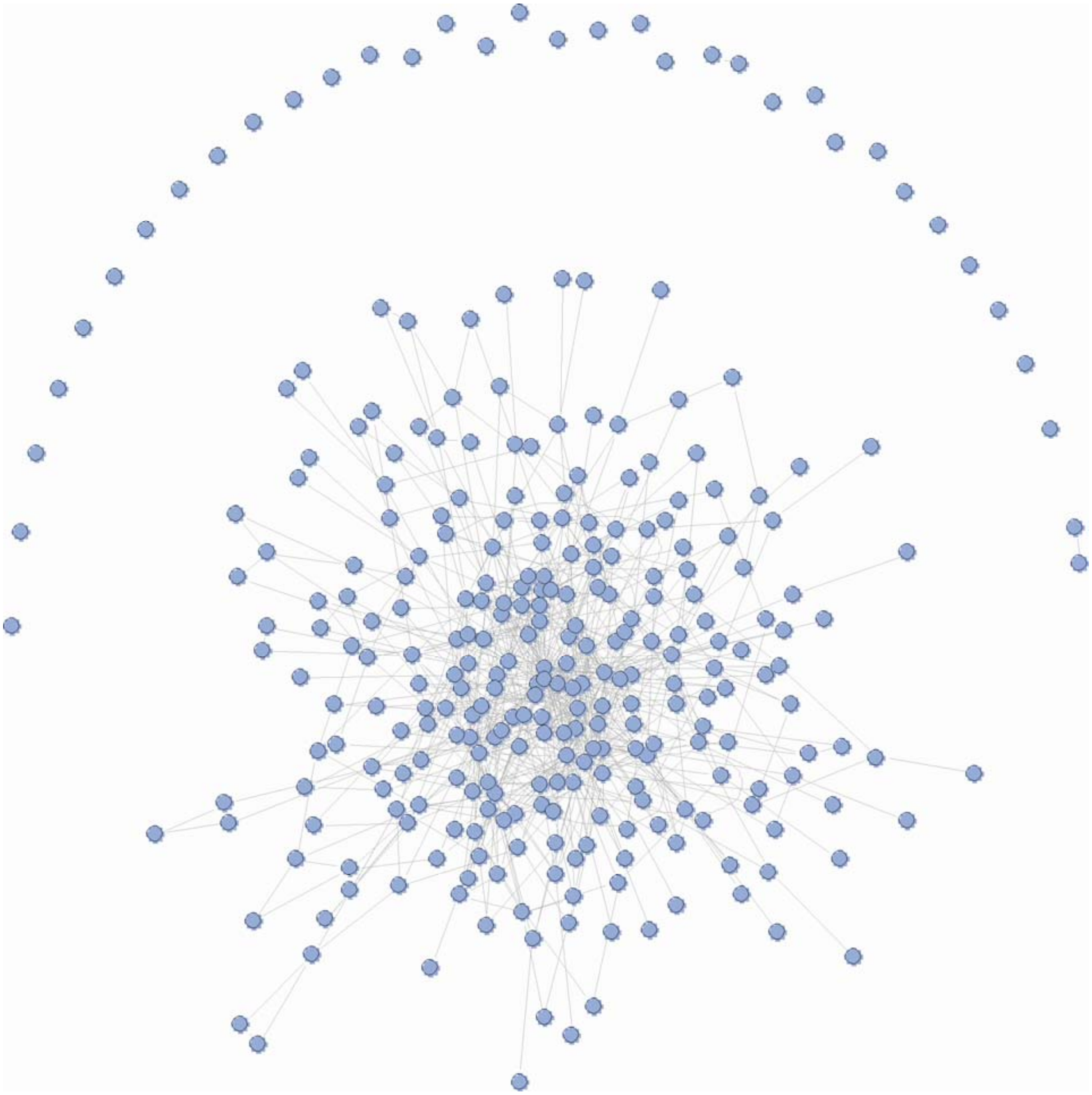

Figure 7: The tissue network of data from SOURCE database has tissues of expression as nodes and immunome genes as edges.

4.2. Tissue network analysis

The idea of the immunome tissue network analysis was to find out the central immunome tissues. First the immunome tissue network was generated and similar community analyses were performed as to the immunome gene network. The community analyses revealed tissue clusters.

4.2.1. Tissue network

Immunome tissue networks were the opposite of immunome gene networks by having tissues of expression as nodes and the immunome genes as edges (Figure 7).

4.2.2. Tissue data

Edge betweenness and fast greedy community analyses were performed to the immunome tissue network in a similar fashion as to immunome gene networks. The following results show the division into communities by the analysis of data from SOURCE and unified data (Figure 8).

4.2.3. Common tissue clusters from edge betweenness community groups and fast greedy community groups

The two community analyses found 203 immunome tissues from SOURCE expression data and 547 immunome tissues from unified data. However these results have less relevance than immunome gene clusters in this study, because important tissues in the immune system are well studied already and this result does not lead to any new experiments.

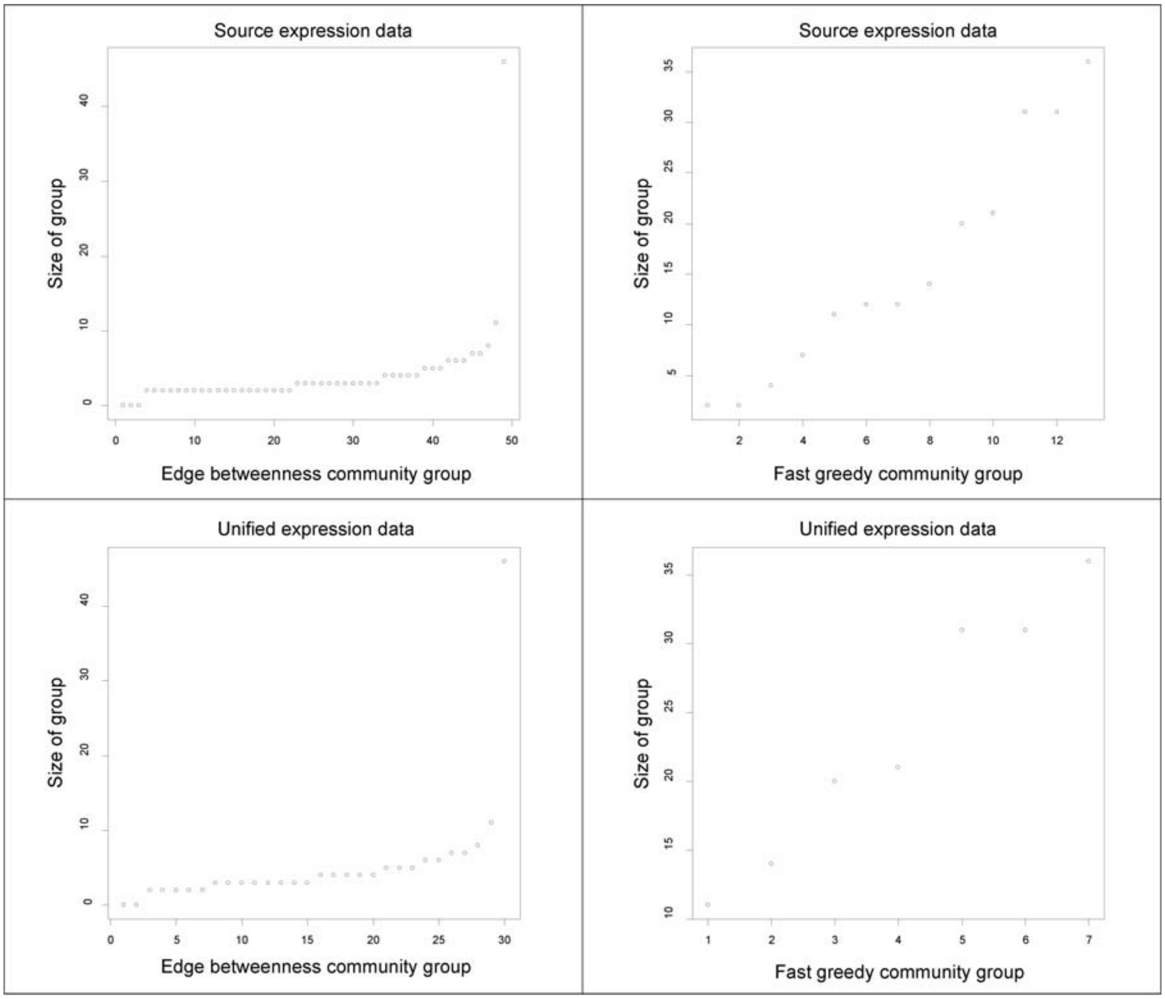

Figure 8: Division of immunome tissues in the immunome tissue network into edge betweenness and fast greedy community groups.

4.2.4. Correlation between degree of tissue and number of genes

The following plots (Figure 9) show the correlation between the degree in the immunome tissue network and the corresponding number of genes in the immunome gene network. There seems to be a high correlation between degree and number of genes in data from SOURCE database (Figure 9A), while degree and number of genes has a weaker correlation in data from HPA database (Figure 9B).

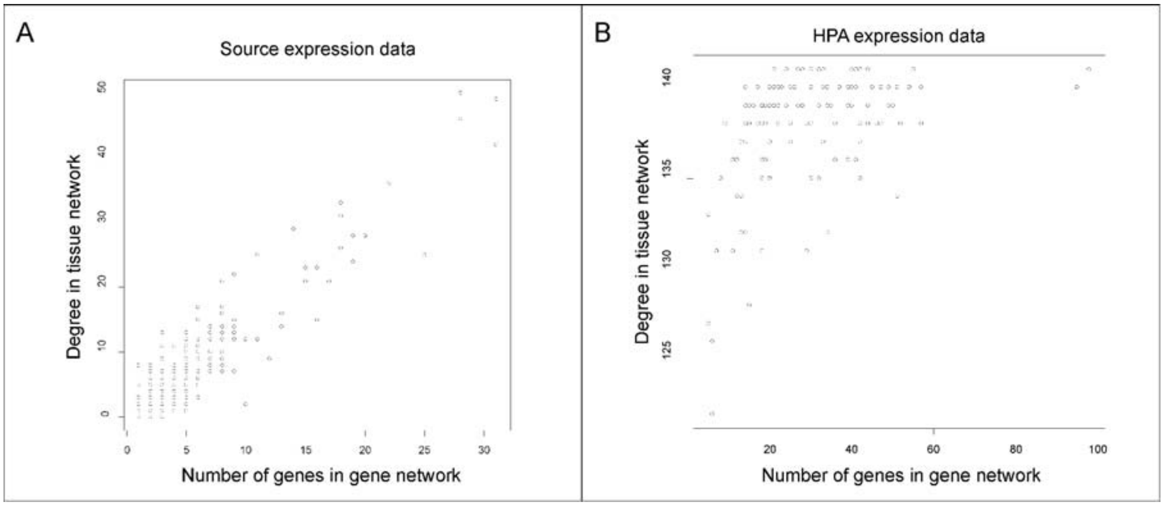

Figure 9: Picture A shows the 0.927 correlation between degree in tissue network and number of genes in the gene network in data from SOURCE database. Picture B shows the 0.347 correlation between degree in tissue network and number of genes in the gene network in data from HPA database (degrees under 100 are excluded).

4.2.5. Tissues with highest degree or closeness

Immunome tissues with highest degree and closeness were collected in the table (Table 7). Lymph, bone marrow and spleen have high degree and closeness in immunome tissue networks from SOURCE expression data and unified data.

Table 7: Immunome tissues with highest degree and closeness

SOURCE expression data		Unified expression data	
degree	closeness	degree	closeness
lymph	lymph	bronchus surface epithelial cells	colon glandular cells
lymphocyte	follicular lymphoma	colon glandular cells	rectum glandular cells
natural killer cells, cell line	lymphocyte	rectum glandular cells	stomach 2 glandular cells
spleen	natural killer cells, cell line	stomach 2 glandular cells	cervix, uterine glandular cells
alveolar macrophage	spleen	cervix, uterine glandular cells	kidney cells in tubuli

SOURCE expression data		Unified expression data	
breast cancer	alveolar macrophage	bone marrow bone marrow poetic cells	bone marrow bone marrow poetic cells
leukocyte	breast cancer	bone marrow bone marrow poetic cells	bone marrow bone marrow poetic cells
leukopheresis	leukocyte	breast glandular cells	breast glandular cells
subchondral bone	leukopheresis	duodenum glandular cell	duodenum glandular cell
thymus	thymus	epidydimis glandular cells	endometrium 2 cells in endometrial stroma/ecm
thymus	thymus	fallopian tube glandular cells	epidydimis glandular cells
pooled	pooled	lung macrophages	fallopian tube glandular cells
spleen	spleen	lymph node follicle cells (cortex)	lung macrophages
prostate	prostate	lymph node non-follicle cells	lymph node follicle cells (cortex)
corresponding non cancerous liver tissue	fetal liver	soft tissue 1 mesenchymal cells	lymph node non-follicle cells (paracortex)
synovial membrane tissue from rheumatoid	synovial membrane tissue from rheumatoid arthritis	soft tissue 2 mesenchymal cells	soft tissue 1 mesenchymal cells
bone marrow	bone marrow	spleen cells in red pulp	soft tissue 2 mesenchymal cells
myeloid cells, 18 pooled cml cases,	myeloid cells, 18 pooled cml cases, bcr/abl	tonsil follicle cells (cortex)	spleen cells in red pulp
lymphoma, follicular mixed small and large	lymphoma, follicular mixed small and large	tonsil non-follicle cells (paracortex)	tonsil follicle cells (cortex)
		urinary bladder surface epithelial cells	tonsil non-follicle cells (paracortex)
			urinary bladder surface epithelial cells

5. Discussion

Immunome gene clusters (Table 5) are the main harvest of this study (Figure 1). Information of the gene clusters was gained by immunome gene network and network analyses. Immunome clustered genes are in relevant part in the function of immune system and could be studied more.

The study started with expression data from HPA (Uhlen et al., 2005) and SOURCE (Diehn et al., 2003) databases. Unfortunately, HPA database did not hold all the immunome genes and expression data for them when this data was collected, and this limited the analysis. HPA expression data is growing all the time, and thus in the future this same analysis could be done with more immunome genes. SOURCE database contained expression data for all the immunome genes. The weakness with SOURCE expression data is that it is collected from various resources and the quality of the data can vary. Errors in microarray expression data can result from the selection of samples, as well as from technical and measuring errors (Churchill, 2002). SOURCE expression data from different sources can base on different techniques and samples (van Bakel and Holstege, 2004). SOURCE expression data cannot be considered as the final truth, it is more like directional data. SOURCE database was the widest at the time of collecting expression, because it covered all the immunome genes. Microarray expression data is increasing rapidly, and in the future this study can be done with supplemental microarray expression data. HPA expression data (Uhlen et al., 2005) is carefully checked by experts and is thus more reliable than SOURCE expression data. Reliability of the expression data depends of the quality of used antibodies too, but the biggest problem is the small size of HPA database.

Gene networks, network theory and features of networks were widely used in this study. Gene networks are scale-free networks, meaning that some nodes have only one or few edges, while some have many. This kind of network structure makes it possible to find out many features of gene network, such as degree, closeness and communities of the network. These features were helping on the way to find out the common immunome gene clusters, which are expressed in the similar way. Immunome gene networks were created from the expression data from SOURCE database, HPA database and their unified data. Modularity of a gene describes its division to communities (Newman and Girvan, 2004). The immunome gene network of SOURCE expression data had strong community structure by its modularity, while the immunome gene network of HPA expression data had a poorer community structure. The poor community structure in the immunome gene network of HPA expression data is due to the fact that genes in the gene network of HPA data are mostly expressed in the same tissues. The community structure of the unified network from SOURCE and HPA expression data is somewhere in the middle of the two individual networks. Edge betweenness and fast greedy community analyses revealed the community structures of the immunome gene networks from SOURCE expression data and unified data, and the lack of community structure in immunome gene network from HPA expression data. Edge betweenness groups of immunome gene networks from SOURCE expression and unified data are divided equally into different groups, each having less than 40 genes, with the exception of one group with over 100 genes. Fast greedy groups of immunome gene networks from SOURCE expression and unified data are divided into groups of sizes varying up to 90 genes. The reason why unified data results look the same as those of the SOURCE expression data is that the SOURCE expression data have more effect on the results than the HPA expression data. These two community analysis methods were the only ones in R that worked with this data, so it was not possible to expand the study to other methods. There are coming new features to R all the time and in the future it probably will be possible to use more community analysis methods and compare those results.

The next step was to reach the main aim: to find out immunome gene clusters. The method for searching immunome gene clusters from these differently divided community groups was to take each gene pair from every edge betweenness group and check if it appears in some fast greedy group. This method uncovered 547 clustered genes from the immunome gene network from SOURCE expression data and 566 clustered genes in the unified data, so only 300 genes were eliminated from the original 847 immunome genes. Immunome gene clusters vary in size from 2 to 32 and there are 88 clusters altogether. Genes inside the same cluster have similar gene expression patterns, in other words they are expressed in the same tissues, which is significant. These discovered immunome gene clusters could be used for further studies to find out their common properties and to research their roles in the tissues that they are commonly expressed in. Their common properties could be studied by checking if they have common ontologies. Some extra data collected from literature might illuminate why these genes appear together.

The immunome gene cluster data was compared to immunome PPI data by their degree, closeness and vulnerability. Immunome gene cluster data of SOURCE and unified data, degree and closeness correlated with each other. Inside the immunome PPI network there is no clear correlation. There was no noticeable correlation between the immunome gene network and the PPI network. Degree and closeness values of immunome clustered genes were also compared to their evolutionary age to find out if they are correlated. This analysis did not uncover any trend. Immunome genes are evenly distributed to evolutionary levels, meaning that the importance of an immunome gene is independent from its evolutionary age.

Immunome genes with highest degree or closeness were collected. These important nodes having high degree in the network are called hubs, and removing or disturbing them affects the network. For example, Özgűr et al. (Özgűr et al., 2008) find out that highest degree, closeness and betweenness genes in the disease gene interaction network were most likely to be related with prostate cancer. Many of the highest degree and closeness immunome genes were overlapping, because there is a strong correlation between degree and closeness. These immunome hub genes have the most central position in the gene networks and thus have most related genes with similar gene expression patterns. One of the main immunome hub genes in this study appeared to be *WAS*. Central role of *WAS* is supported (Maglott et al., 2005) by the fact that mutations in the gene affect actin polymerization and cause Wiskott-Aldrich syndrome. Hutton et al. (Hutton et al., 2004) made the similar analysis defining highly expressed genes of the mouse immunome. Mouse immunome hub genes were by tissue: activated T cells, 17 genes: *CTSZ*, *KPNB1*, *TNFRSF9*, *TNFRSF4*, *MYC*, *MCM2*, *MCM5*, *MCM6*, *MCM7*, *GZMB*, *NCF4*, *GAPD*, *CCl4*, *PCNA*, *RPl1*3, *CD86*, *ICSBP1*; thymus, 7 genes: *SATB1*, *HDAC7A*, *SGPl1*, *ABCA1*, *PRSS16*, *ABCG1*, *C1QG*; stimulated lymph node, 4 genes: *STK10*, *IRF5*, *CXCl9*, *TNFRSF1* (Hutton et al., 2004). When you compare these to human immunome hub genes, only two common genes are found: *TNFRSF9* and *CD86.* Mouse and human are not so far from each other in the history of evolution and they should have common orthologs. Different results could simply be the result of different methods or unreliable results of one or both of the studies.

Immunome tissue networks were created, in which tissues are nodes and immunome genes are edges. Edge betweenness and fast greedy community methods revealed the community structures of tissue networks. Edge betweenness community of SOURCE expression data and unified data revealed communities in which tissues are evenly distributed, with the exception of one bigger group having over 40 tissues. Fast greedy community of SOURCE expression data and unified data has fewer groups and they vary more in size. The tissue clusters were found with same method than the gene clusters. There were 203 immunome tissues from SOURCE expression data, and 547 immunome tissues from unified data. Tissue clusters are however a less significant part of this study than gene clusters, because important tissues of the immune system are already common knowledge. Results indicate that immunome genes are expressed in a wide variety of tissues. Tissues with highest degree and closeness were also overlapping. Lymph, bone marrow and spleen turned out to be the immunome hub tissues. Bone marrow is known to be the place of B cell maturation in mammals, while spleen and lymph nodes are secondary lymphoid organs (Male et al., 2006).

There is a correlation between the degree of a tissue in the immunome tissue network and the number of genes with that tissue in the immunome gene network. This is not a surprise as the degree of a tissue in a tissue network shows how many genes are expressed in that tissue, and the number of genes in a tissue in the gene network should express the same thing. The correlation is stronger in networks of SOURCE expression data than with HPA expression data.

This type of a large scale study of human immunome gene clusters has not been done before and all the achieved results are novel. Methods developed in this search of immunome gene clusters could be used likewise for other types of analyses. After this study I would recommend to use additional expression data and other methods to verify or reject these results. HPA is still growing and it could be used when there is data for all the immunome genes. Properties of R are increasing promptly and there could also be other computational methods to use for this kind of study.

There is still plenty to study something as fascinating and complex as the human immune system. This study is one part of exposing the mysteries of the immune system. It has been a pleasure to be involved in this kind of a project. Perhaps in the future it is possible to build a realistic working model of the human immune system with all the immune system parts involved and all the affecting forces considered. This can be far in the future, but I will enthusiastically follow development of clarifying the immune system function.

6. Conclusion

The main aim of this project was to find out immunome gene clusters. The first part started by creating immunome gene networks of the expression data from HPA and SOURCE databases and their unified data. Modularities of these networks revealed the strong community structure in the network created from SOURCE expression data and the lack of it in network created from HPA expression data. The network created from unified data had modularity between these two. Edge betweenness and fast greedy community analyses were used to reveal community structures of the networks. The network created from HPA expression data was excluded, because it did not have a clear community structure. Common immunome gene clusters were found by searching for gene pairs which appear together in the same community group in both community analyses. There were 547 clustered genes in data from SOURCE database and 566 clustered genes in unified data, so only about 300 genes were eliminated this way. Immunome clustered genes were divided to 88 gene clusters with varying size from 2 to 32 genes. This was the main yield of this study. Genes belonging to the same cluster have similar gene expression patterns and they can be studied further.

Immunome gene cluster data was compared to earlier achieved data of immunome PPI data and the evolutionary age of the genes. Degree and closeness values between immunome clusters genes and immunome PPIs had no apparent correlation, while there was a high correlation between degree and closeness values in immunome clustered genes. There was no trend between degree or closeness and evolution level, which indicated that the location and importance of a gene in the gene network are independent of the evolutionary age of the gene.

The second part started by creating immunome tissue networks of the expression data from HPA and SOURCE databases and their unified data. Edge betweenness and fast greedy community analyses were used to reveal the community structure of tissue networks. Common gene clusters from these two community analyses resulted in 203 immunome tissues from SOURCE expression data and 547 immunome tissues from unified data. These results have, however, less relevance than immunome gene clusters in

this study, because the important tissues of immune system are already common knowledge.
Next the correlation between the degree in the immunome tissue network and the corresponding number of genes in the immunome gene network was checked. There should be a strong correlation between them and that indeed was found to be true.

Immunome hub genes with the highest degree and closeness in the gene network were collected. These central genes of the gene network have most related genes with similar gene expression patterns. Two of the most central genes were *WAS* and *TNFRSF9*. Immunome hub tissues in this analysis were lymph node, bone marrow and spleen, having high degree and closeness in immunome tissue networks from SOURCE expression data and unified data.

An aim of this study was to find tissue ontologies for immunome tissues. They were not utilized in this work, but they could be used further for these results in order to find out information about important immunological tissues.

The main objective of this study was to identify immunome gene groups with similar tissue specificity pattern in their gene expression using these various network analysis tools. This objective was reached successfully by finding 88 immunome gene clusters with 507 genes. Each cluster has genes from two to 32 which are expressed in the same tissues. We could further look for evidence in the literature about the expression of these clustered genes.

7. References

Aderem, A., & Smith, K. D. (2004). A systems approach to dissecting immunity and inflammation. *Seminars in Immunology, 16*(1), 55-67.

Agrawal, A., Eastman, Q. M., & Schatz, D. G. (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature, 394*(6695), 744-751.

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

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Barabási, A., & Oltvai, Z. N. (2004). Network biology: Understanding the Cell´s functional organization. *Nature Reviews Genetics, 5*, 101-113.

Boguski, M. S., Lowe, T. M. J., & Tolstoshev, C. M. (1993). dbEST database for "expressed sequence tags". *Nature Genetics, 4*(4), 332-333.

Camilla Reali, Monica Curto, Valeria Sogos, Franca Scintu, Susanne Pauly, Herbert Schwarz,Fulvia Gremo,. (2003). Expression of CD137 and its ligand in human neurons, astrocytes, and microglia: Modulation by FGF-2. *Journal of Neuroscience Research, 74*(1), 67-73.

Chaplin, D. D. (2006). 1. overview of the human immune response. *Journal of Allergy and Clinical Immunology, 117*(2, Supplement 2), S430-S435.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics, 32*, 490-495.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), 70*(6), 066111.

Cohen, I. R. (2007). Modeling immune behavior for experimentalists. *Immunological Reviews, 216*(1), 232-236.

Csardi, G. (2008). The igraph package. http://cran.stat.auckland.ac.nz/web/packages/igraph/igraph.pdf

Dekker, A. (2007). The eurovision song contest as a "friendship" network. *CONNECTIONS, 27*(3), 53-60.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tomé, P., et al. (1998). A physical map of 30,000 human genes. *Science, 282*(5389), 744-746.

Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J. C., Hernandez-Boussard, T., et al. (2003). SOURCE: A unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Research, 31*(1), 219-223.

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acat. Sci., 5*, 17-61.

Ewing, R. M., Chu, P., Elisma, F., & et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. *Molecular Systems Biology, 3*(89)

Forrest, S., & Beauchemin, C. (2007). Computer immunology. *Immunological Reviews, 216*(1), 176-197.

Freeman, L. (1978). Centrality in social networks conceptual clarification. *Social Networks,* (1), 215-239.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences, 99*(12), 7821-7826.

Granovetter, M. (1973). The strength of weak ties. *The American Journal of Sociology, 78*(6), 1360-1380.

Havlicek, J., & Roberts, S. C. (2009). MHC-correlated mate choice in humans: A review. *Psychoneuroendocrinology, 34*(4), 497-512.

Hutton, J., Jegga, A., Kong, S., Gupta, A., Ebert, C., Williams, S., et al. (2004). Microarray and comparative genomics-based identification of genes and gene regulatory regions of the mouse immune system. *BMC Genomics, 5*(1), 82.

Imai, K., Nonoyama, S & Ochs H. D. (2003). WASP (Wiskott-Aldrich syndrome protein) gene mutations and phenotype. *Current Opinion in Allergy and Clinical Immunology, 3*, 427-436.

Jung, H. W., Choi, S. W., Choi, J. I., & Kwon, B. S. (2004). Serum concentrations of soluble 4-1BB and 4-1BB ligand correlated with the disease severity in rheumatoid arthritis. *Experimental and Molecular Medicine, 36*(1), 13-22.

Kelley, J., de Bono, B., & Trowsdale, J. (2005). IRIS: A database surveying known human immune system genes. *Genomics, 85*(4), 503-511.

Kim, Y., Han, M., & Broxmeyer, H. E. (2008). 4-1BB regulates NKG2D costimulation in human cord blood CD8+ T cells. *Blood, 111*(3), 1378-1386.

Laderach, D., Movassagh, M., Johnson, A., Mittler, R. S., & Galy, A. (2002). 4-1BB co-stimulation enhances human CD8+ T cell priming by augmenting the proliferation and survival of effector CD8+ T cells. *International Immunology, 14*(10), 1155-1167.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., et al. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America, 94*(24), 13057-13062.

Latora, V., & Marchiori, M. (2002). Is the boston subway a small-world network? *Physica A: Statistical Mechanics and its Applications, 314*(1-4), 109-113.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabó, G., Rual, J., et al. (2006). A Protein–Protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell, 125*(4), 801-814.

Louzoun, Y. (2007). The evolution of mathematical immunology. *Immunological Reviews, 216*(1), 9-20.

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2005). Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Research, 33*(suppl_1), D54-58.

Male, D., Brostoff, J., Roth, D., & Roitt, I. (2006). *Immunology - 7th edition*

Milgram, S. (1967). The small world problem. *Psychology Today, 1*, 61-67.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*(2), 026113.

Ortutay, C., Siermala, M., & Vihinen, M. (2007B). Molecular characterization of the immune system: Emergence of proteins, processes, and domains. *Immunogenetics, 59*(5), 333-348.

Ortutay, C., & Vihinen, M. (2007A). Immunome: A reference set of genes and proteins for systems biology of the human immune system. *Cellular Immunology, 244*(2), 87-89.

Ortutay, C., & Vihinen, M. (2008). Efficiency of the immunome protein interaction network increases during evolution. *Immunome Research, 4*(1), 4.

Ozgur, A., Vu, T., Erkan, G., & Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics, 24*(13), i277-285.

Peri, S., Navarro, J. D., Kristiansen, T. Z., Amanchy, R., Surendranath, V., Muthusamy, B., et al. (2004). Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research, 32*(suppl_1), D497-501.

Piirilä, H., Väliaho, J., & Vihinen, M. (2006). Immunodeficiency mutation databases (IDbases). *Human Mutation, 27*(12), 1200-1208.

R Development Core Team. (2005). R: A language and environment for statistical computing. *R Foundation for Statistical Computing,*

Rinkevich, B. (2004). Primitive immune systems: Are your ways my ways? *Immunological Reviews, 198*(1), 25-35.

Rodriguez-Tome, P., & Lijnzaad, P. (2001). RHdb: The radiation hybrid database. *Nucl.Acids Res., 29*(1), 165-166.

Samarghitean, C., Valiaho, J., & Vihinen, M. (2007). IDR knowledge base for primary immunodeficiencies. *Immunome Research, 3*(1), 6.

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science, 270*(5235), 467-470.

Shaw, H. (1954). A study of popular and unpopular children. *Educational Review, 6*(3), 208-220.

Smith, C. M., Finger, J. H., Hayamizu, T. F., McCright, I. J., Eppig, J. T., Kadin, J. A., et al. (2007). The mouse gene expression database (GXD): 2007 update. *Nucl.Acids Res., 35*(suppl_1), D618-623.

Teitell, M., & Richardson, B. (2003). DNA methylation in the immune system. *Clinical Immunology, 109*(1), 2-5.

The UniProt Consortium. (2009). The universal protein resource (UniProt) 2009. *Nucl.Acids Res., 37*(suppl_1), D169-174.

Tsuboi, S., & Meerloo, J. (2007). Wiskott-aldrich syndrome protein is a key regulator of the phagocytic cup formation in macrophages. *Journal of Biological Chemistry, 282*(47), 34194-34203.

Uhlen, M., Bjorling, E., Agaton, C., Szigyarto, C. A., Amini, B., Andersen, E., et al. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular Cellular Proteomics, 4*(12), 1920-1932.

van Bakel, H., & Holstege, F. (2004). In control: Systematic assessment of microarray performance. *EMBO Reports, 5*(10), 964-969.

von Bertalanffy, L. (1950). An outline of general system theory. *The British Journal for the Philosophy of Science, 1*(2), 134-165.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature, 393*, 440-442.

Wheeler, D., Barrett, T., Benson, D., Bryant, S., Canese, K., Chetvernin, V., et al. (2008). Database resources of the national center for biotechnology information. *Nucleic Acids Research, 36*(Database issue), D13-21.

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., et al. (2003). Database resources of the national center for biotechnology. *Nucl.Acids Res., 31*(1), 28-33.

Williams, P. L., Warwick, R., Dyson M., & Bannister, L. H. (1996). *Gray's anatomy 37th edition*. Edinburgh: Churchill Livingstone.